LPM vs. Logit/Probit

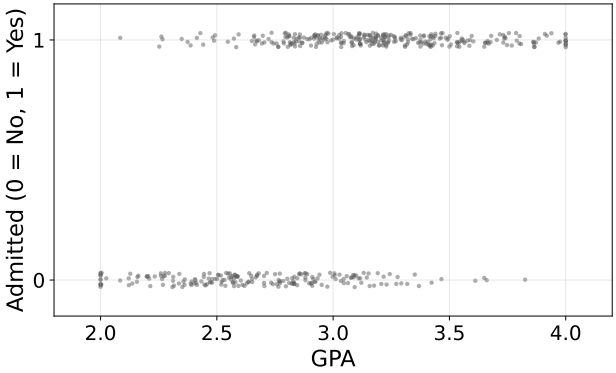
Jake Anderson

May 16, 2026

Outline

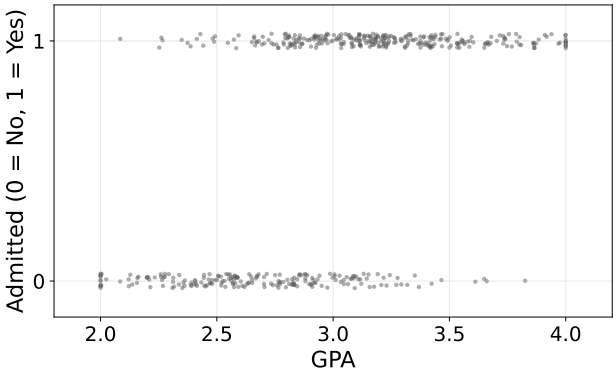
The Data

A university admissions office records **GPA** and **admission outcome** (admit/reject) for 500 applicants.



The Data

A university admissions office records **GPA** and **admission outcome** (admit/reject) for 500 applicants.



The outcome is binary: 0 (rejected) or 1 (admitted). How do we model the probability of admission?

Natural Instinct: Run OLS

The simplest approach: regress the 0/1 outcome on GPA, just like any other regression.

Natural Instinct: Run OLS

The simplest approach: regress the 0/1 outcome on GPA, just like any other regression.

This is the **Linear Probability Model** (LPM):

$$P(\text{Admit}_i = 1 \mid \text{GPA}_i) = \beta_0 + \beta_1 \, \text{GPA}_i$$

Natural Instinct: Run OLS

The simplest approach: regress the 0/1 outcome on GPA, just like any other regression.

This is the **Linear Probability Model** (LPM):

$$P(\text{Admit}_i = 1 \mid \text{GPA}_i) = \beta_0 + \beta_1 \, \text{GPA}_i$$

The coefficients have a direct interpretation:

- β_1 = change in the *probability of admission* for a one-unit increase in GPA

Natural Instinct: Run OLS

The simplest approach: regress the 0/1 outcome on GPA, just like any other regression.

This is the **Linear Probability Model** (LPM):

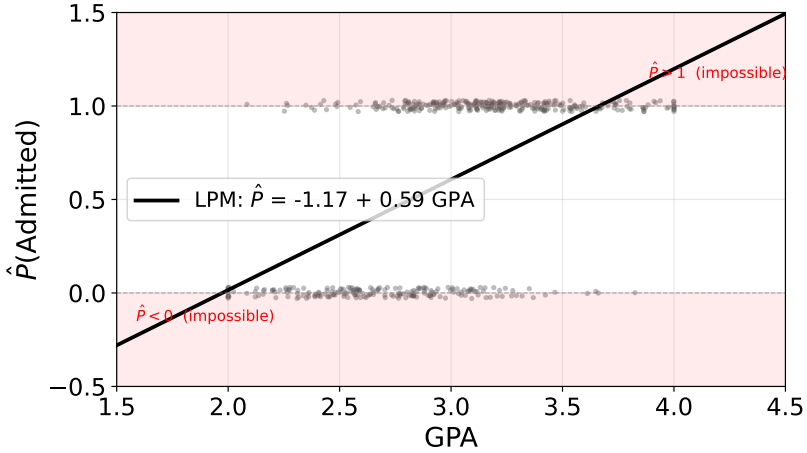
$$P(\text{Admit}_i = 1 \mid \text{GPA}_i) = \beta_0 + \beta_1 \, \text{GPA}_i$$

The coefficients have a direct interpretation:

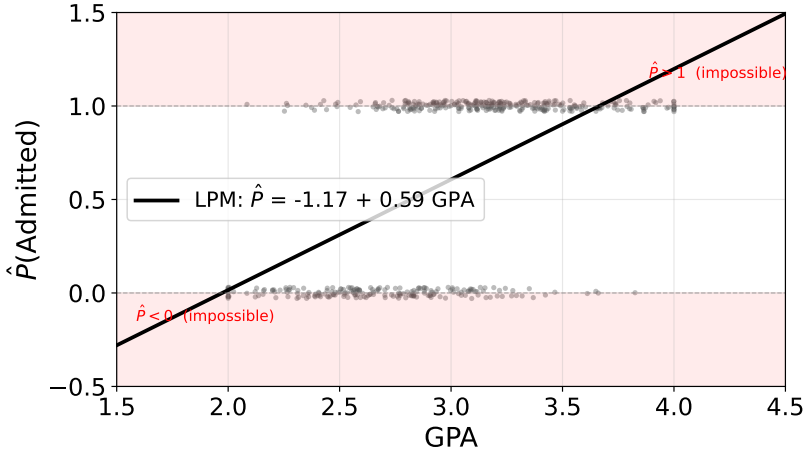
- β_1 = change in the *probability of admission* for a one-unit increase in GPA

Sounds reasonable. Let's see what happens.

LPM Fit

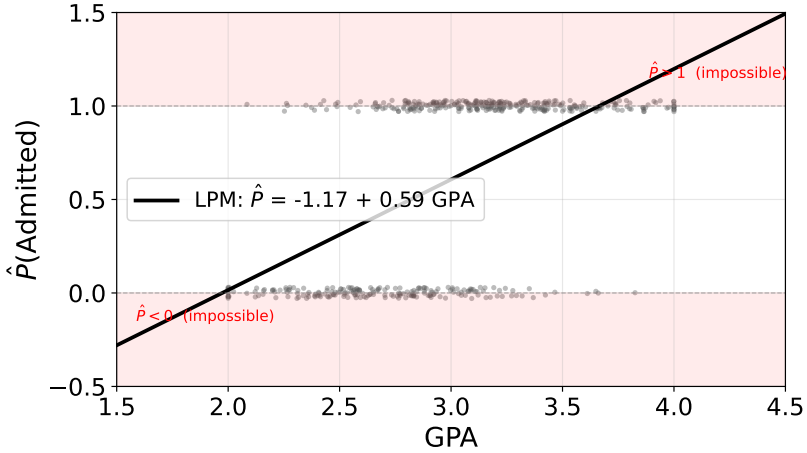

1.5
1.0
0.5
0.0
−0.5
$\hat{P}$(Admitted)
1.5
2.0
2.5
3.0
3.5
4.0
4.5
GPA
LPM: $\hat{P}$ = -1.17 + 0.59 GPA
$\hat{P} > 1$ (impossible)
$\hat{P} < 0$ (impossible)

LPM Fit



$\hat{P}(\text{Admit}) = -1.17 + 0.59 \cdot \text{GPA}$. At GPA $= 4.0$: $\hat{P} = 1.20$. At GPA $= 2.0$: $\hat{P} = 0.02$.

LPM Fit



$\hat{P}(\text{Admit}) = -1.17 + 0.59 \cdot \text{GPA}$. At GPA $= 4.0$: $\hat{P} = 1.20$. At GPA $= 2.0$: $\hat{P} = 0.02$.

$\Longrightarrow$ Probabilities **must** lie in $[0, 1]$. A straight line cannot respect this constraint.

Problem 1: Impossible Predictions

A probability model should produce $\hat{P} \in [0, 1]$ for all observations. The LPM violates this.

Problem 1: Impossible Predictions

A probability model should produce $\hat{P} \in [0, 1]$ for all observations. The LPM violates this.

For any linear function $\hat{P} = \beta_0 + \beta_1 x$:

- If x is large enough $\implies \hat{P} > 1$
- If x is small enough $\implies \hat{P} < 0$

Problem 1: Impossible Predictions

A probability model should produce $\hat{P} \in [0, 1]$ for all observations. The LPM violates this.

For any linear function $\hat{P} = \beta_0 + \beta_1 x$:

- If x is large enough $\implies \hat{P} > 1$
- If x is small enough $\implies \hat{P} < 0$

In our data: applicants with GPA above ≈ 3.7 get predicted probabilities exceeding 1.

Problem 1: Impossible Predictions

A probability model should produce $\hat{P} \in [0, 1]$ for all observations. The LPM violates this.

For any linear function $\hat{P} = \beta_0 + \beta_1 x$:

- If x is large enough $\implies \hat{P} > 1$
- If x is small enough $\implies \hat{P} < 0$

In our data: applicants with GPA above ≈ 3.7 get predicted probabilities exceeding 1.

$\implies$ The LPM is a line forced through inherently nonlinear data. It works in the middle but fails in the tails.

Problem 2: Constant Marginal Effects

The LPM says: each additional GPA point increases admission probability by **0.59**, regardless of where you start.

Problem 2: Constant Marginal Effects

The LPM says: each additional GPA point increases admission probability by **0.59**, regardless of where you start.

- Going from GPA 2.0 to 3.0: $+0.59$ probability
- Going from GPA 3.0 to 4.0: $+0.59$ probability

Problem 2: Constant Marginal Effects

The LPM says: each additional GPA point increases admission probability by **0.59**, regardless of where you start.

- Going from GPA 2.0 to 3.0: $+0.59$ probability
- Going from GPA 3.0 to 4.0: $+0.59$ probability

Is that realistic?

Problem 2: Constant Marginal Effects

The LPM says: each additional GPA point increases admission probability by **0.59**, regardless of where you start.

- Going from GPA 2.0 to 3.0: $+0.59$ probability
- Going from GPA 3.0 to 4.0: $+0.59$ probability

Is that realistic?

No. Consider the S-shaped relationship we expect:

- Near the middle of the GPA range, the probability is changing rapidly (steep part of the curve)
- At the extremes, the probability is near 0 or near 1, so an extra GPA point makes little difference (flat parts of the curve)

Problem 2: Constant Marginal Effects

The LPM says: each additional GPA point increases admission probability by **0.59**, regardless of where you start.

- Going from GPA 2.0 to 3.0: $+0.59$ probability
- Going from GPA 3.0 to 4.0: $+0.59$ probability

Is that realistic?

No. Consider the S-shaped relationship we expect:

- Near the middle of the GPA range, the probability is changing rapidly (steep part of the curve)
- At the extremes, the probability is near 0 or near 1, so an extra GPA point makes little difference (flat parts of the curve)

$\Longrightarrow$ Marginal effects should be **largest near the midpoint** and diminish in the tails, not constant everywhere.

Problem 3: Heteroskedastic Errors

When y can only be 0 or 1, its variance is the variance of a Bernoulli random variable:

$$\text{Var}(y_i \mid x_i) = P(x_i)\big(1 - P(x_i)\big)$$

Problem 3: Heteroskedastic Errors

When y can only be 0 or 1, its variance is the variance of a Bernoulli random variable:

$$\text{Var}(y_i \mid x_i) = P(x_i)\big(1 - P(x_i)\big)$$

This varies with x by construction $\implies$ **heteroskedasticity is guaranteed**.

Problem 3: Heteroskedastic Errors

When y can only be 0 or 1, its variance is the variance of a Bernoulli random variable:

$$\text{Var}(y_i \mid x_i) = P(x_i)\big(1 - P(x_i)\big)$$

This varies with x by construction $\implies$ **heteroskedasticity is guaranteed**.

Consequences:

- OLS coefficients are still **unbiased**
- But OLS standard errors are **wrong** (too small or too large)
- Hypothesis tests and confidence intervals are unreliable

Problem 3: Heteroskedastic Errors

When y can only be 0 or 1, its variance is the variance of a Bernoulli random variable:

$$\text{Var}(y_i \mid x_i) = P(x_i)\big(1 - P(x_i)\big)$$

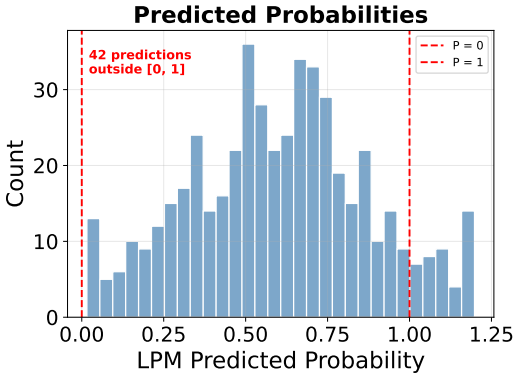
This varies with x by construction $\implies$ **heteroskedasticity is guaranteed**.

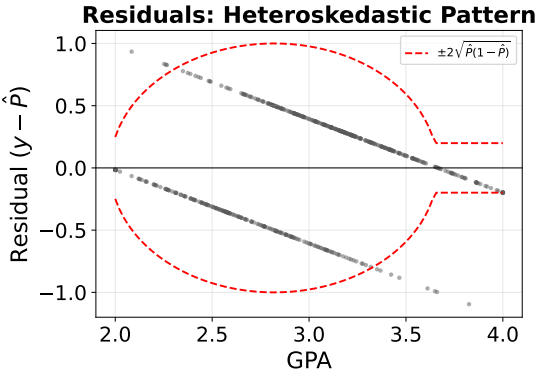
Consequences:

- OLS coefficients are still **unbiased**
- But OLS standard errors are **wrong** (too small or too large)
- Hypothesis tests and confidence intervals are unreliable

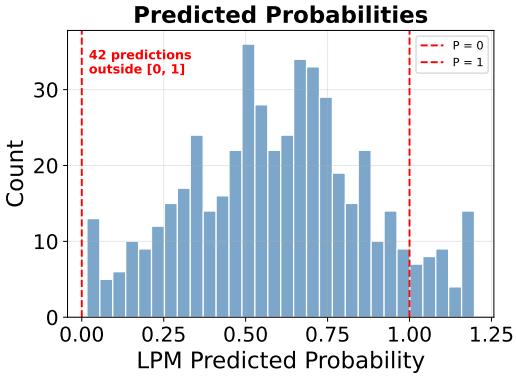
This problem is fixable: robust standard errors correct the SEs. But the impossible predictions and constant marginal effects remain.

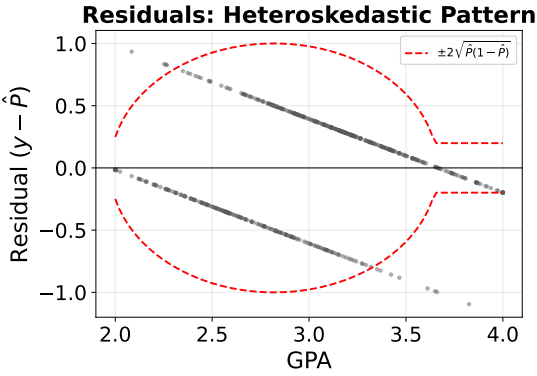
LPM Problems: Visualized





LPM Problems: Visualized





Left: some predictions fall outside $[0, 1]$. Right: residuals fan out, confirming heteroskedasticity.

The Root Cause

All three LPM problems share one structural mismatch:

The Root Cause

All three LPM problems share one structural mismatch:

A straight line has no bounds, but a probability does.

The Root Cause

All three LPM problems share one structural mismatch:

A straight line has no bounds, but a probability does.

- Problems 1 and 2 are two faces of this mismatch:
  - **Out of bounds** $\Longrightarrow$ the line overshoots $[0, 1]$
  - **Constant slope** $\Longrightarrow$ the line cannot flatten as it approaches 0 or 1

The Root Cause

All three LPM problems share one structural mismatch:

A straight line has no bounds, but a probability does.

- Problems 1 and 2 are two faces of this mismatch:
  - **Out of bounds** $\Longrightarrow$ the line overshoots $[0, 1]$
  - **Constant slope** $\Longrightarrow$ the line cannot flatten as it approaches 0 or 1
- Problem 3 (heteroskedasticity) is a byproduct, and can be patched with robust SEs

The Root Cause

All three LPM problems share one structural mismatch:

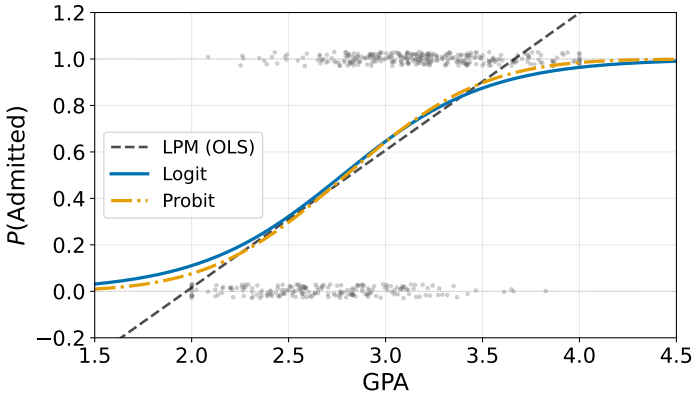
A straight line has no bounds, but a probability does.

- Problems 1 and 2 are two faces of this mismatch:
  - **Out of bounds** $\Longrightarrow$ the line overshoots $[0, 1]$
  - **Constant slope** $\Longrightarrow$ the line cannot flatten as it approaches 0 or 1
- Problem 3 (heteroskedasticity) is a byproduct, and can be patched with robust SEs

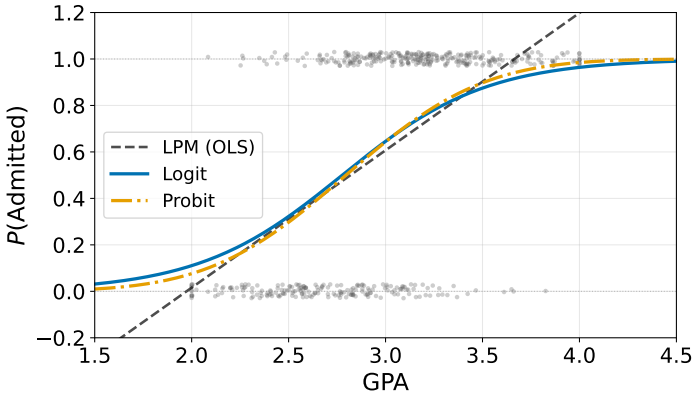
$\Longrightarrow$ We need a **curve**, not a line: something that starts near 0, rises steeply through the middle, and flattens near 1.

Outline

LPM vs. Logit vs. Probit



LPM vs. Logit vs. Probit



The LPM (dashed) overshoots at both ends. Logit and probit replace the line with an S-shaped curve that stays in $[0, 1]$.

Where Does the S-Curve Come From?

Imagine each applicant has a **latent** (unobserved) "admissibility" score:

$$y_i^* = \beta_0 + \beta_1 \, \text{GPA}_i + \varepsilon_i$$

Where Does the S-Curve Come From?

Imagine each applicant has a **latent** (unobserved) "admissibility" score:

$$y_i^* = \beta_0 + \beta_1 \, \text{GPA}_i + \varepsilon_i$$

Latent just means we never see y_i^* directly. We only observe the binary decision:

$$\text{Admit}_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Where Does the S-Curve Come From?

Imagine each applicant has a **latent** (unobserved) "admissibility" score:

$$y_i^* = \beta_0 + \beta_1 \, \text{GPA}_i + \varepsilon_i$$

Latent just means we never see y_i^* directly. We only observe the binary decision:

$$\text{Admit}_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

What is the probability of admission?

$$\begin{aligned} P(\text{Admit} = 1 \mid \text{GPA}) &= P(y^* > 0) \\ &= P(\varepsilon > -\beta_0 - \beta_1 \, \text{GPA}) \end{aligned}$$

Where Does the S-Curve Come From?

Imagine each applicant has a **latent** (unobserved) "admissibility" score:

$$y_i^* = \beta_0 + \beta_1 \, \text{GPA}_i + \varepsilon_i$$

Latent just means we never see y_i^* directly. We only observe the binary decision:

$$\text{Admit}_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

What is the probability of admission?

$$\begin{aligned} P(\text{Admit} = 1 \mid \text{GPA}) &= P(y^* > 0) \\ &= P(\varepsilon > -\beta_0 - \beta_1 \, \text{GPA}) \end{aligned}$$

$\Longrightarrow$ The distribution we assume for ε determines the shape of the curve.

The Latent Variable and Choice Probability

Starting from:

$$P(\text{Admit} = 1) = P(\varepsilon > -\beta_0 - \beta_1 \, \text{GPA})$$

The Latent Variable and Choice Probability

Starting from:

$$P(\text{Admit} = 1) = P(\varepsilon > -\beta_0 - \beta_1 \,\text{GPA})$$

If ε has a *symmetric* distribution (by symmetry of the CDF):

$$P(\varepsilon > -z) = P(\varepsilon < z) = F(z)$$

The Latent Variable and Choice Probability

Starting from:

$$P(\text{Admit} = 1) = P(\varepsilon > -\beta_0 - \beta_1 \, \text{GPA})$$

If ε has a *symmetric* distribution (by symmetry of the CDF):

$$P(\varepsilon > -z) = P(\varepsilon < z) = F(z)$$

So the probability is just the CDF of ε evaluated at $\beta_0 + \beta_1 \, \text{GPA}$:

$$P(\text{Admit} = 1 \mid \text{GPA}) = F(\beta_0 + \beta_1 \, \text{GPA})$$

The Latent Variable and Choice Probability

Starting from:

$$P(\text{Admit} = 1) = P(\varepsilon > -\beta_0 - \beta_1 \,\text{GPA})$$

If ε has a *symmetric* distribution (by symmetry of the CDF):

$$P(\varepsilon > -z) = P(\varepsilon < z) = F(z)$$

So the probability is just the CDF of ε evaluated at $\beta_0 + \beta_1 \,\text{GPA}$:

$$P(\text{Admit} = 1 \mid \text{GPA}) = F(\beta_0 + \beta_1 \,\text{GPA})$$

$\implies$ Any CDF maps $(-\infty, +\infty) \to [0, 1]$, which is exactly what we need. Two standard choices give us two models.

Two Distributions, Two Models

Logistic distribution for ε:

$$P(\text{Admit} = 1 \mid \text{GPA}) = \Lambda(\beta_0 + \beta_1 \text{ GPA}) = \frac{e^{\beta_0 + \beta_1 \text{ GPA}}}{1 + e^{\beta_0 + \beta_1 \text{ GPA}}}$$

This is the **logit** model.

Two Distributions, Two Models

Logistic distribution for ε:

$$P(\text{Admit} = 1 \mid \text{GPA}) = \Lambda(\beta_0 + \beta_1 \, \text{GPA}) = \frac{e^{\beta_0 + \beta_1 \, \text{GPA}}}{1 + e^{\beta_0 + \beta_1 \, \text{GPA}}}$$

This is the **logit** model.

Standard normal distribution for ε:

$$P(\text{Admit} = 1 \mid \text{GPA}) = \Phi(\beta_0 + \beta_1 \, \text{GPA})$$

This is the **probit** model.

Two Distributions, Two Models

Logistic distribution for ε:

$$P(\text{Admit} = 1 \mid \text{GPA}) = \Lambda(\beta_0 + \beta_1 \,\text{GPA}) = \frac{e^{\beta_0 + \beta_1 \,\text{GPA}}}{1 + e^{\beta_0 + \beta_1 \,\text{GPA}}}$$

This is the **logit** model.

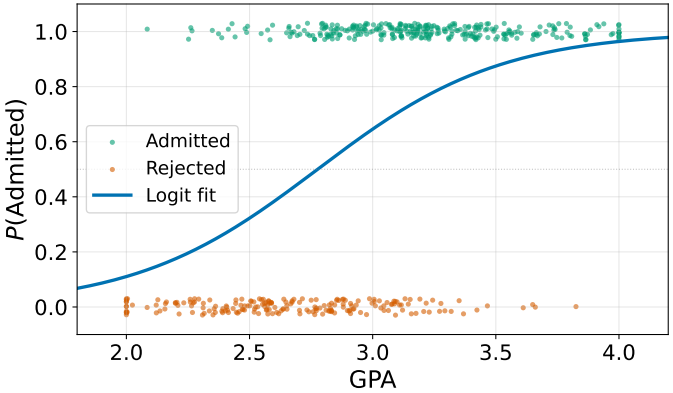
Standard normal distribution for ε:

$$P(\text{Admit} = 1 \mid \text{GPA}) = \Phi(\beta_0 + \beta_1 \,\text{GPA})$$

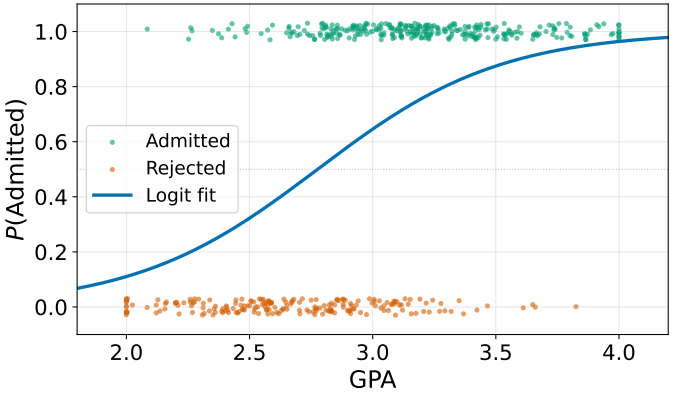
This is the **probit** model.

Both produce S-shaped curves bounded in $[0, 1]$. The logistic CDF (Λ) has slightly heavier tails than the normal CDF (Φ), but in practice the two are nearly indistinguishable.

The Logit Model: Fitted Curve



The Logit Model: Fitted Curve



The logit curve passes through the middle of the data, stays in $[0, 1]$, and has the steepest slope near $P = 0.5$.

Outline

The Interpretation Problem

Our logit estimates: $\hat{\beta}_0 = -7.46$, $\hat{\beta}_1 = 2.69$.

The Interpretation Problem

Our logit estimates: $\hat{\beta}_0 = -7.46$, $\hat{\beta}_1 = 2.69$.

How would you interpret $\hat{\beta}_1 = 2.69$?

The Interpretation Problem

Our logit estimates: $\hat{\beta}_0 = -7.46$, $\hat{\beta}_1 = 2.69$.

How would you interpret $\hat{\beta}_1 = 2.69$?

Tempting but wrong: "A one-unit increase in GPA raises the probability of admission by 2.69."

The Interpretation Problem

Our logit estimates: $\hat{\beta}_0 = -7.46$, $\hat{\beta}_1 = 2.69$.

How would you interpret $\hat{\beta}_1 = 2.69$?

Tempting but wrong: "A one-unit increase in GPA raises the probability of admission by 2.69."

Why wrong? Because the logit is **nonlinear**. The coefficient 2.69 operates on the **log-odds** scale, not the probability scale. A probability change of 2.69 is not even possible.

The Interpretation Problem

Our logit estimates: $\hat{\beta}_0 = -7.46$, $\hat{\beta}_1 = 2.69$.

How would you interpret $\hat{\beta}_1 = 2.69$?

Tempting but wrong: "A one-unit increase in GPA raises the probability of admission by 2.69."

Why wrong? Because the logit is **nonlinear**. The coefficient 2.69 operates on the **log-odds** scale, not the probability scale. A probability change of 2.69 is not even possible.

$\Longrightarrow$ To interpret logit coefficients, we need to understand what they actually measure.

Log-Odds: What the Coefficient Measures

Define the **odds** of admission:

$$\text{Odds} = \frac{P(\text{Admit} = 1)}{P(\text{Admit} = 0)} = \frac{P}{1 - P}$$

Log-Odds: What the Coefficient Measures

Define the **odds** of admission:

$$\text{Odds} = \frac{P(\text{Admit} = 1)}{P(\text{Admit} = 0)} = \frac{P}{1 - P}$$

The logit model is **linear in log-odds**:

$$\underbrace{\ln\left(\frac{P}{1 - P}\right)}_{\text{log-odds}} = \beta_0 + \beta_1 \,\text{GPA}$$

Log-Odds: What the Coefficient Measures

Define the **odds** of admission:

$$\text{Odds} = \frac{P(\text{Admit} = 1)}{P(\text{Admit} = 0)} = \frac{P}{1 - P}$$

The logit model is **linear in log-odds**:

$$\underbrace{\ln\left(\frac{P}{1-P}\right)}_{\text{log-odds}} = \beta_0 + \beta_1 \,\text{GPA}$$

$\implies$ $\beta_1 = 2.69$ means: a one-unit increase in GPA raises the **log-odds** of admission by 2.69.

Log-Odds: What the Coefficient Measures

Define the **odds** of admission:

$$\text{Odds} = \frac{P(\text{Admit} = 1)}{P(\text{Admit} = 0)} = \frac{P}{1 - P}$$

The logit model is **linear in log-odds**:

$$\underbrace{\ln\left(\frac{P}{1-P}\right)}_{\text{log-odds}} = \beta_0 + \beta_1 \,\text{GPA}$$

$\implies$ $\beta_1 = 2.69$ means: a one-unit increase in GPA raises the **log-odds** of admission by 2.69.

Equivalently, the **odds ratio**:

$$e^{\beta_1} = e^{2.69} \approx 14.7$$

A one-unit increase in GPA **multiplies** the odds of admission by ≈ 14.7. For example, going from GPA 2.5 to 3.5 multiplies the odds by this factor.

Marginal Effects: What We Actually Want

The effect on *probability* depends on where you start:

$$\underbrace{\frac{\partial P}{\partial \,\text{GPA}}}_{\text{marginal effect}} = \beta_1 \cdot \Lambda(\beta_0 + \beta_1 \,\text{GPA}) \cdot \big(1 - \Lambda(\beta_0 + \beta_1 \,\text{GPA})\big)$$

Marginal Effects: What We Actually Want

The effect on *probability* depends on where you start:

$$\underbrace{\frac{\partial P}{\partial \,\text{GPA}}}_{\text{marginal effect}} = \beta_1 \cdot \Lambda(\beta_0 + \beta_1 \,\text{GPA}) \cdot \big(1 - \Lambda(\beta_0 + \beta_1 \,\text{GPA})\big)$$

This equals $\beta_1 \cdot P \cdot (1 - P)$, which is largest when $P = 0.5$.

Marginal Effects: What We Actually Want

The effect on *probability* depends on where you start:

$$\underbrace{\frac{\partial P}{\partial \,\text{GPA}}}_{\text{marginal effect}} = \beta_1 \cdot \Lambda(\beta_0 + \beta_1 \,\text{GPA}) \cdot \big(1 - \Lambda(\beta_0 + \beta_1 \,\text{GPA})\big)$$

This equals $\beta_1 \cdot P \cdot (1 - P)$, which is largest when $P = 0.5$.

GPA	$\hat{P}$(Admit)	**Marginal Effect**
2.0	0.11	0.26
2.5	0.32	0.59
3.0	0.65	**0.61**
3.5	0.87	0.29
4.0	0.96	0.09

Marginal Effects: What We Actually Want

The effect on *probability* depends on where you start:

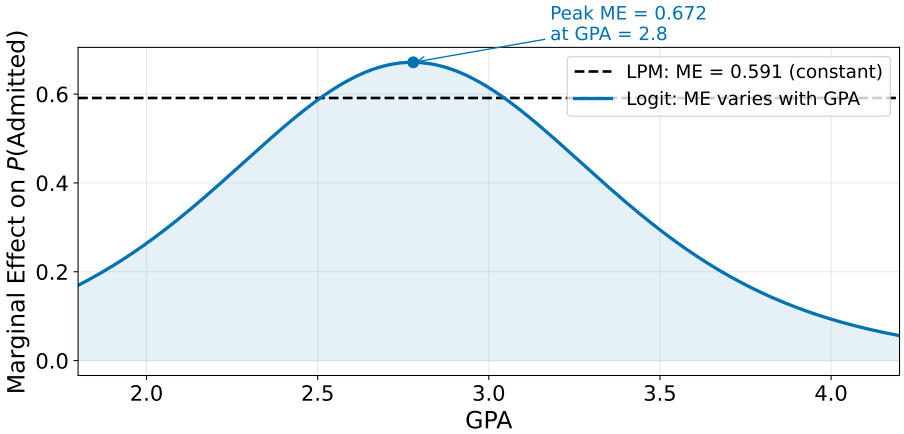
$$\underbrace{\frac{\partial P}{\partial \,\text{GPA}}}_{\text{marginal effect}} = \beta_1 \cdot \Lambda(\beta_0 + \beta_1 \,\text{GPA}) \cdot \big(1 - \Lambda(\beta_0 + \beta_1 \,\text{GPA})\big)$$

This equals $\beta_1 \cdot P \cdot (1 - P)$, which is largest when $P = 0.5$.

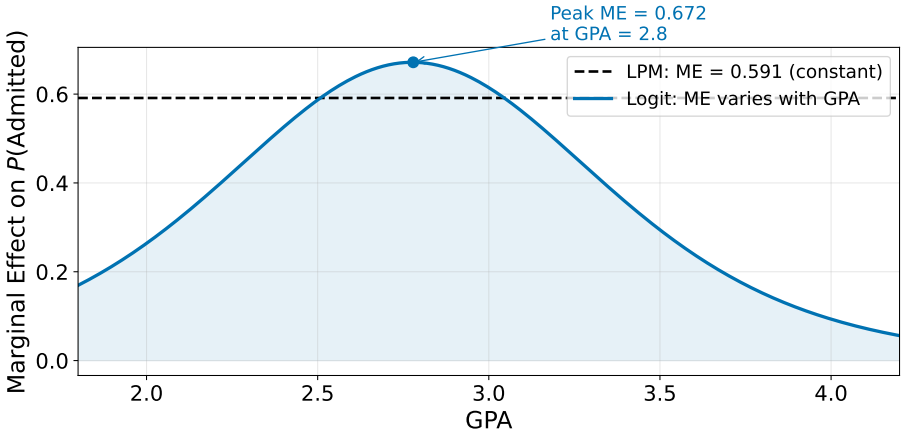
GPA	$\hat{P}$(Admit)	**Marginal Effect**
2.0	0.11	0.26
2.5	0.32	0.59
3.0	0.65	**0.61**
3.5	0.87	0.29
4.0	0.96	0.09

$\Longrightarrow$ The same one-unit GPA increase has roughly 7x more impact near the middle than at the top.

Marginal Effects: Visualized



Marginal Effects: Visualized



The LPM assumes a constant effect (dashed). The logit captures the realistic bell shape: largest effect near $P = 0.5$, vanishing in the tails.

Average Marginal Effect (AME)

Reporting a marginal effect at a single GPA is incomplete. Researchers typically report the **Average Marginal Effect**:

Average Marginal Effect (AME)

Reporting a marginal effect at a single GPA is incomplete. Researchers typically report the **Average Marginal Effect**:

$$\text{AME} = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_1 \cdot \hat{P}_i \cdot (1 - \hat{P}_i)$$

Average Marginal Effect (AME)

Reporting a marginal effect at a single GPA is incomplete. Researchers typically report the **Average Marginal Effect**:

$$\text{AME} = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_1 \cdot \hat{P}_i \cdot (1 - \hat{P}_i)$$

Compute the marginal effect *at each observation's actual GPA*, then average.

Average Marginal Effect (AME)

Reporting a marginal effect at a single GPA is incomplete. Researchers typically report the **Average Marginal Effect**:

$$\text{AME} = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_1 \cdot \hat{P}_i \cdot (1 - \hat{P}_i)$$

Compute the marginal effect *at each observation's actual GPA*, then average.

In our data: AME ≈ 0.49.

"On average, a one-unit increase in GPA is associated with a 49 percentage point increase in the probability of admission."

Average Marginal Effect (AME)

Reporting a marginal effect at a single GPA is incomplete. Researchers typically report the **Average Marginal Effect**:

$$\text{AME} = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_1 \cdot \hat{P}_i \cdot (1 - \hat{P}_i)$$

Compute the marginal effect *at each observation's actual GPA*, then average.

In our data: AME ≈ 0.49.

"On average, a one-unit increase in GPA is associated with a 49 percentage point increase in the probability of admission."

A full GPA point is a large change (e.g., 2.5 to 3.5), so this large AME makes sense in context.

Average Marginal Effect (AME)

Reporting a marginal effect at a single GPA is incomplete. Researchers typically report the **Average Marginal Effect**:

$$\text{AME} = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_1 \cdot \hat{P}_i \cdot (1 - \hat{P}_i)$$

Compute the marginal effect *at each observation's actual GPA*, then average.

In our data: AME ≈ 0.49.

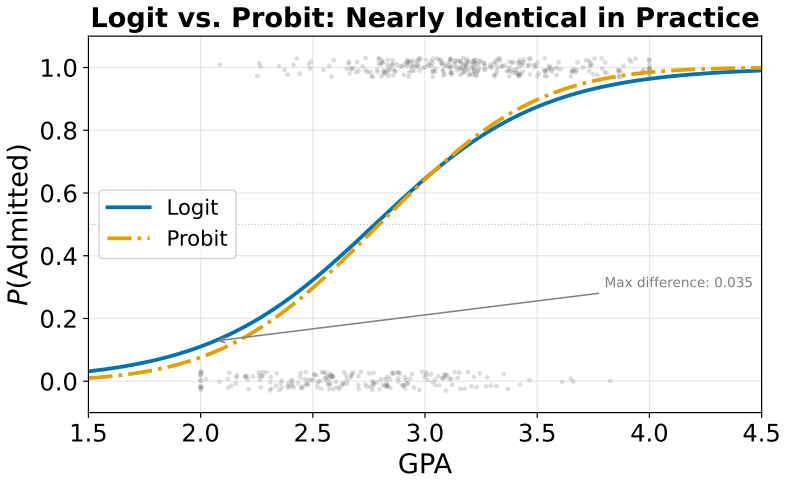
"On average, a one-unit increase in GPA is associated with a 49 percentage point increase in the probability of admission."

A full GPA point is a large change (e.g., 2.5 to 3.5), so this large AME makes sense in context.

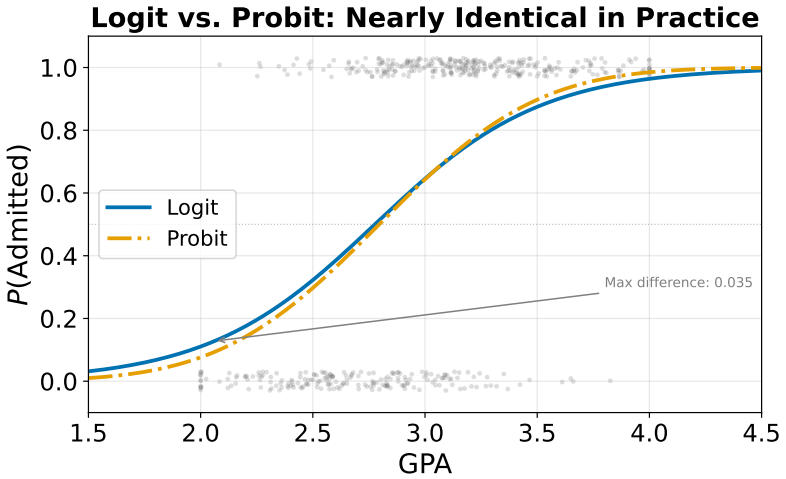
$\implies$ AME gives a single summary number comparable to the LPM coefficient (0.59). The LPM overstates the average effect because it ignores diminishing returns.

Outline

Logit vs. Probit: Nearly Identical



Logit vs. Probit: Nearly Identical



The two curves are almost indistinguishable. The largest difference is in the tails, where both curves are near 0 or 1.

Logit vs. Probit: Coefficients

The logit and probit coefficients are on different scales:

Logit vs. Probit: Coefficients

The logit and probit coefficients are on different scales:

	$\hat{\beta}_0$	$\hat{\beta}_1$	**Scale**
Logit	-7.46	2.69	Log-odds
Probit	-5.03	1.80	z-score (std. normal)

Logit vs. Probit: Coefficients

The logit and probit coefficients are on different scales:

	$\hat{\beta}_0$	$\hat{\beta}_1$	**Scale**
Logit	-7.46	2.69	Log-odds
Probit	-5.03	1.80	z-score (std. normal)

Three numbers you may see for the logit/probit coefficient ratio:

- $\sqrt{\pi^2/3} \approx 1.81$: the *theoretical* ratio, from the fact that the logistic distribution has variance $\pi^2/3$ while the standard normal has variance 1
- ≈ 1.6: a coarser textbook approximation (Amemiya)
- Here: $2.69/1.80 = 1.49$: the actual ratio in this finite sample

Logit vs. Probit: Coefficients

The logit and probit coefficients are on different scales:

	$\hat{\beta}_0$	$\hat{\beta}_1$	**Scale**
Logit	-7.46	2.69	Log-odds
Probit	-5.03	1.80	z-score (std. normal)

Three numbers you may see for the logit/probit coefficient ratio:

- $\sqrt{\pi^2/3} \approx 1.81$: the *theoretical* ratio, from the fact that the logistic distribution has variance $\pi^2/3$ while the standard normal has variance 1
- ≈ 1.6: a coarser textbook approximation (Amemiya)
- Here: $2.69/1.80 = 1.49$: the actual ratio in this finite sample

$\Longrightarrow$ Marginal effects and predicted probabilities are nearly identical regardless. The choice between logit and probit rarely changes conclusions. Logit is more common in economics because of the odds-ratio interpretation.

Outline

In Defense of the LPM

Despite its problems, the LPM is widely used in applied research. When is it acceptable?

In Defense of the LPM

Despite its problems, the LPM is widely used in applied research. When is it acceptable?

The LPM works well when:

1. Predicted probabilities fall in $[0.2, 0.8]$ for most observations

   $\Longrightarrow$ The S-curve is approximately linear in this range

In Defense of the LPM

Despite its problems, the LPM is widely used in applied research. When is it acceptable?

The LPM works well when:

1. Predicted probabilities fall in $[0.2, 0.8]$ for most observations

   $\Longrightarrow$ The S-curve is approximately linear in this range

2. You only need the **average** effect, not predictions at extremes

   $\Longrightarrow$ LPM coefficient $\approx$ AME from logit

In Defense of the LPM

Despite its problems, the LPM is widely used in applied research. When is it acceptable?

The LPM works well when:

1. Predicted probabilities fall in $[0.2, 0.8]$ for most observations

   $\Longrightarrow$ The S-curve is approximately linear in this range

2. You only need the **average** effect, not predictions at extremes

   $\Longrightarrow$ LPM coefficient $\approx$ AME from logit

3. With robust standard errors to correct heteroskedasticity

In Defense of the LPM

Despite its problems, the LPM is widely used in applied research. When is it acceptable?

The LPM works well when:

1. Predicted probabilities fall in $[0.2, 0.8]$ for most observations

   $\Longrightarrow$ The S-curve is approximately linear in this range

2. You only need the **average** effect, not predictions at extremes

   $\Longrightarrow$ LPM coefficient $\approx$ AME from logit

3. With robust standard errors to correct heteroskedasticity

The LPM fails when:

1. You need predictions (e.g., credit scoring, medical diagnosis)
2. The outcome is rare or very common (P near 0 or 1)
3. You have covariates that push predictions far from 0.5

LPM vs. Logit: Decision Framework

	LPM	**Logit / Probit**
Estimation	OLS	MLE
Predictions in $[0, 1]$?	No	Yes
Marginal effects	Constant	Vary with x
Coefficient = ME?	Yes	No (need AME)
Heteroskedasticity	Built in	Handled by MLE
Speed / simplicity	Fastest	Slightly more complex
With FE (many groups)	Easy	Bias risk (incidental parameters problem)

LPM vs. Logit: Decision Framework

	LPM	**Logit / Probit**
Estimation	OLS	MLE
Predictions in $[0, 1]$?	No	Yes
Marginal effects	Constant	Vary with x
Coefficient = ME?	Yes	No (need AME)
Heteroskedasticity	Built in	Handled by MLE
Speed / simplicity	Fastest	Slightly more complex
With FE (many groups)	Easy	Bias risk (incidental parameters problem)

Incidental parameters: with many FE, logit MLE estimates one parameter per group $\implies$ biased coefficients in short panels.

LPM vs. Logit: Decision Framework

	LPM	**Logit / Probit**
Estimation	OLS	MLE
Predictions in $[0, 1]$?	No	Yes
Marginal effects	Constant	Vary with x
Coefficient = ME?	Yes	No (need AME)
Heteroskedasticity	Built in	Handled by MLE
Speed / simplicity	Fastest	Slightly more complex
With FE (many groups)	Easy	Bias risk (incidental parameters problem)

Incidental parameters: with many FE, logit MLE estimates one parameter per group $\implies$ biased coefficients in short panels.

$\implies$ LPM + robust SEs is reasonable as a baseline; switch to logit/probit when predictions or nonlinear effects are central.

Outline

Why Not OLS?

OLS minimizes the sum of squared residuals. With binary y, this creates the problems we saw: impossible predictions, heteroskedasticity.

Why Not OLS?

OLS minimizes the sum of squared residuals. With binary y, this creates the problems we saw: impossible predictions, heteroskedasticity.

Logit and probit use a different estimation criterion: **Maximum Likelihood Estimation** (MLE).

Why Not OLS?

OLS minimizes the sum of squared residuals. With binary y, this creates the problems we saw: impossible predictions, heteroskedasticity.

Logit and probit use a different estimation criterion: **Maximum Likelihood Estimation** (MLE).

MLE idea: Find the parameters β_0, β_1 that make the *observed data* most probable.

Why Not OLS?

OLS minimizes the sum of squared residuals. With binary y, this creates the problems we saw: impossible predictions, heteroskedasticity.

Logit and probit use a different estimation criterion: **Maximum Likelihood Estimation** (MLE).

MLE idea: Find the parameters β_0, β_1 that make the *observed data* most probable.

- For an admitted applicant ($y_i = 1$): we want P_i to be **high**
- For a rejected applicant ($y_i = 0$): we want P_i to be **low**

Why Not OLS?

OLS minimizes the sum of squared residuals. With binary y, this creates the problems we saw: impossible predictions, heteroskedasticity.

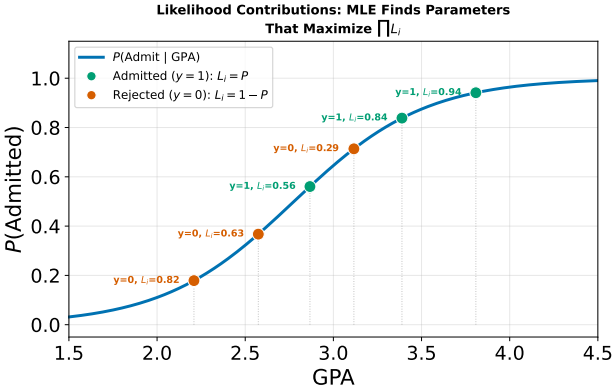
Logit and probit use a different estimation criterion: **Maximum Likelihood Estimation** (MLE).

MLE idea: Find the parameters β_0, β_1 that make the *observed data* most probable.

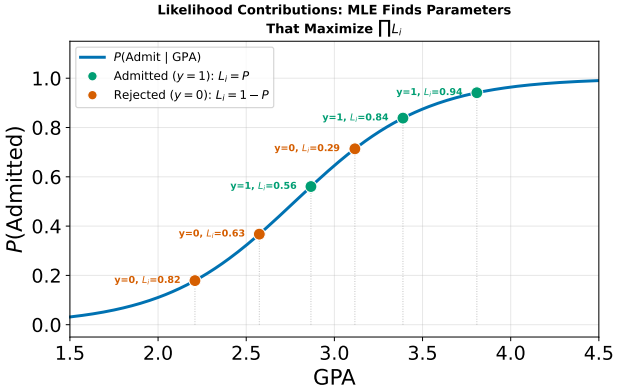
- For an admitted applicant ($y_i = 1$): we want P_i to be **high**
- For a rejected applicant ($y_i = 0$): we want P_i to be **low**

$\Longrightarrow$ MLE finds the S-curve that best separates the admitted from the rejected.

MLE: How It Works



MLE: How It Works



Each observation contributes P_i (if admitted) or $1 - P_i$ (if rejected) to the likelihood. MLE maximizes the product of these contributions.

The Log-Likelihood

The **likelihood** for one observation:

$$L_i = P_i^{y_i} \cdot (1 - P_i)^{1-y_i}$$

The Log-Likelihood

The **likelihood** for one observation:

$$L_i = P_i^{y_i} \cdot (1 - P_i)^{1-y_i}$$

The total **log-likelihood** (sum over all observations):

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big[y_i \ln P_i + (1 - y_i) \ln(1 - P_i) \Big]$$

The Log-Likelihood

The **likelihood** for one observation:

$$L_i = P_i^{y_i} \cdot (1 - P_i)^{1-y_i}$$

The total **log-likelihood** (sum over all observations):

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big[y_i \ln P_i + (1 - y_i) \ln(1 - P_i) \Big]$$

where $P_i = \Lambda(\beta_0 + \beta_1 \, \text{GPA}_i)$ for logit.

The Log-Likelihood

The **likelihood** for one observation:

$$L_i = P_i^{y_i} \cdot (1 - P_i)^{1-y_i}$$

The total **log-likelihood** (sum over all observations):

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big[y_i \ln P_i + (1 - y_i) \ln(1 - P_i) \Big]$$

where $P_i = \Lambda(\beta_0 + \beta_1 \, \text{GPA}_i)$ for logit.

No closed-form solution $\implies$ solved numerically (Newton-Raphson, gradient ascent). Software handles this automatically.

The Log-Likelihood

The **likelihood** for one observation:

$$L_i = P_i^{y_i} \cdot (1 - P_i)^{1-y_i}$$

The total **log-likelihood** (sum over all observations):

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big[y_i \ln P_i + (1 - y_i) \ln(1 - P_i) \Big]$$

where $P_i = \Lambda(\beta_0 + \beta_1 \, \text{GPA}_i)$ for logit.

No closed-form solution $\implies$ solved numerically (Newton-Raphson, gradient ascent). Software handles this automatically.

$\implies$ MLE is the standard estimation method for logit and probit. The resulting $\hat{\beta}$ values are the ones that maximize this log-likelihood.

Thank you!
jakeanderson@g.ucla.edu

Ordered Probit / Ordered Logit

Modeling Outcomes That Have a Ranking but Not a Scale

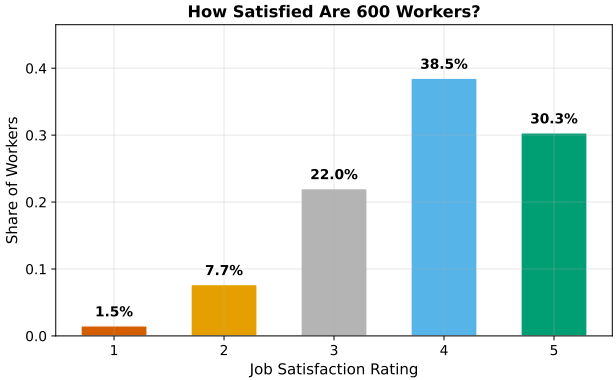
Jake Anderson

May 16, 2026

Outline

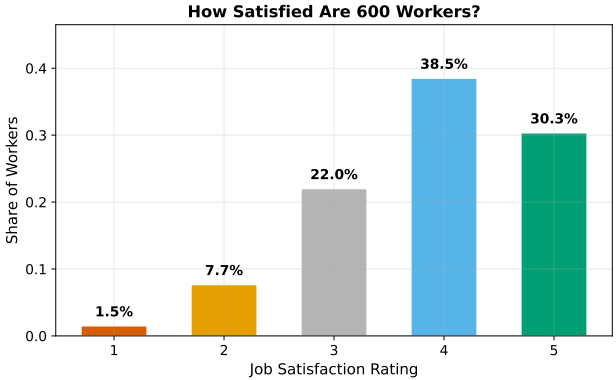
The Data

A firm surveys **600 workers** about job satisfaction. Each worker reports a rating from 1 (very dissatisfied) to 5 (very satisfied).



The Data

A firm surveys **600 workers** about job satisfaction. Each worker reports a rating from 1 (very dissatisfied) to 5 (very satisfied).



The outcome is **ordinal**: $5 > 4 > 3 > 2 > 1$, but the distances between categories are not meaningful. Is the gap from 1 to 2 the same as 4 to 5?

Does Wage Predict Satisfaction?

Does Wage Predict Satisfaction?

Higher-paid workers tend to report higher ratings. But the outcome takes only five discrete values. How should we model this?

First Instinct: Run OLS

Treat the rating as a continuous number and regress it on wage:

$$\text{Rating}_i = \beta_0 + \beta_1 \text{Wage}_i + \varepsilon_i$$

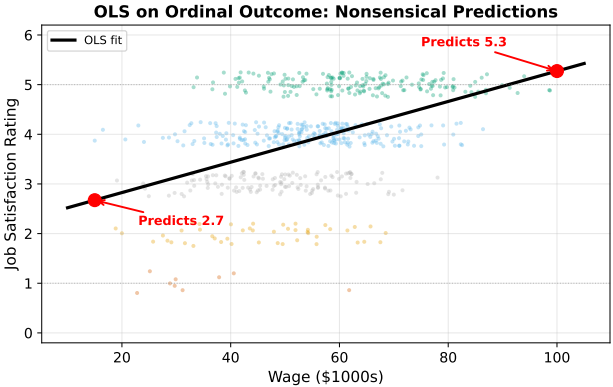
First Instinct: Run OLS

Treat the rating as a continuous number and regress it on wage:

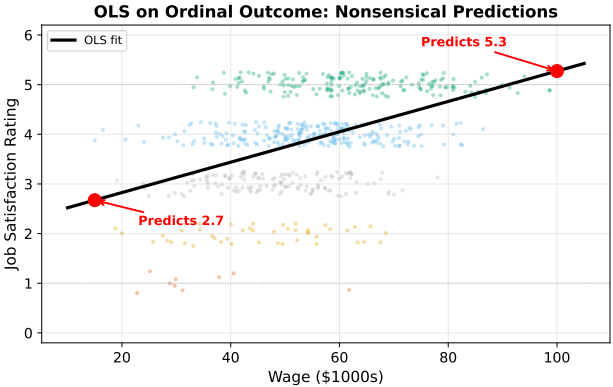
$$\text{Rating}_i = \beta_0 + \beta_1 \, \text{Wage}_i + \varepsilon_i$$

This is fast and gives a slope you can interpret. What could go wrong?

OLS on Ordinal Outcomes: The Failure



OLS on Ordinal Outcomes: The Failure



At high wages, OLS predicts **5.3**. But the scale only goes to 5. At low wages, it predicts non-integer values between categories. OLS treats ordinal categories as if they were continuous and equally spaced.

What Goes Wrong with OLS on Ordinal Data

Three problems:

1. **Predictions outside the valid range.** OLS can predict 0.5 or 5.3 when the outcome only takes values 1–5

What Goes Wrong with OLS on Ordinal Data

Three problems:

1. **Predictions outside the valid range.** OLS can predict 0.5 or 5.3 when the outcome only takes values 1–5
2. **Equal-spacing assumption.** OLS treats the jump from 1 to 2 as identical to the jump from 4 to 5. The numerical labels are arbitrary; replacing 1–5 with 2, 4, 6, 8, 10 would change the slope

What Goes Wrong with OLS on Ordinal Data

Three problems:

1. **Predictions outside the valid range.** OLS can predict 0.5 or 5.3 when the outcome only takes values 1–5
2. **Equal-spacing assumption.** OLS treats the jump from 1 to 2 as identical to the jump from 4 to 5. The numerical labels are arbitrary; replacing 1–5 with 2, 4, 6, 8, 10 would change the slope
3. **Constant marginal effects.** OLS forces each \$1k wage increase to add the same amount to the predicted rating, regardless of where on the scale the worker starts

What Goes Wrong with OLS on Ordinal Data

Three problems:

1. **Predictions outside the valid range.** OLS can predict 0.5 or 5.3 when the outcome only takes values 1–5
2. **Equal-spacing assumption.** OLS treats the jump from 1 to 2 as identical to the jump from 4 to 5. The numerical labels are arbitrary; replacing 1–5 with 2, 4, 6, 8, 10 would change the slope
3. **Constant marginal effects.** OLS forces each \$1k wage increase to add the same amount to the predicted rating, regardless of where on the scale the worker starts

$\Longrightarrow$ We need a model that respects the ordinal nature of the outcome: categories have a ranking, but the distances between them are unknown.

What Would a Better Model Look Like?

A proper model for ordinal outcomes should:

What Would a Better Model Look Like?

A proper model for ordinal outcomes should:

1. **Produce valid probabilities.** $P(\text{Rating} = j) \in (0, 1)$, summing to 1.

What Would a Better Model Look Like?

A proper model for ordinal outcomes should:

1. **Produce valid probabilities.** $P(\text{Rating} = j) \in (0, 1)$, summing to 1.
2. **Respect the ordering.** Use $5 > 4 > 3 > 2 > 1$ without assuming equal spacing.

What Would a Better Model Look Like?

A proper model for ordinal outcomes should:

1. **Produce valid probabilities.** $P(\text{Rating} = j) \in (0, 1)$, summing to 1.
2. **Respect the ordering.** Use $5 > 4 > 3 > 2 > 1$ without assuming equal spacing.
3. **Allow flexible marginal effects.** Effect of a wage increase can depend on where the worker sits on the scale.

What Would a Better Model Look Like?

A proper model for ordinal outcomes should:

1. **Produce valid probabilities.** $P(\text{Rating} = j) \in (0, 1)$, summing to 1.
2. **Respect the ordering.** Use $5 > 4 > 3 > 2 > 1$ without assuming equal spacing.
3. **Allow flexible marginal effects.** Effect of a wage increase can depend on where the worker sits on the scale.

$\Longrightarrow$ Where can we find a model with these properties?

The Latent Continuous Variable Behind Ordered Ratings

Think about what the 1–5 scale really represents.

The Latent Continuous Variable Behind Ordered Ratings

Think about what the 1–5 scale really represents.

Satisfaction is probably **continuous** in a worker's mind. A worker who reports "4" and one who reports "5" might differ by a hair; another pair of 4 and 5 reporters might differ enormously.

The Latent Continuous Variable Behind Ordered Ratings

Think about what the 1–5 scale really represents.

Satisfaction is probably **continuous** in a worker's mind. A worker who reports "4" and one who reports "5" might differ by a hair; another pair of 4 and 5 reporters might differ enormously.

The ratings are a **coarse discretization** of something continuous. What if we modeled that continuous variable directly and let the discrete ratings emerge from it?

The Latent Continuous Variable Behind Ordered Ratings

Think about what the 1–5 scale really represents.

Satisfaction is probably **continuous** in a worker's mind. A worker who reports "4" and one who reports "5" might differ by a hair; another pair of 4 and 5 reporters might differ enormously.

The ratings are a **coarse discretization** of something continuous. What if we modeled that continuous variable directly and let the discrete ratings emerge from it?

$\Longrightarrow$ This is the latent variable approach: posit an unobserved continuous satisfaction level, then map it to the observed categories through thresholds.

Outline

The Idea: A Continuous Variable Behind Discrete Ratings

Imagine each worker has a latent (unobserved) satisfaction level y_i^* that is continuous:

$$y_i^* = \beta_1 \,\text{Wage}_i + \beta_2 \,\text{Hours}_i + \beta_3 \,\text{Support}_i + \varepsilon_i$$

(no intercept; it is absorbed into the cutpoints, explained shortly)

The Idea: A Continuous Variable Behind Discrete Ratings

Imagine each worker has a latent (unobserved) satisfaction level y_i^* that is continuous:

$$y_i^* = \beta_1 \, \text{Wage}_i + \beta_2 \, \text{Hours}_i + \beta_3 \, \text{Support}_i + \varepsilon_i$$

(no intercept; it is absorbed into the cutpoints, explained shortly)

- y_i^* can take any real value (no boundary problems)
- We never observe y_i^* directly
- We observe only which **category** it falls into

The Idea: A Continuous Variable Behind Discrete Ratings

Imagine each worker has a latent (unobserved) satisfaction level y_i^* that is continuous:

$$y_i^* = \beta_1 \, \text{Wage}_i + \beta_2 \, \text{Hours}_i + \beta_3 \, \text{Support}_i + \varepsilon_i$$

(no intercept; it is absorbed into the cutpoints, explained shortly)

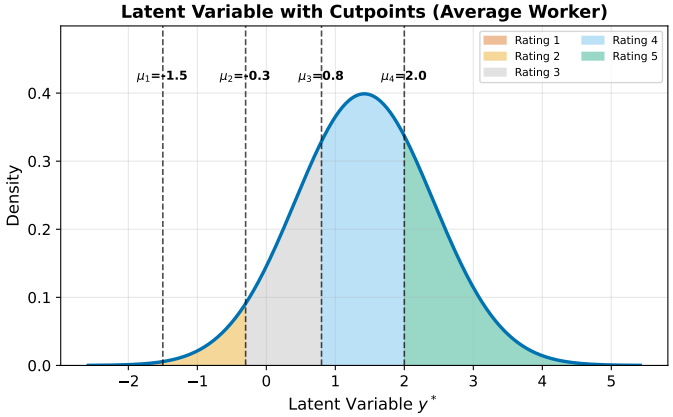
- y_i^* can take any real value (no boundary problems)
- We never observe y_i^* directly
- We observe only which **category** it falls into

The mapping from y_i^* to the observed rating uses **cutpoints** $\mu_1 < \mu_2 < \mu_3 < \mu_4$:

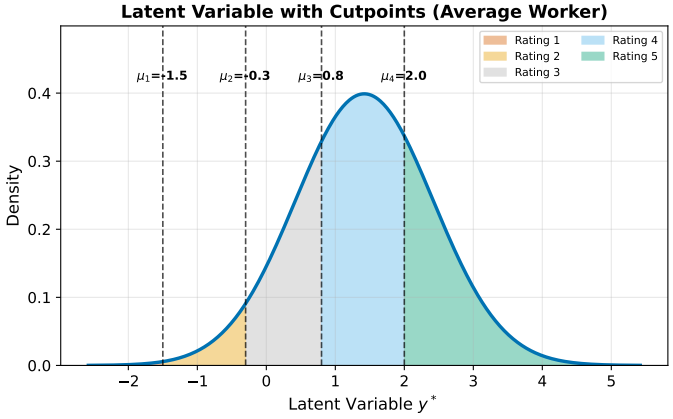
$$\text{Rating}_i = j \quad \Longleftrightarrow \quad \mu_{j-1} < y_i^* \leq \mu_j$$

where $\mu_0 = -\infty$ and $\mu_5 = +\infty$.

Cutpoints Partition the Latent Variable



Cutpoints Partition the Latent Variable



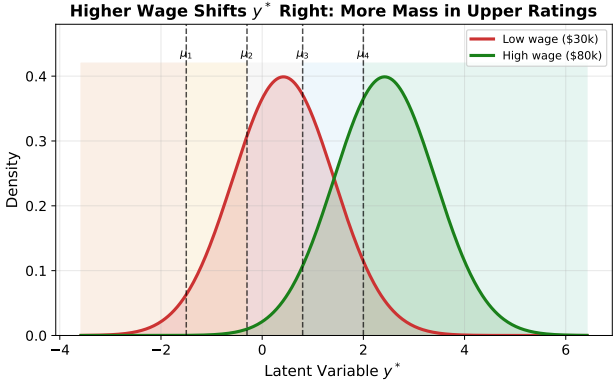
The density of y_i^* is divided into five regions by four cutpoints. The area in each region equals the probability of that rating.

How Covariates Shift the Distribution

Define the **linear index**: $\text{XB}_i = \beta_1 \text{Wage}_i + \beta_2 \text{Hours}_i + \beta_3 \text{Support}_i$.

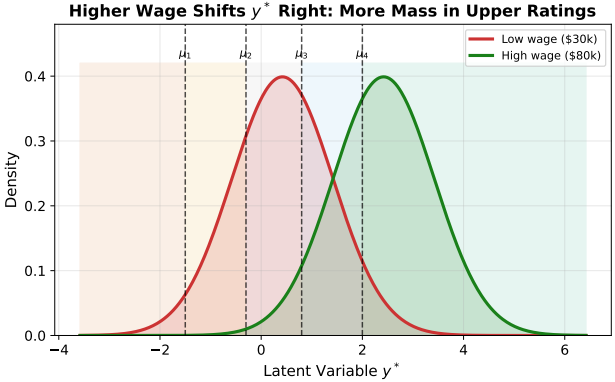
How Covariates Shift the Distribution

Define the **linear index**: $\text{XB}_i = \beta_1 \text{Wage}_i + \beta_2 \text{Hours}_i + \beta_3 \text{Support}_i$.
A higher wage increases XB_i, which shifts the entire density of y_i^* to the right.



How Covariates Shift the Distribution

Define the **linear index**: $\text{XB}_i = \beta_1 \text{Wage}_i + \beta_2 \text{Hours}_i + \beta_3 \text{Support}_i$.
A higher wage increases XB_i, which shifts the entire density of y_i^* to the right.



The cutpoints stay fixed; the density moves. Positive β raises P(highest) and lowers P(lowest); middle categories are ambiguous.

The Probability Formula: Derivation

The probability of observing rating j depends on where y_i^* falls:

$$P(\text{Rating}_i = j) = P(\mu_{j-1} < y_i^* \leq \mu_j)$$

The Probability Formula: Derivation

The probability of observing rating j depends on where y_i^* falls:

$$P(\text{Rating}_i = j) = P(\mu_{j-1} < y_i^* \leq \mu_j)$$

Substitute $y_i^* = \text{XB}_i + \varepsilon_i$ and rearrange:

$$= P\big(\mu_{j-1} - \text{XB}_i < \varepsilon_i \leq \mu_j - \text{XB}_i\big)$$

The Probability Formula: Derivation

The probability of observing rating j depends on where y_i^* falls:

$$P(\text{Rating}_i = j) = P(\mu_{j-1} < y_i^* \leq \mu_j)$$

Substitute $y_i^* = \text{XB}_i + \varepsilon_i$ and rearrange:

$$= P\big(\mu_{j-1} - \text{XB}_i < \varepsilon_i \leq \mu_j - \text{XB}_i\big)$$

Since ε_i has CDF $F(\cdot)$:

$$= F(\mu_j - \text{XB}_i) - F(\mu_{j-1} - \text{XB}_i)$$

The Probability Formula: Derivation

The probability of observing rating j depends on where y_i^* falls:

$$P(\text{Rating}_i = j) = P(\mu_{j-1} < y_i^* \leq \mu_j)$$

Substitute $y_i^* = \text{XB}_i + \varepsilon_i$ and rearrange:

$$= P\big(\mu_{j-1} - \text{XB}_i < \varepsilon_i \leq \mu_j - \text{XB}_i\big)$$

Since ε_i has CDF $F(\cdot)$:

$$= F(\mu_j - \text{XB}_i) - F(\mu_{j-1} - \text{XB}_i)$$

with boundary conditions $F(-\infty) = 0$ and $F(+\infty) = 1$.

Probit vs. Logit: Choosing $F(\cdot)$

The general formula works with any CDF. The two standard choices:

Probit vs. Logit: Choosing $F(\cdot)$

The general formula works with any CDF. The two standard choices:

- **Ordered probit:** $F = \Phi$ (standard normal CDF), so $\varepsilon_i \sim N(0, 1)$

Probit vs. Logit: Choosing $F(\cdot)$

The general formula works with any CDF. The two standard choices:

- **Ordered probit:** $F = \Phi$ (standard normal CDF), so $\varepsilon_i \sim N(0, 1)$
- **Ordered logit:** $F = \Lambda$ (logistic CDF), so $\varepsilon_i \sim$ Logistic

Probit vs. Logit: Choosing $F(\cdot)$

The general formula works with any CDF. The two standard choices:

- **Ordered probit:** $F = \Phi$ (standard normal CDF), so $\varepsilon_i \sim N(0, 1)$
- **Ordered logit:** $F = \Lambda$ (logistic CDF), so $\varepsilon_i \sim$ Logistic

$\Longrightarrow$ Same idea, different distributional assumption on ε_i. Results are usually similar in practice. Economists tend to use ordered probit; biostatisticians often prefer ordered logit.

What Gets Estimated

The model estimates two sets of parameters:

What Gets Estimated

The model estimates two sets of parameters:

1. **Coefficients** $(\beta_1, \beta_2, \beta_3)$, i.e. $(\beta_{\text{wage}}, \beta_{\text{hours}}, \beta_{\text{support}})$
   How each predictor shifts the latent variable y_i^*

What Gets Estimated

The model estimates two sets of parameters:

1. **Coefficients** $(\beta_1, \beta_2, \beta_3)$, i.e. $(\beta_{\text{wage}}, \beta_{\text{hours}}, \beta_{\text{support}})$
   How each predictor shifts the latent variable y_i^*
2. **Cutpoints** $\mu_1, \mu_2, \mu_3, \mu_4$
   Where the boundaries between adjacent categories lie

What Gets Estimated

The model estimates two sets of parameters:

1. **Coefficients** $(\beta_1, \beta_2, \beta_3)$, i.e. $(\beta_{\text{wage}}, \beta_{\text{hours}}, \beta_{\text{support}})$
   How each predictor shifts the latent variable y_i^*
2. **Cutpoints** $\mu_1, \mu_2, \mu_3, \mu_4$
   Where the boundaries between adjacent categories lie

There is **no intercept** in the equation. An intercept would be absorbed into the cutpoints (you cannot separately identify both), so we normalize by omitting it.

The Parallel Regressions Assumption

The ordered model assumes that the coefficients $(\beta_1, \beta_2, \beta_3)$ are the **same for every cutpoint**.
Visually: when a covariate changes, the density shifts horizontally, but all cutpoints stay fixed.

The Parallel Regressions Assumption

The ordered model assumes that the coefficients $(\beta_1, \beta_2, \beta_3)$ are the **same for every cutpoint**. Visually: when a covariate changes, the density shifts horizontally, but all cutpoints stay fixed.

Where the name comes from: imagine running separate binary logits for each cumulative split:

- Rating ≤ 1 vs. > 1
- Rating ≤ 2 vs. > 2
- Rating ≤ 3 vs. > 3
- Rating ≤ 4 vs. > 4

The Parallel Regressions Assumption

The ordered model assumes that the coefficients $(\beta_1, \beta_2, \beta_3)$ are the **same for every cutpoint**. Visually: when a covariate changes, the density shifts horizontally, but all cutpoints stay fixed.

Where the name comes from: imagine running separate binary logits for each cumulative split:

- Rating ≤ 1 vs. > 1
- Rating ≤ 2 vs. > 2
- Rating ≤ 3 vs. > 3
- Rating ≤ 4 vs. > 4

The "parallel regressions" assumption says the slope on each predictor would be the **same across all four splits**. Only the intercept (cutpoint) differs.

The Parallel Regressions Assumption

The ordered model assumes that the coefficients $(\beta_1, \beta_2, \beta_3)$ are the **same for every cutpoint**. Visually: when a covariate changes, the density shifts horizontally, but all cutpoints stay fixed.

Where the name comes from: imagine running separate binary logits for each cumulative split:

- Rating ≤ 1 vs. > 1
- Rating ≤ 2 vs. > 2
- Rating ≤ 3 vs. > 3
- Rating ≤ 4 vs. > 4

The "parallel regressions" assumption says the slope on each predictor would be the **same across all four splits**. Only the intercept (cutpoint) differs.

$\Longrightarrow$ This is what makes the ordered model parsimonious: one set of slopes instead of four.

Numeric Example: Computing Probabilities

Suppose a worker earns \$55k, works 42 hours, and has supervisor support ($= 1$). With the ordered probit coefficients $\beta_{\text{wage}} = 0.04$, $\beta_{\text{hours}} = -0.03$, $\beta_{\text{support}} = 0.80$:

Numeric Example: Computing Probabilities

Suppose a worker earns \$55k, works 42 hours, and has supervisor support ($= 1$). With the ordered probit coefficients $\beta_{\text{wage}} = 0.04$, $\beta_{\text{hours}} = -0.03$, $\beta_{\text{support}} = 0.80$:

$$\text{XB}_i = 0.04 \times 55 + (-0.03) \times 42 + 0.80 \times 1 = 2.20 - 1.26 + 0.80 = 1.74$$

Numeric Example: Computing Probabilities

Suppose a worker earns \$55k, works 42 hours, and has supervisor support ($= 1$). With the ordered probit coefficients $\beta_{\text{wage}} = 0.04$, $\beta_{\text{hours}} = -0.03$, $\beta_{\text{support}} = 0.80$:

$$\text{XB}_i = 0.04 \times 55 + (-0.03) \times 42 + 0.80 \times 1 = 2.20 - 1.26 + 0.80 = 1.74$$

With cutpoints $\mu_1 = -1.5$, $\mu_2 = -0.3$, $\mu_3 = 0.8$, $\mu_4 = 2.0$:

Numeric Example: Computing Probabilities

Suppose a worker earns \$55k, works 42 hours, and has supervisor support ($= 1$). With the ordered probit coefficients $\beta_{\text{wage}} = 0.04$, $\beta_{\text{hours}} = -0.03$, $\beta_{\text{support}} = 0.80$:

$$\text{XB}_i = 0.04 \times 55 + (-0.03) \times 42 + 0.80 \times 1 = 2.20 - 1.26 + 0.80 = 1.74$$

With cutpoints $\mu_1 = -1.5$, $\mu_2 = -0.3$, $\mu_3 = 0.8$, $\mu_4 = 2.0$:

$$\begin{aligned} P(\text{Rating} = 4) &= \Phi(\mu_4 - 1.74) - \Phi(\mu_3 - 1.74) \\ &= \Phi(2.0 - 1.74) - \Phi(0.8 - 1.74) \\ &= \Phi(0.26) - \Phi(-0.94) \\ &= 0.603 - 0.174 = 0.429 \end{aligned}$$

Numeric Example: Computing Probabilities

Suppose a worker earns \$55k, works 42 hours, and has supervisor support ($= 1$). With the ordered probit coefficients $\beta_{\text{wage}} = 0.04$, $\beta_{\text{hours}} = -0.03$, $\beta_{\text{support}} = 0.80$:

$$\text{XB}_i = 0.04 \times 55 + (-0.03) \times 42 + 0.80 \times 1 = 2.20 - 1.26 + 0.80 = 1.74$$

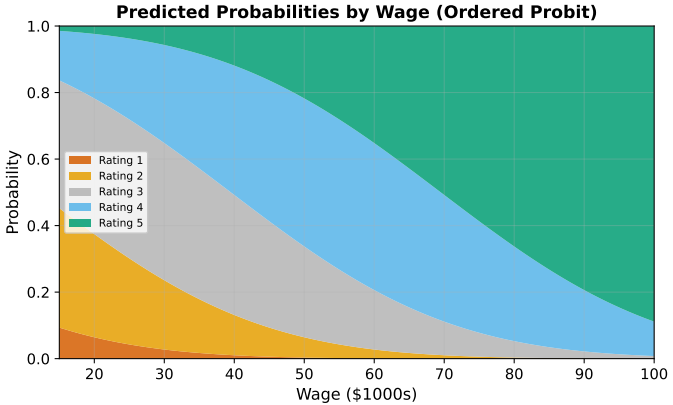
With cutpoints $\mu_1 = -1.5$, $\mu_2 = -0.3$, $\mu_3 = 0.8$, $\mu_4 = 2.0$:

$$\begin{aligned} P(\text{Rating} = 4) &= \Phi(\mu_4 - 1.74) - \Phi(\mu_3 - 1.74) \\ &= \Phi(2.0 - 1.74) - \Phi(0.8 - 1.74) \\ &= \Phi(0.26) - \Phi(-0.94) \\ &= 0.603 - 0.174 = 0.429 \end{aligned}$$

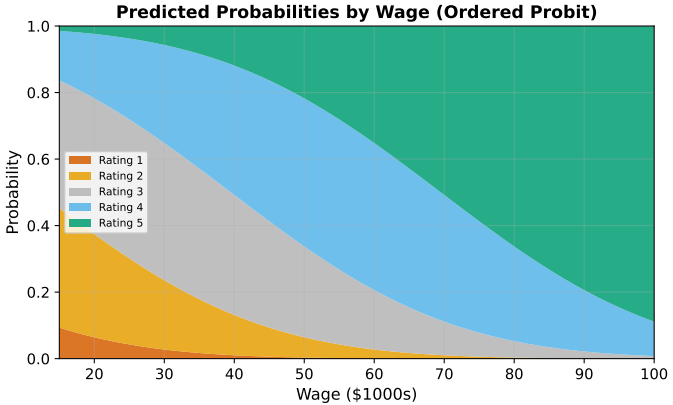
$\Longrightarrow$ This worker has a 42.9% chance of reporting "Satisfied" (Rating 4). Software computes all five probabilities simultaneously.

Outline

Predicted Probabilities



Predicted Probabilities



As wage increases, probability shifts from lower ratings to higher ratings. At every wage level, the five probabilities sum to 1.

Coefficient Interpretation: Sign Only

The estimated coefficients tell us the **direction** in which a predictor shifts y_i^*:

Coefficient Interpretation: Sign Only

The estimated coefficients tell us the **direction** in which a predictor shifts y_i^*:

Variable	$\hat{\beta}$	**Interpretation**
Wage ($1k)	$+$	Higher wage $\Longrightarrow$ y^* shifts right $\Longrightarrow$ higher satisfaction
Hours	$-$	More hours $\Longrightarrow$ y^* shifts left $\Longrightarrow$ lower satisfaction
Supervisor support	$+$	Support $\Longrightarrow$ y^* shifts right $\Longrightarrow$ higher satisfaction

Coefficient Interpretation: Sign Only

The estimated coefficients tell us the **direction** in which a predictor shifts y_i^*:

Variable	$\hat{\beta}$	**Interpretation**
Wage ($1k)	$+$	Higher wage $\Longrightarrow$ y^* shifts right $\Longrightarrow$ higher satisfaction
Hours	$-$	More hours $\Longrightarrow$ y^* shifts left $\Longrightarrow$ lower satisfaction
Supervisor support	$+$	Support $\Longrightarrow$ y^* shifts right $\Longrightarrow$ higher satisfaction

$\Longrightarrow$ You can interpret the **sign**, but not the magnitude directly. Saying "a $1k raise increases satisfaction by 0.04" is wrong because 0.04 is in latent-variable units, which have no natural scale.

Coefficient Interpretation: Sign Only

The estimated coefficients tell us the **direction** in which a predictor shifts y_i^*:

Variable	$\hat{\beta}$	**Interpretation**
Wage ($1k)	$+$	Higher wage $\Longrightarrow$ y^* shifts right $\Longrightarrow$ higher satisfaction
Hours	$-$	More hours $\Longrightarrow$ y^* shifts left $\Longrightarrow$ lower satisfaction
Supervisor support	$+$	Support $\Longrightarrow$ y^* shifts right $\Longrightarrow$ higher satisfaction

$\Longrightarrow$ You can interpret the **sign**, but not the magnitude directly. Saying "a $1k raise increases satisfaction by 0.04" is wrong because 0.04 is in latent-variable units, which have no natural scale.

For magnitudes, compute **marginal effects** on probabilities.

Marginal Effects: The Formula

The marginal effect on any single category depends on how much density sits near the two cutpoints that bracket that category.

Marginal Effects: The Formula

The marginal effect on any single category depends on how much density sits near the two cutpoints that bracket that category.

The marginal effect of variable x_k on $P(\text{Rating} = j)$ is:

$$\frac{\partial P(\text{Rating} = j)}{\partial x_k} = \Big[f(\mu_{j-1} - \text{XB}_i) - f(\mu_j - \text{XB}_i) \Big] \cdot \beta_k$$

Marginal Effects: The Formula

The marginal effect on any single category depends on how much density sits near the two cutpoints that bracket that category.

The marginal effect of variable x_k on $P(\text{Rating} = j)$ is:

$$\frac{\partial P(\text{Rating} = j)}{\partial x_k} = \Big[f(\mu_{j-1} - \text{XB}_i) - f(\mu_j - \text{XB}_i) \Big] \cdot \beta_k$$

where $f(\cdot)$ is the density (derivative of F).

Marginal Effects: The Formula

The marginal effect on any single category depends on how much density sits near the two cutpoints that bracket that category.

The marginal effect of variable x_k on $P(\text{Rating} = j)$ is:

$$\frac{\partial P(\text{Rating} = j)}{\partial x_k} = \Big[f(\mu_{j-1} - \text{XB}_i) - f(\mu_j - \text{XB}_i) \Big] \cdot \beta_k$$

where $f(\cdot)$ is the density (derivative of F).

For the **extreme categories**, one of the boundary terms drops out:

- Rating 1: $\text{ME} = -f(\mu_1 - \text{XB}_i) \cdot \beta_k$ ($\beta_k > 0 \implies$ negative)
- Rating 5: $\text{ME} = f(\mu_4 - \text{XB}_i) \cdot \beta_k$ ($\beta_k > 0 \implies$ positive)

Marginal Effects: The Formula

The marginal effect on any single category depends on how much density sits near the two cutpoints that bracket that category.

The marginal effect of variable x_k on $P(\text{Rating} = j)$ is:

$$\frac{\partial P(\text{Rating} = j)}{\partial x_k} = \Big[f(\mu_{j-1} - \text{XB}_i) - f(\mu_j - \text{XB}_i) \Big] \cdot \beta_k$$

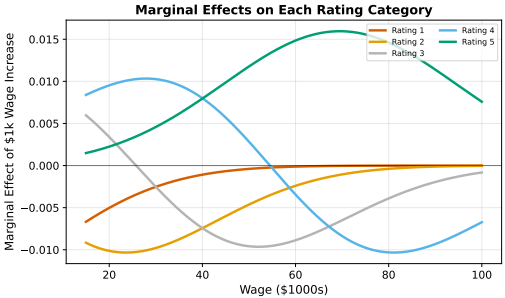
where $f(\cdot)$ is the density (derivative of F).

For the **extreme categories**, one of the boundary terms drops out:

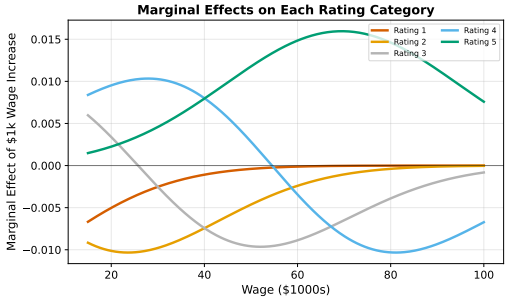
- Rating 1: $\text{ME} = -f(\mu_1 - \text{XB}_i) \cdot \beta_k$ ($\beta_k > 0 \implies$ negative)
- Rating 5: $\text{ME} = f(\mu_4 - \text{XB}_i) \cdot \beta_k$ ($\beta_k > 0 \implies$ positive)

$\implies$ A positive coefficient always decreases $P(\text{lowest})$ and increases $P(\text{highest})$. But what about the middle categories?

Middle Categories: Ambiguous Marginal Effects

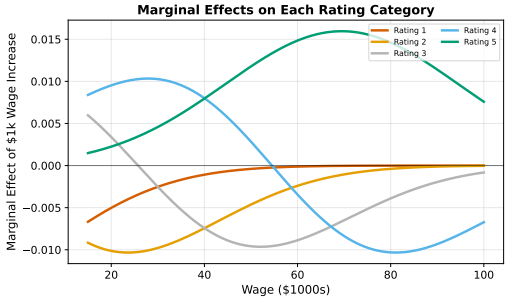


Middle Categories: Ambiguous Marginal Effects



The marginal effect on **middle categories changes sign** with the worker's starting point: at low wages a raise increases $P(\text{Rating} = 4)$; at high wages it decreases it (workers there are already shifting into Rating 5).

Middle Categories: Ambiguous Marginal Effects



The marginal effect on **middle categories changes sign** with the worker's starting point: at low wages a raise increases $P(\text{Rating} = 4)$; at high wages it decreases it (workers there are already shifting into Rating 5).

$\Longrightarrow$ Report Average Marginal Effects (AMEs) for each category, not just the coefficient.

Estimation: Maximum Likelihood

Each worker contributes one term to the likelihood: the probability of the rating they actually reported.

Estimation: Maximum Likelihood

Each worker contributes one term to the likelihood: the probability of the rating they actually reported.

The log-likelihood:

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{5} d_{ij} \ln P(\text{Rating}_i = j)$$

where $d_{ij} = 1$ if worker i reported rating j.

Estimation: Maximum Likelihood

Each worker contributes one term to the likelihood: the probability of the rating they actually reported.

The log-likelihood:

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{5} d_{ij} \ln P(\text{Rating}_i = j)$$

where $d_{ij} = 1$ if worker i reported rating j.

- $P(\text{Rating}_i = j) = F(\mu_j - \text{XB}_i) - F(\mu_{j-1} - \text{XB}_i)$
- Parameters: $(\beta_1, \beta_2, \beta_3)$ and $\mu_1, \ldots, \mu_4$
- No closed-form solution $\implies$ numerical optimization (same as binary logit)

Estimation: Maximum Likelihood

Each worker contributes one term to the likelihood: the probability of the rating they actually reported.

The log-likelihood:

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{5} d_{ij} \ln P(\text{Rating}_i = j)$$

where $d_{ij} = 1$ if worker i reported rating j.

- $P(\text{Rating}_i = j) = F(\mu_j - \text{XB}_i) - F(\mu_{j-1} - \text{XB}_i)$
- Parameters: $(\beta_1, \beta_2, \beta_3)$ and $\mu_1, \ldots, \mu_4$
- No closed-form solution $\implies$ numerical optimization (same as binary logit)

$\implies$ Software estimates the coefficients and the cutpoints jointly.

Outline

This Is Not Multinomial Logit

Ordered choice and multinomial logit both handle categorical outcomes. But they are built for different situations.

This Is Not Multinomial Logit

Ordered choice and multinomial logit both handle categorical outcomes. But they are built for different situations.

Multinomial logit: categories have **no natural ordering** (Car, Bus, Bike, Walk). Each alternative gets its **own coefficient** (β_j per mode).

This Is Not Multinomial Logit

Ordered choice and multinomial logit both handle categorical outcomes. But they are built for different situations.

Multinomial logit: categories have **no natural ordering** (Car, Bus, Bike, Walk). Each alternative gets its **own coefficient** (β_j per mode).

Ordered choice: categories have a **natural ranking** ($5 > 4 > 3 > 2 > 1$). One set of coefficients ($\beta_1, \ldots, \beta_K$) shifts the entire latent distribution; the ordering is exploited rather than discarded.

This Is Not Multinomial Logit

Ordered choice and multinomial logit both handle categorical outcomes. But they are built for different situations.

Multinomial logit: categories have **no natural ordering** (Car, Bus, Bike, Walk). Each alternative gets its **own coefficient** (β_j per mode).

Ordered choice: categories have a **natural ranking** ($5 > 4 > 3 > 2 > 1$). One set of coefficients ($\beta_1, \ldots, \beta_K$) shifts the entire latent distribution; the ordering is exploited rather than discarded.

$\Longrightarrow$ Using multinomial logit on ordered data wastes the ordering information and estimates far more parameters than necessary ($J - 1$ coefficient vectors instead of one).

When Parallel Regressions Fails

Recall the parallel regressions assumption: the same slopes apply at every cumulative split. When does this break down?

When Parallel Regressions Fails

Recall the parallel regressions assumption: the same slopes apply at every cumulative split. When does this break down?

Example: suppose wage has a strong effect on moving from "Dissatisfied" to "Neutral," but no effect on moving from "Satisfied" to "Very Satisfied." The ordered model cannot capture this because it forces a single β_{wage}.

When Parallel Regressions Fails

Recall the parallel regressions assumption: the same slopes apply at every cumulative split. When does this break down?

Example: suppose wage has a strong effect on moving from "Dissatisfied" to "Neutral," but no effect on moving from "Satisfied" to "Very Satisfied." The ordered model cannot capture this because it forces a single β_{wage}.

Test: the Brant test checks whether the slopes are stable across cutpoints (note: applies specifically to ordered logit). If it rejects, consider:

- A **generalized ordered logit** (which allows slopes to vary by cutpoint)
- Multinomial logit as a fallback

When Parallel Regressions Fails

Recall the parallel regressions assumption: the same slopes apply at every cumulative split. When does this break down?

Example: suppose wage has a strong effect on moving from "Dissatisfied" to "Neutral," but no effect on moving from "Satisfied" to "Very Satisfied." The ordered model cannot capture this because it forces a single β_{wage}.

Test: the Brant test checks whether the slopes are stable across cutpoints (note: applies specifically to ordered logit). If it rejects, consider:

- A **generalized ordered logit** (which allows slopes to vary by cutpoint)
- Multinomial logit as a fallback

$\Longrightarrow$ Always check parallel regressions before reporting ordered model results.

Ordered Probit vs. Ordered Logit

The only difference is the assumed distribution of ε_i:

Ordered Probit vs. Ordered Logit

The only difference is the assumed distribution of ε_i:

	Ordered Probit	**Ordered Logit**
Error distribution	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim$ Logistic
CDF used	$\Phi(\cdot)$	$\Lambda(\cdot) = \frac{e^{(\cdot)}}{1+e^{(\cdot)}}$
Tails	Thinner	Slightly heavier
Predicted probabilities	Nearly identical	Nearly identical

Ordered Probit vs. Ordered Logit

The only difference is the assumed distribution of ε_i:

	Ordered Probit	**Ordered Logit**
Error distribution	$\varepsilon_i \sim N(0,1)$	$\varepsilon_i \sim$ Logistic
CDF used	$\Phi(\cdot)$	$\Lambda(\cdot) = \frac{e^{(\cdot)}}{1+e^{(\cdot)}}$
Tails	Thinner	Slightly heavier
Predicted probabilities	Nearly identical	Nearly identical

$\Longrightarrow$ In practice, both give very similar results. Choose based on convention in your field.

Decision Framework: Which Model to Use

1. **Binary** (yes/no, 0/1) $\Longrightarrow$ Binary logit or probit

Decision Framework: Which Model to Use

1. **Binary** (yes/no, 0/1) $\Longrightarrow$ Binary logit or probit
2. **Categorical, no natural order** (car, bus, bike, walk) $\Longrightarrow$ Multinomial logit

Decision Framework: Which Model to Use

1. **Binary** (yes/no, 0/1) $\implies$ Binary logit or probit
2. **Categorical, no natural order** (car, bus, bike, walk) $\implies$ Multinomial logit
3. **Categorical, natural ranking** (strongly disagree $\rightarrow$ strongly agree) $\implies$ Ordered probit or ordered logit

Decision Framework: Which Model to Use

1. **Binary** (yes/no, 0/1) $\implies$ Binary logit or probit
2. **Categorical, no natural order** (car, bus, bike, walk) $\implies$ Multinomial logit
3. **Categorical, natural ranking** (strongly disagree $\rightarrow$ strongly agree) $\implies$ Ordered probit or ordered logit
4. **Parallel regressions assumption fails** $\implies$ Generalized ordered logit, or fall back to multinomial logit

Decision Framework: Which Model to Use

1. **Binary** (yes/no, 0/1) $\implies$ Binary logit or probit
2. **Categorical, no natural order** (car, bus, bike, walk) $\implies$ Multinomial logit
3. **Categorical, natural ranking** (strongly disagree $\rightarrow$ strongly agree) $\implies$ Ordered probit or ordered logit
4. **Parallel regressions assumption fails** $\implies$ Generalized ordered logit, or fall back to multinomial logit

$\implies$ The ordered model is more efficient than multinomial logit when the ordering is genuine, because it estimates fewer parameters while exploiting the ranking structure.

Summary: Back to Job Satisfaction

Recall the three OLS problems. How does ordered probit address each one?

Summary: Back to Job Satisfaction

Recall the three OLS problems. How does ordered probit address each one?

1. **Predictions outside the valid range** $\implies$ probabilities $\in (0, 1)$ that sum to 1. No predictions of 5.3.

Summary: Back to Job Satisfaction

Recall the three OLS problems. How does ordered probit address each one?

1. **Predictions outside the valid range** $\implies$ probabilities $\in (0, 1)$ that sum to 1. No predictions of 5.3.
2. **Equal-spacing assumption** $\implies$ cutpoints $\mu_1, \ldots, \mu_4$ are estimated freely; no equal-distance assumption.

Summary: Back to Job Satisfaction

Recall the three OLS problems. How does ordered probit address each one?

1. **Predictions outside the valid range** $\Longrightarrow$ probabilities $\in (0, 1)$ that sum to 1. No predictions of 5.3.
2. **Equal-spacing assumption** $\Longrightarrow$ cutpoints $\mu_1, \ldots, \mu_4$ are estimated freely; no equal-distance assumption.
3. **Constant marginal effects** $\Longrightarrow$ marginal effects depend on the worker's current position on the satisfaction scale.

Summary: Back to Job Satisfaction

Recall the three OLS problems. How does ordered probit address each one?

1. **Predictions outside the valid range** $\Longrightarrow$ probabilities $\in (0, 1)$ that sum to 1. No predictions of 5.3.
2. **Equal-spacing assumption** $\Longrightarrow$ cutpoints $\mu_1, \ldots, \mu_4$ are estimated freely; no equal-distance assumption.
3. **Constant marginal effects** $\Longrightarrow$ marginal effects depend on the worker's current position on the satisfaction scale.

$\Longrightarrow$ Interpret the sign of $\hat{\beta}$ for direction; compute AMEs for magnitude. Check the parallel regressions assumption before reporting.

Thank you!
jakeanderson@g.ucla.edu

Multinomial Logit / Conditional Logit

Why You Can't Just Run Four Logits

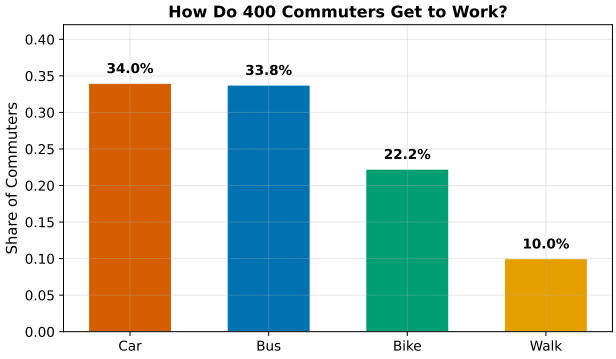
Jake Anderson

May 16, 2026

Outline

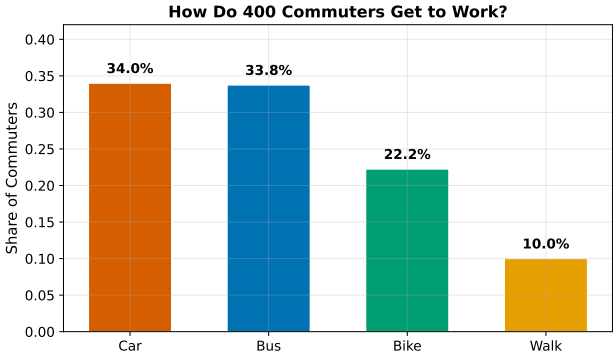
The Data

A city surveys **400 commuters** about how they get to work. Each person picks one of four modes: **Car**, **Bus**, **Bike**, or **Walk**.



The Data

A city surveys **400 commuters** about how they get to work. Each person picks one of four modes: **Car**, **Bus**, **Bike**, or **Walk**.



The outcome is **categorical with 4 levels**. Binary logit handles 2 options. How do we handle 4?

First Instinct: Run Separate Binary Logits

You already know binary logit. Why not run one for each mode?

First Instinct: Run Separate Binary Logits

You already know binary logit. Why not run one for each mode?

1. Logit for Car vs. Not Car
2. Logit for Bus vs. Not Bus
3. Logit for Bike vs. Not Bike
4. Logit for Walk vs. Not Walk

First Instinct: Run Separate Binary Logits

You already know binary logit. Why not run one for each mode?

1. Logit for Car vs. Not Car
2. Logit for Bus vs. Not Bus
3. Logit for Bike vs. Not Bike
4. Logit for Walk vs. Not Walk

Each model gives $\hat{P}(\text{mode})$ as a function of income. Seems straightforward.

First Instinct: Run Separate Binary Logits

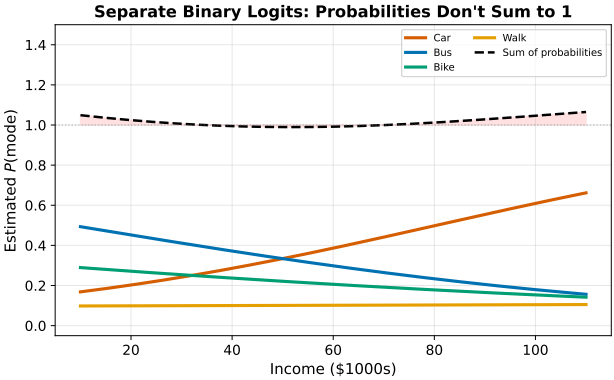
You already know binary logit. Why not run one for each mode?

1. Logit for Car vs. Not Car
2. Logit for Bus vs. Not Bus
3. Logit for Bike vs. Not Bike
4. Logit for Walk vs. Not Walk

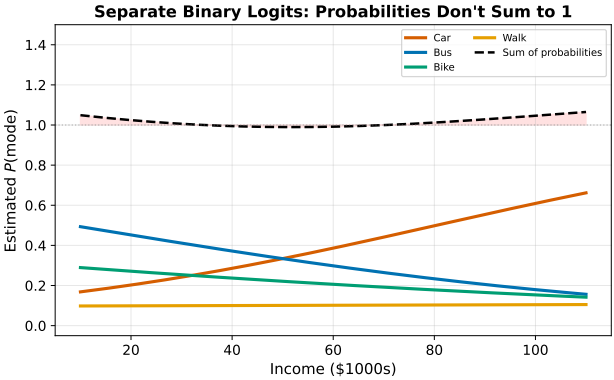
Each model gives $\hat{P}(\text{mode})$ as a function of income. Seems straightforward.

But there is a structural problem: four separate models know nothing about each other.

Separate Binary Logits: The Failure

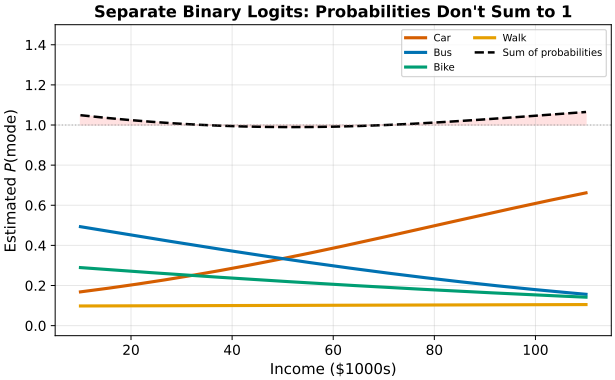


Separate Binary Logits: The Failure



The four probabilities don't sum to 1. At some income levels, the total exceeds 1; at others, it falls short.

Separate Binary Logits: The Failure



The four probabilities don't sum to 1. At some income levels, the total exceeds 1; at others, it falls short.

$\Longrightarrow$ Separate binary models violate a basic requirement: a commuter must choose **exactly one** mode, so probabilities must sum to 1 across alternatives.

What Went Wrong, and What We Need

The problem: separate binary logits model each mode in isolation.

- Each model is free to assign any probability between 0 and 1
- Nothing forces the four probabilities to coordinate
- $\Longrightarrow$ They don't sum to 1, so they aren't valid choice probabilities

What Went Wrong, and What We Need

The problem: separate binary logits model each mode in isolation.

- Each model is free to assign any probability between 0 and 1
- Nothing forces the four probabilities to coordinate
- $\Longrightarrow$ They don't sum to 1, so they aren't valid choice probabilities

The requirement: we need a single model that assigns probabilities to *all four modes simultaneously*, with:

- Every $P(y_i = j) \in (0, 1)$
- $\sum_{j=1}^{4} P(y_i = j) = 1$ by construction

What Went Wrong, and What We Need

The problem: separate binary logits model each mode in isolation.

- Each model is free to assign any probability between 0 and 1
- Nothing forces the four probabilities to coordinate
- $\Longrightarrow$ They don't sum to 1, so they aren't valid choice probabilities

The requirement: we need a single model that assigns probabilities to *all four modes simultaneously*, with:

- Every $P(y_i = j) \in (0, 1)$
- $\sum_{j=1}^{4} P(y_i = j) = 1$ by construction

$\Longrightarrow$ The multinomial logit model achieves exactly this.

Outline

Random Utility Framework

Each commuter i assigns a **utility** to each mode j:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

Random Utility Framework

Each commuter i assigns a **utility** to each mode j:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

- V_{ij} = the part we can model (income, travel time, cost)
- ε_{ij} = unobserved taste variation (idiosyncratic preferences)

Random Utility Framework

Each commuter i assigns a **utility** to each mode j:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

- V_{ij} = the part we can model (income, travel time, cost)
- ε_{ij} = unobserved taste variation (idiosyncratic preferences)

Decision rule: commuter i chooses mode j if $U_{ij} > U_{ik}$ for every other mode k.

Random Utility Framework

Each commuter i assigns a **utility** to each mode j:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

- V_{ij} = the part we can model (income, travel time, cost)
- ε_{ij} = unobserved taste variation (idiosyncratic preferences)

Decision rule: commuter i chooses mode j if $U_{ij} > U_{ik}$ for every other mode k.

$\Longrightarrow$ We never observe utility directly. We observe the *choice*, which reveals which mode had the highest utility for that person.

The Multinomial Logit Probability

If the ε_{ij} follow an i.i.d. Type I Extreme Value (Gumbel) distribution (chosen for mathematical convenience), the choice probability takes a clean form:

The Multinomial Logit Probability

If the ε_{ij} follow an i.i.d. Type I Extreme Value (Gumbel) distribution (chosen for mathematical convenience), the choice probability takes a clean form:

$$P(y_i = j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}$$

The Multinomial Logit Probability

If the ε_{ij} follow an i.i.d. Type I Extreme Value (Gumbel) distribution (chosen for mathematical convenience), the choice probability takes a clean form:

$$P(y_i = j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}$$

This is the **softmax** function (the same function used in machine learning classifiers). Note:

- Every $P(y_i = j) \in (0, 1)$
- $\sum_{j=1}^{J} P(y_i = j) = 1$ by construction

The Multinomial Logit Probability

If the ε_{ij} follow an i.i.d. Type I Extreme Value (Gumbel) distribution (chosen for mathematical convenience), the choice probability takes a clean form:

$$P(y_i = j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}$$

This is the **softmax** function (the same function used in machine learning classifiers). Note:

- Every $P(y_i = j) \in (0, 1)$
- $\sum_{j=1}^{J} P(y_i = j) = 1$ by construction

$\Longrightarrow$ The softmax forces probabilities to be positive and sum to 1. This is the structural constraint that separate binary logits violate.

What Drives Mode Choice?

We observe two types of variables:

What Drives Mode Choice?

We observe two types of variables:

1. **Individual-specific** (same value across all alternatives):
   - Income, age, household size
   - Vary across *people*, not across modes

What Drives Mode Choice?

We observe two types of variables:

1. **Individual-specific** (same value across all alternatives):
   - Income, age, household size
   - Vary across *people*, not across modes
2. **Alternative-specific** (different value for each mode):
   - Travel time, travel cost
   - The same person faces different times/costs for Car vs. Bus vs. Bike

What Drives Mode Choice?

We observe two types of variables:

1. **Individual-specific** (same value across all alternatives):
   - Income, age, household size
   - Vary across *people*, not across modes
2. **Alternative-specific** (different value for each mode):
   - Travel time, travel cost
   - The same person faces different times/costs for Car vs. Bus vs. Bike

Higher income $\implies$ more likely to drive. Longer bus time $\implies$ less likely to take the bus.

MNL with Individual-Specific Variables

When the regressors vary across *people* but not across alternatives (e.g., income), the systematic utility is:

$$V_{ij} = \alpha_j + \beta_j x_i$$

where x_i is a person-level variable (e.g., income in \$10k).

MNL with Individual-Specific Variables

When the regressors vary across *people* but not across alternatives (e.g., income), the systematic utility is:

$$V_{ij} = \alpha_j + \beta_j x_i$$

where x_i is a person-level variable (e.g., income in \$10k).

Each alternative gets its **own intercept** α_j and its **own slope** β_j.

MNL with Individual-Specific Variables

When the regressors vary across *people* but not across alternatives (e.g., income), the systematic utility is:

$$V_{ij} = \alpha_j + \beta_j x_i$$

where x_i is a person-level variable (e.g., income in \$10k).

Each alternative gets its **own intercept** α_j and its **own slope** β_j.

- α_j = baseline appeal of mode j ("alternative-specific constant")
- β_j = how income shifts the probability of choosing mode j

MNL with Individual-Specific Variables

When the regressors vary across *people* but not across alternatives (e.g., income), the systematic utility is:

$$V_{ij} = \alpha_j + \beta_j x_i$$

where x_i is a person-level variable (e.g., income in \$10k).

Each alternative gets its **own intercept** α_j and its **own slope** β_j.

- α_j = baseline appeal of mode j ("alternative-specific constant")
- β_j = how income shifts the probability of choosing mode j

$\Longrightarrow$ The same variable (income) has a *different coefficient for each alternative*. This is what makes the probability curves fan out differently.

Numeric Example: Computing Utilities

Suppose Income = \$50k, so $x_i = 5$ (in \$10k units). With Bus as the base ($\alpha_{\text{Bus}} = 0$, $\beta_{\text{Bus}} = 0$):

Numeric Example: Computing Utilities

Suppose Income = \$50k, so $x_i = 5$ (in \$10k units). With Bus as the base ($\alpha_{\text{Bus}} = 0$, $\beta_{\text{Bus}} = 0$):

$$
\begin{aligned}
V_{\text{Car}} &= -0.3 + 0.15 \times 5 = 0.45 \\
V_{\text{Bus}} &= 0 + 0 \times 5 = 0 \quad \text{(base)} \\
V_{\text{Bike}} &= -0.6 + (-0.03) \times 5 = -0.75 \\
V_{\text{Walk}} &= -1.1 + (-0.08) \times 5 = -1.50
\end{aligned}
$$

Numeric Example: Computing Utilities

Suppose Income = \$50k, so $x_i = 5$ (in \$10k units). With Bus as the base ($\alpha_{\text{Bus}} = 0$, $\beta_{\text{Bus}} = 0$):

$$
\begin{aligned}
V_{\text{Car}} &= -0.3 + 0.15 \times 5 = 0.45 \\
V_{\text{Bus}} &= 0 + 0 \times 5 = 0 \quad \text{(base)} \\
V_{\text{Bike}} &= -0.6 + (-0.03) \times 5 = -0.75 \\
V_{\text{Walk}} &= -1.1 + (-0.08) \times 5 = -1.50
\end{aligned}
$$

Apply the softmax: $e^{0.45} + e^{0} + e^{-0.75} + e^{-1.50} = 3.26$

Numeric Example: Computing Utilities

Suppose Income = \$50k, so $x_i = 5$ (in \$10k units). With Bus as the base ($\alpha_{\text{Bus}} = 0$, $\beta_{\text{Bus}} = 0$):

$$
\begin{aligned}
V_{\text{Car}} &= -0.3 + 0.15 \times 5 = 0.45 \\
V_{\text{Bus}} &= 0 + 0 \times 5 = 0 \quad \text{(base)} \\
V_{\text{Bike}} &= -0.6 + (-0.03) \times 5 = -0.75 \\
V_{\text{Walk}} &= -1.1 + (-0.08) \times 5 = -1.50
\end{aligned}
$$

Apply the softmax: $e^{0.45} + e^{0} + e^{-0.75} + e^{-1.50} = 3.26$

$$
P(\text{Car}) = \frac{e^{0.45}}{3.26} = 0.48, \quad P(\text{Bus}) = 0.31, \quad P(\text{Bike}) = 0.14, \quad P(\text{Walk}) = 0.07
$$

Numeric Example: Computing Utilities

Suppose Income = \$50k, so $x_i = 5$ (in \$10k units). With Bus as the base ($\alpha_{\text{Bus}} = 0$, $\beta_{\text{Bus}} = 0$):

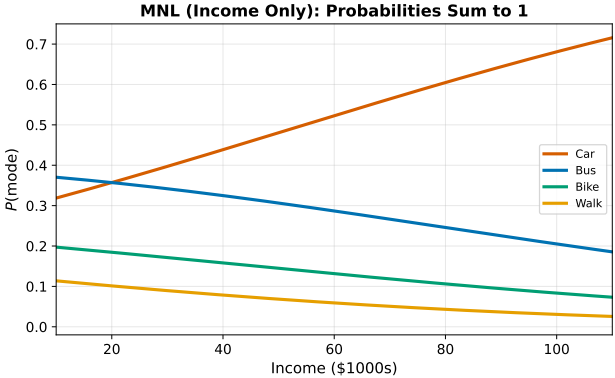
$$
\begin{aligned}
V_{\text{Car}} &= -0.3 + 0.15 \times 5 = 0.45 \\
V_{\text{Bus}} &= 0 + 0 \times 5 = 0 \quad \text{(base)} \\
V_{\text{Bike}} &= -0.6 + (-0.03) \times 5 = -0.75 \\
V_{\text{Walk}} &= -1.1 + (-0.08) \times 5 = -1.50
\end{aligned}
$$

Apply the softmax: $e^{0.45} + e^{0} + e^{-0.75} + e^{-1.50} = 3.26$

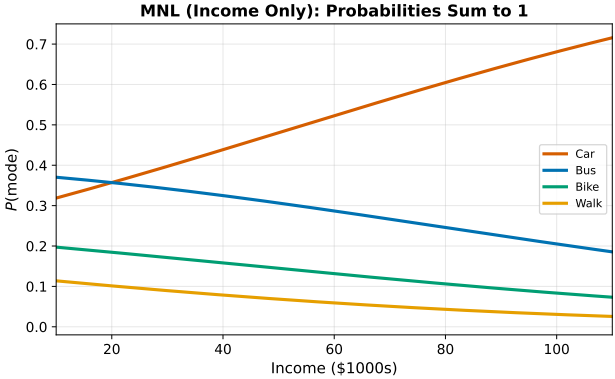
$$
P(\text{Car}) = \frac{e^{0.45}}{3.26} = 0.48, \quad P(\text{Bus}) = 0.31, \quad P(\text{Bike}) = 0.14, \quad P(\text{Walk}) = 0.07
$$

$\implies$ Car has a lower baseline than Bus ($\alpha_{\text{Car}} < 0$), but the positive income effect ($\beta_{\text{Car}} = 0.15$) makes Car most likely at this income level.

MNL Probabilities: Visualized



MNL Probabilities: Visualized



As income rises, Car probability increases while Bike and Walk decline. At every income level, the four curves sum to 1.

Normalization: Choosing a Base Category

Absolute utility levels are not identified; only **differences** are identified. Here is why:

Normalization: Choosing a Base Category

Absolute utility levels are not identified; only **differences** are identified. Here is why:

If we add the same constant c to every V_{ij}:

$$\frac{e^{V_{ij}+c}}{\sum_k e^{V_{ik}+c}} = \frac{e^c \cdot e^{V_{ij}}}{e^c \cdot \sum_k e^{V_{ik}}} = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

Normalization: Choosing a Base Category

Absolute utility levels are not identified; only **differences** are identified. Here is why:

If we add the same constant c to every V_{ij}:

$$\frac{e^{V_{ij}+c}}{\sum_k e^{V_{ik}+c}} = \frac{e^c \cdot e^{V_{ij}}}{e^c \cdot \sum_k e^{V_{ik}}} = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

$\implies$ The probabilities are unchanged. We cannot separately identify α_{Car} and α_{Bus}, only their difference.

Normalization: Choosing a Base Category

Absolute utility levels are not identified; only **differences** are identified. Here is why:

If we add the same constant c to every V_{ij}:

$$\frac{e^{V_{ij}+c}}{\sum_k e^{V_{ik}+c}} = \frac{e^c \cdot e^{V_{ij}}}{e^c \cdot \sum_k e^{V_{ik}}} = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

$\Longrightarrow$ The probabilities are unchanged. We cannot separately identify α_{Car} and α_{Bus}, only their difference.

Fix: set one alternative as the **base category** (e.g., Bus) and normalize:

$$\alpha_{\text{Bus}} = 0, \quad \beta_{\text{Bus}} = 0$$

Normalization: Choosing a Base Category

Absolute utility levels are not identified; only **differences** are identified. Here is why:

If we add the same constant c to every V_{ij}:

$$\frac{e^{V_{ij}+c}}{\sum_k e^{V_{ik}+c}} = \frac{e^c \cdot e^{V_{ij}}}{e^c \cdot \sum_k e^{V_{ik}}} = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

$\Longrightarrow$ The probabilities are unchanged. We cannot separately identify α_{Car} and α_{Bus}, only their difference.

Fix: set one alternative as the **base category** (e.g., Bus) and normalize:

$$\alpha_{\text{Bus}} = 0, \quad \beta_{\text{Bus}} = 0$$

All estimated coefficients are then interpreted *relative to Bus*:

- $\beta_{\text{Car}} > 0$: higher income increases Car utility *relative to Bus*

Interpreting Coefficients

With Bus as the base category, the log-odds ratio is:

$$\ln\left(\frac{P(y_i = j)}{P(y_i = \text{Bus})}\right) = \alpha_j + \beta_j x_i$$

Interpreting Coefficients

With Bus as the base category, the log-odds ratio is:

$$\ln\left(\frac{P(y_i = j)}{P(y_i = \text{Bus})}\right) = \alpha_j + \beta_j x_i$$

(This follows directly from dividing the softmax probabilities for mode j and Bus.)

Interpreting Coefficients

With Bus as the base category, the log-odds ratio is:

$$\ln\left(\frac{P(y_i = j)}{P(y_i = \text{Bus})}\right) = \alpha_j + \beta_j x_i$$

(This follows directly from dividing the softmax probabilities for mode j and Bus.)

$\implies$ β_j measures how income changes the **log-odds of mode j relative to Bus**.

Interpreting Coefficients

With Bus as the base category, the log-odds ratio is:

$$\ln\left(\frac{P(y_i = j)}{P(y_i = \text{Bus})}\right) = \alpha_j + \beta_j x_i$$

(This follows directly from dividing the softmax probabilities for mode j and Bus.)

$\implies$ β_j measures how income changes the **log-odds of mode j relative to Bus**.

Mode	$\hat{\alpha}_j$	$\hat{\beta}_j$	**Interpretation**
Bus (base)	0	0	(normalized)
Car	$-$	$+$	Lower baseline than Bus, but income pulls toward Car
Bike	$-$	$-$	Lower baseline, income pushes further away
Walk	$-$	$-$	Lowest baseline, income pushes further away

Interpreting Coefficients

With Bus as the base category, the log-odds ratio is:

$$\ln\left(\frac{P(y_i = j)}{P(y_i = \text{Bus})}\right) = \alpha_j + \beta_j x_i$$

(This follows directly from dividing the softmax probabilities for mode j and Bus.)

$\implies$ β_j measures how income changes the **log-odds of mode j relative to Bus**.

Mode	$\hat{\alpha}_j$	$\hat{\beta}_j$	**Interpretation**
Bus (base)	0	0	(normalized)
Car	$-$	$+$	Lower baseline than Bus, but income pulls toward Car
Bike	$-$	$-$	Lower baseline, income pushes further away
Walk	$-$	$-$	Lowest baseline, income pushes further away

The sign of β_j tells us the direction *relative to Bus*. It does not directly give the marginal effect on probability (same nonlinearity issue as binary logit).

Marginal Effects in the Multinomial Logit

For an individual-specific variable x_i, the marginal effect on $P(y_i = j)$ is:

$$\frac{\partial P(y_i = j)}{\partial x_i} = P(y_i = j) \left[\beta_j - \sum_{k=1}^{J} P(y_i = k) \beta_k \right]$$

Marginal Effects in the Multinomial Logit

For an individual-specific variable x_i, the marginal effect on $P(y_i = j)$ is:

$$\frac{\partial P(y_i = j)}{\partial x_i} = P(y_i = j) \left[\beta_j - \sum_{k=1}^{J} P(y_i = k) \beta_k \right]$$

This looks complicated, but the intuition is simple:

- A variable can increase $P(j)$ even if $\beta_j = 0$, as long as it *decreases* the probability of other alternatives

Marginal Effects in the Multinomial Logit

For an individual-specific variable x_i, the marginal effect on $P(y_i = j)$ is:

$$\frac{\partial P(y_i = j)}{\partial x_i} = P(y_i = j) \left[\beta_j - \sum_{k=1}^{J} P(y_i = k) \beta_k \right]$$

This looks complicated, but the intuition is simple:

- A variable can increase $P(j)$ even if $\beta_j = 0$, as long as it *decreases* the probability of other alternatives
- The effect depends on *all* alternatives' probabilities, not just mode j

Marginal Effects in the Multinomial Logit

For an individual-specific variable x_i, the marginal effect on $P(y_i = j)$ is:

$$\frac{\partial P(y_i = j)}{\partial x_i} = P(y_i = j) \left[\beta_j - \sum_{k=1}^{J} P(y_i = k)\, \beta_k \right]$$

This looks complicated, but the intuition is simple:

- A variable can increase $P(j)$ even if $\beta_j = 0$, as long as it *decreases* the probability of other alternatives
- The effect depends on *all* alternatives' probabilities, not just mode j

Example: suppose $\beta_{\text{Car}} = 0.5$, $\beta_{\text{Bike}} = 0$, $\beta_{\text{Walk}} = -0.3$, and income rises by \$1k. Even though $\beta_{\text{Bike}} = 0$, income may *increase* $P(\text{Bike})$ if the probability-weighted average $\bar{\beta} = \sum_k P_k \beta_k$ is negative. In practice, $\bar{\beta}$ is usually dominated by the positive Car coefficient, so $P(\text{Bike})$ falls.

Marginal Effects in the Multinomial Logit

For an individual-specific variable x_i, the marginal effect on $P(y_i = j)$ is:

$$\frac{\partial P(y_i = j)}{\partial x_i} = P(y_i = j) \left[\beta_j - \sum_{k=1}^{J} P(y_i = k)\, \beta_k \right]$$

This looks complicated, but the intuition is simple:

- A variable can increase $P(j)$ even if $\beta_j = 0$, as long as it *decreases* the probability of other alternatives
- The effect depends on *all* alternatives' probabilities, not just mode j

Example: suppose $\beta_{\text{Car}} = 0.5$, $\beta_{\text{Bike}} = 0$, $\beta_{\text{Walk}} = -0.3$, and income rises by \$1k. Even though $\beta_{\text{Bike}} = 0$, income may *increase* $P(\text{Bike})$ if the probability-weighted average $\bar{\beta} = \sum_k P_k \beta_k$ is negative. In practice, $\bar{\beta}$ is usually dominated by the positive Car coefficient, so $P(\text{Bike})$ falls.

$\Longrightarrow$ Report Average Marginal Effects (AMEs) for each alternative, just as in binary logit.

Estimation: Maximum Likelihood

Define an indicator $d_{ij} = 1$ if person i chose mode j, and 0 otherwise.

Estimation: Maximum Likelihood

Define an indicator $d_{ij} = 1$ if person i chose mode j, and 0 otherwise.

Each person contributes one term to the likelihood: the probability of the mode they actually chose.

Estimation: Maximum Likelihood

Define an indicator $d_{ij} = 1$ if person i chose mode j, and 0 otherwise.

Each person contributes one term to the likelihood: the probability of the mode they actually chose.

The log-likelihood:

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ij} \ln P(y_i = j)$$

Estimation: Maximum Likelihood

Define an indicator $d_{ij} = 1$ if person i chose mode j, and 0 otherwise.

Each person contributes one term to the likelihood: the probability of the mode they actually chose.

The log-likelihood:

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ij} \ln P(y_i = j)$$

where $P(y_i = j) = \dfrac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$.

Estimation: Maximum Likelihood

Define an indicator $d_{ij} = 1$ if person i chose mode j, and 0 otherwise.

Each person contributes one term to the likelihood: the probability of the mode they actually chose.

The log-likelihood:

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ij} \ln P(y_i = j)$$

where $P(y_i = j) = \dfrac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$.

No closed-form solution $\implies$ solved numerically, just like binary logit. Software handles this automatically.

Outline

The Limitation Multinomial Logit Cannot Overcome

MNL models **who** chooses what. But it ignores the attributes of the alternatives themselves.

The Limitation Multinomial Logit Cannot Overcome

MNL models **who** chooses what. But it ignores the attributes of the alternatives themselves.

Concrete example: two people with identical incomes face different bus routes.

- Person A: bus takes 20 minutes
- Person B: bus takes 50 minutes

The Limitation Multinomial Logit Cannot Overcome

MNL models **who** chooses what. But it ignores the attributes of the alternatives themselves.

Concrete example: two people with identical incomes face different bus routes.

- Person A: bus takes 20 minutes
- Person B: bus takes 50 minutes

MNL (income only) gives them the **same** probability of taking the bus:

$$V_{\text{Bus},A} = \alpha_{\text{Bus}} + \beta_{\text{Bus}} \cdot \text{Income} = V_{\text{Bus},B}$$

The Limitation Multinomial Logit Cannot Overcome

MNL models **who** chooses what. But it ignores the attributes of the alternatives themselves.

Concrete example: two people with identical incomes face different bus routes.

- Person A: bus takes 20 minutes
- Person B: bus takes 50 minutes

MNL (income only) gives them the **same** probability of taking the bus:

$$V_{\text{Bus},A} = \alpha_{\text{Bus}} + \beta_{\text{Bus}} \cdot \text{Income} = V_{\text{Bus},B}$$

Travel time does not appear in the model. Person B, facing a 50-minute bus ride, gets the same $P(\text{Bus})$ as Person A with a 20-minute ride.

The Limitation Multinomial Logit Cannot Overcome

MNL models **who** chooses what. But it ignores the attributes of the alternatives themselves.

Concrete example: two people with identical incomes face different bus routes.

- Person A: bus takes 20 minutes
- Person B: bus takes 50 minutes

MNL (income only) gives them the **same** probability of taking the bus:

$$V_{\text{Bus},A} = \alpha_{\text{Bus}} + \beta_{\text{Bus}} \cdot \text{Income} = V_{\text{Bus},B}$$

Travel time does not appear in the model. Person B, facing a 50-minute bus ride, gets the same $P(\text{Bus})$ as Person A with a 20-minute ride.

$\Longrightarrow$ We need to incorporate **alternative-specific variables**: attributes that vary across both people *and* modes (travel time, travel cost).

Conditional Logit: Alternative-Specific Variables

Conditional logit (McFadden, 1974) enters alternative-specific variables with a **single coefficient** shared across all modes:

$$V_{ij} = \beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}$$

Conditional Logit: Alternative-Specific Variables

Conditional logit (McFadden, 1974) enters alternative-specific variables with a **single coefficient** shared across all modes:

$$V_{ij} = \beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}$$

- Time_{ij} = travel time person i faces for mode j
- β_{time} = the same for all modes

Conditional Logit: Alternative-Specific Variables

Conditional logit (McFadden, 1974) enters alternative-specific variables with a **single coefficient** shared across all modes:

$$V_{ij} = \beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}$$

- Time_{ij} = travel time person i faces for mode j
- β_{time} = the same for all modes

(In practice, we also include alternative-specific constants α_j. Here we focus on the distinctive feature: alternative-varying regressors with shared coefficients.)

Conditional Logit: Alternative-Specific Variables

Conditional logit (McFadden, 1974) enters alternative-specific variables with a **single coefficient** shared across all modes:

$$V_{ij} = \beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}$$

- Time_{ij} = travel time person i faces for mode j
- β_{time} = the same for all modes

(In practice, we also include alternative-specific constants α_j. Here we focus on the distinctive feature: alternative-varying regressors with shared coefficients.)

This is the opposite pattern from multinomial logit:

	Regressors vary by...	**Coefficients...**
Multinomial logit	Individual only	Differ by alternative
Conditional logit	Individual *and* alternative	Same across alternatives

Conditional Logit: Alternative-Specific Variables

Conditional logit (McFadden, 1974) enters alternative-specific variables with a **single coefficient** shared across all modes:

$$V_{ij} = \beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}$$

- Time_{ij} = travel time person i faces for mode j
- β_{time} = the same for all modes

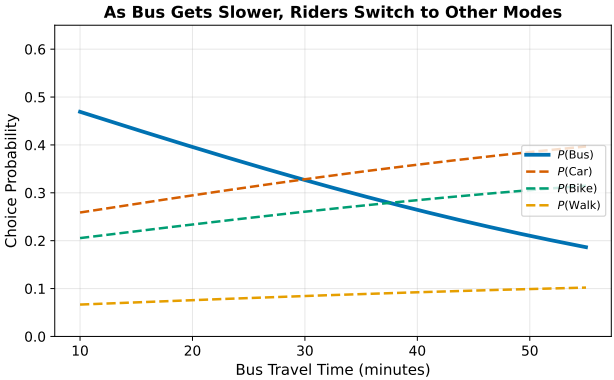
(In practice, we also include alternative-specific constants α_j. Here we focus on the distinctive feature: alternative-varying regressors with shared coefficients.)

This is the opposite pattern from multinomial logit:

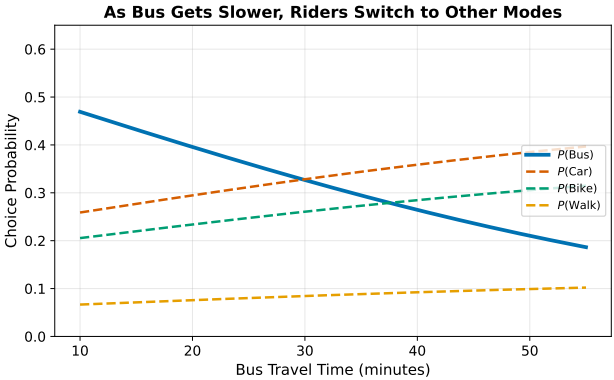
	Regressors vary by...	**Coefficients...**
Multinomial logit	Individual only	Differ by alternative
Conditional logit	Individual *and* alternative	Same across alternatives

$\Longrightarrow$ One extra minute of travel time reduces utility by β_{time}, regardless of whether it is a minute on the bus or a minute in the car.

Conditional Logit: Travel Time Effect



Conditional Logit: Travel Time Effect



As bus travel time increases, $P(\text{Bus})$ falls and the other modes absorb the lost share. The rate of substitution depends on each mode's current probability.

The Full Model: Combining Both Variable Types

In practice, we combine both types of variables in one model. We add a superscript to distinguish income coefficients from the travel-time coefficient:

$$V_{ij} = \underbrace{\alpha_j + \beta_j^{\text{inc}} \cdot \text{Income}_i}_{\text{individual-specific}} + \underbrace{\beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}}_{\text{alternative-specific}}$$

The Full Model: Combining Both Variable Types

In practice, we combine both types of variables in one model. We add a superscript to distinguish income coefficients from the travel-time coefficient:

$$V_{ij} = \underbrace{\alpha_j + \beta_j^{\text{inc}} \cdot \text{Income}_i}_{\text{individual-specific}} + \underbrace{\beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}}_{\text{alternative-specific}}$$

- $\alpha_j, \beta_j^{\text{inc}}$: vary by alternative (one per mode, base category normalized)
- $\beta_{\text{time}}, \beta_{\text{cost}}$: shared across all modes (one coefficient each)

The Full Model: Combining Both Variable Types

In practice, we combine both types of variables in one model. We add a superscript to distinguish income coefficients from the travel-time coefficient:

$$V_{ij} = \underbrace{\alpha_j + \beta_j^{\text{inc}} \cdot \text{Income}_i}_{\text{individual-specific}} + \underbrace{\beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}}_{\text{alternative-specific}}$$

- $\alpha_j, \beta_j^{\text{inc}}$: vary by alternative (one per mode, base category normalized)
- $\beta_{\text{time}}, \beta_{\text{cost}}$: shared across all modes (one coefficient each)

The probability formula is the same softmax:

$$P(y_i = j) = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

The Full Model: Combining Both Variable Types

In practice, we combine both types of variables in one model. We add a superscript to distinguish income coefficients from the travel-time coefficient:

$$V_{ij} = \underbrace{\alpha_j + \beta_j^{\text{inc}} \cdot \text{Income}_i}_{\text{individual-specific}} + \underbrace{\beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}}_{\text{alternative-specific}}$$

- $\alpha_j, \beta_j^{\text{inc}}$: vary by alternative (one per mode, base category normalized)
- $\beta_{\text{time}}, \beta_{\text{cost}}$: shared across all modes (one coefficient each)

The probability formula is the same softmax:

$$P(y_i = j) = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

$\Longrightarrow$ This combined model is often called the "conditional logit" or "McFadden's choice model" in applied work, though terminology varies across textbooks. It handles both types of variation simultaneously.

Model Comparison: What Each Specification Handles

	MNL	**CL**	**Combined**
Individual-specific variables (income)	✓		✓
Alternative-specific variables (time, cost)		✓	✓
Alternative-specific constants (α_j)	✓		✓
Alternative-varying slopes (β_j)	✓		✓
Shared slopes (β)		✓	✓

Model Comparison: What Each Specification Handles

	MNL	CL	Combined
Individual-specific variables (income)	✓		✓
Alternative-specific variables (time, cost)		✓	✓
Alternative-specific constants (α_j)	✓		✓
Alternative-varying slopes (β_j)	✓		✓
Shared slopes (β)		✓	✓

$\Longrightarrow$ The combined model subsumes both. It reduces to pure MNL when there are no alternative-specific variables, and to pure CL when there are no individual-specific variables.

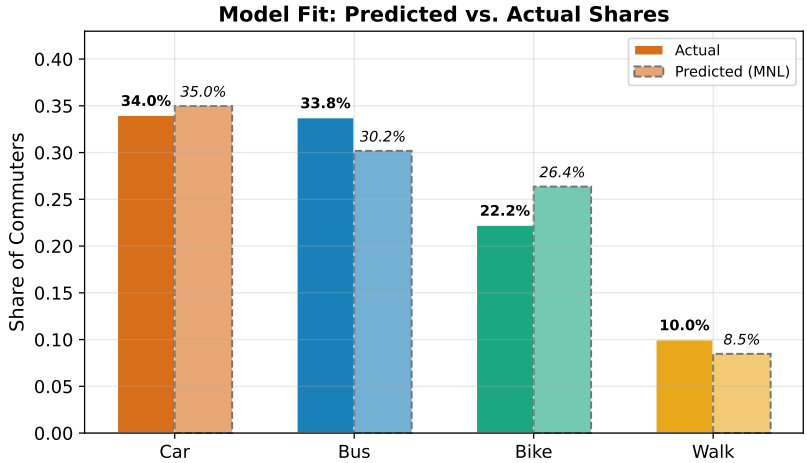
Model Comparison: What Each Specification Handles

	MNL	CL	Combined
Individual-specific variables (income)	✓		✓
Alternative-specific variables (time, cost)		✓	✓
Alternative-specific constants (α_j)	✓		✓
Alternative-varying slopes (β_j)	✓		✓
Shared slopes (β)		✓	✓

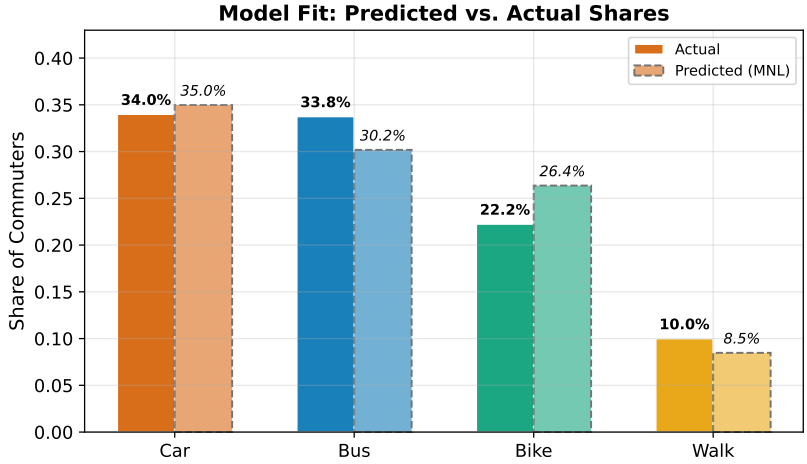
$\Longrightarrow$ The combined model subsumes both. It reduces to pure MNL when there are no alternative-specific variables, and to pure CL when there are no individual-specific variables.

All three use the same softmax probability and the same MLE estimator. The only difference is what goes into V_{ij}.

Model Fit: Predicted vs. Actual



Model Fit: Predicted vs. Actual



The multinomial/conditional logit closely matches the observed mode shares. The model aggregates well even though individual predictions are probabilistic.

Outline

The IIA Assumption

The multinomial logit probability has a special structure. Take the ratio of any two alternatives:

$$\frac{P(y_i = j)}{P(y_i = k)} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = e^{V_{ij} - V_{ik}}$$

The IIA Assumption

The multinomial logit probability has a special structure. Take the ratio of any two alternatives:

$$\frac{P(y_i = j)}{P(y_i = k)} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = e^{V_{ij} - V_{ik}}$$

This ratio depends **only on** j **and** k. Adding or removing a third alternative does not change it.

The IIA Assumption

The multinomial logit probability has a special structure. Take the ratio of any two alternatives:

$$\frac{P(y_i = j)}{P(y_i = k)} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = e^{V_{ij} - V_{ik}}$$

This ratio depends **only on** j **and** k. Adding or removing a third alternative does not change it.

$\implies$ **Independence of Irrelevant Alternatives (IIA):** the relative odds between any two modes are unaffected by what other modes exist.

The IIA Assumption

The multinomial logit probability has a special structure. Take the ratio of any two alternatives:

$$\frac{P(y_i = j)}{P(y_i = k)} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = e^{V_{ij} - V_{ik}}$$

This ratio depends **only on** j **and** k. Adding or removing a third alternative does not change it.

$\Longrightarrow$ **Independence of Irrelevant Alternatives (IIA):** the relative odds between any two modes are unaffected by what other modes exist.

This is both a strength (clean, tractable) and a weakness (unrealistic in some settings). Note: this property holds given fixed parameters, which is the standard presentation of IIA.

The Red Bus / Blue Bus Problem

A classic thought experiment exposes when IIA fails.

The Red Bus / Blue Bus Problem

A classic thought experiment exposes when IIA fails.

Before: two options with equal market shares:

- Car: 50% Bus: 50%

The Red Bus / Blue Bus Problem

A classic thought experiment exposes when IIA fails.

Before: two options with equal market shares:

- Car: 50% Bus: 50%

Now the city introduces a **Red Bus** that is identical to the existing (Blue) Bus. Under IIA, the multinomial logit predicts:

The Red Bus / Blue Bus Problem

A classic thought experiment exposes when IIA fails.

Before: two options with equal market shares:

- Car: 50% Bus: 50%

Now the city introduces a **Red Bus** that is identical to the existing (Blue) Bus. Under IIA, the multinomial logit predicts:

- Car: 33%, Red Bus: 33%, Blue Bus: 33%

The Red Bus / Blue Bus Problem

A classic thought experiment exposes when IIA fails.

Before: two options with equal market shares:

- Car: 50% Bus: 50%

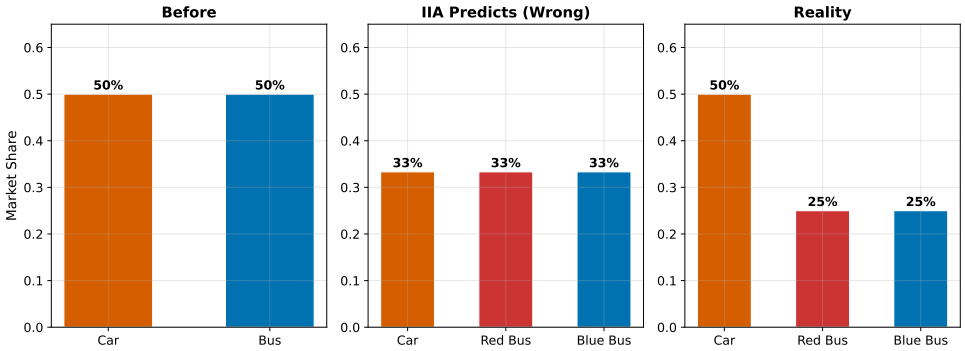
Now the city introduces a **Red Bus** that is identical to the existing (Blue) Bus. Under IIA, the multinomial logit predicts:

- Car: 33%, Red Bus: 33%, Blue Bus: 33%

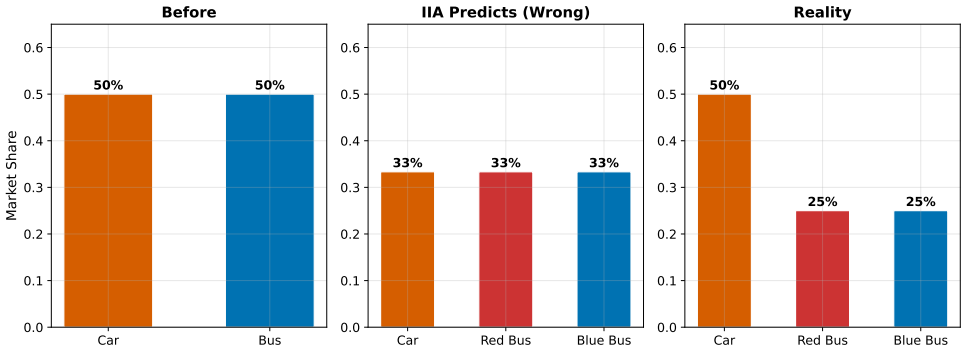
But common sense says bus riders split between Red and Blue, while car drivers are unaffected:

- Car: 50%, Red Bus: 25%, Blue Bus: 25%

Red Bus / Blue Bus: Visualized



Red Bus / Blue Bus: Visualized



IIA says the relative odds of Car to Blue Bus stay the same after Red Bus enters. In reality, Red Bus steals from Blue Bus (a close substitute), not proportionally from all modes.

When Does IIA Fail?

IIA is problematic when alternatives are **close substitutes** for some people but not for others.

When Does IIA Fail?

IIA is problematic when alternatives are **close substitutes** for some people but not for others.

IIA fails when:

- Two alternatives share unobserved attributes (Red Bus $\approx$ Blue Bus)
- The error terms ε_{ij} are **correlated** across alternatives

When Does IIA Fail?

IIA is problematic when alternatives are **close substitutes** for some people but not for others.

IIA fails when:

- Two alternatives share unobserved attributes (Red Bus $\approx$ Blue Bus)
- The error terms ε_{ij} are **correlated** across alternatives

IIA is reasonable when:

- Alternatives are genuinely distinct (car, bus, bike, walk are quite different)
- The observed variables capture the similarities between alternatives

When Does IIA Fail?

IIA is problematic when alternatives are **close substitutes** for some people but not for others.

IIA fails when:

- Two alternatives share unobserved attributes (Red Bus $\approx$ Blue Bus)
- The error terms ε_{ij} are **correlated** across alternatives

IIA is reasonable when:

- Alternatives are genuinely distinct (car, bus, bike, walk are quite different)
- The observed variables capture the similarities between alternatives

Hausman test for IIA: estimate the model on a subset of alternatives. If IIA holds, the coefficients should not change significantly when you drop one alternative.

Beyond Multinomial Logit: Relaxing IIA

You don't need to estimate these models for this course, but you should know they exist.

Beyond Multinomial Logit: Relaxing IIA

You don't need to estimate these models for this course, but you should know they exist.

Model	**How it relaxes IIA**
Nested logit	Groups similar alternatives into "nests" (e.g., Motorized: {Car, Bus} vs. Non-motorized: {Bike, Walk}). Allows correlation within nests
Mixed logit	Coefficients vary randomly across individuals. Generates flexible substitution patterns
Multinomial probit	Normal errors with a full covariance structure. Most general, but computationally expensive

Beyond Multinomial Logit: Relaxing IIA

You don't need to estimate these models for this course, but you should know they exist.

Model	**How it relaxes IIA**
Nested logit	Groups similar alternatives into "nests" (e.g., Motorized: {Car, Bus} vs. Non-motorized: {Bike, Walk}). Allows correlation within nests
Mixed logit	Coefficients vary randomly across individuals. Generates flexible substitution patterns
Multinomial probit	Normal errors with a full covariance structure. Most general, but computationally expensive

$\Longrightarrow$ Start with multinomial logit and test IIA. If it fails, move to nested or mixed logit.

Decision Framework: Which Model to Use

1. **Only individual-specific variables?** (income, age, etc.)
   $\Longrightarrow$ Multinomial logit (MNL)

Decision Framework: Which Model to Use

1. **Only individual-specific variables?** (income, age, etc.)
   $\Longrightarrow$ Multinomial logit (MNL)
2. **Only alternative-specific variables?** (time, cost, etc.)
   $\Longrightarrow$ Conditional logit (CL)

Decision Framework: Which Model to Use

1. **Only individual-specific variables?** (income, age, etc.)
   $\Longrightarrow$ Multinomial logit (MNL)
2. **Only alternative-specific variables?** (time, cost, etc.)
   $\Longrightarrow$ Conditional logit (CL)
3. **Both types of variables?**
   $\Longrightarrow$ Combined model (the usual case in practice)

Decision Framework: Which Model to Use

1. **Only individual-specific variables?** (income, age, etc.)
   $\Longrightarrow$ Multinomial logit (MNL)
2. **Only alternative-specific variables?** (time, cost, etc.)
   $\Longrightarrow$ Conditional logit (CL)
3. **Both types of variables?**
   $\Longrightarrow$ Combined model (the usual case in practice)
4. **Alternatives are close substitutes?** (Red Bus / Blue Bus concern)
   $\Longrightarrow$ Test IIA. If it fails, consider nested logit or mixed logit

Decision Framework: Which Model to Use

1. **Only individual-specific variables?** (income, age, etc.)
   $\Longrightarrow$ Multinomial logit (MNL)
2. **Only alternative-specific variables?** (time, cost, etc.)
   $\Longrightarrow$ Conditional logit (CL)
3. **Both types of variables?**
   $\Longrightarrow$ Combined model (the usual case in practice)
4. **Alternatives are close substitutes?** (Red Bus / Blue Bus concern)
   $\Longrightarrow$ Test IIA. If it fails, consider nested logit or mixed logit

$\Longrightarrow$ In most applied work, start with the combined model. It nests the other two as special cases.

Thank you!
jakeanderson@g.ucla.edu

Count Data Models

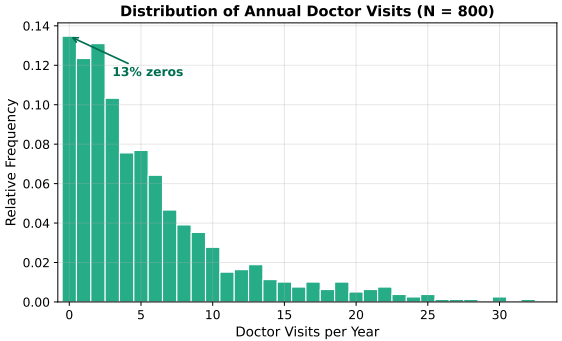
Jake Anderson

May 16, 2026

Outline

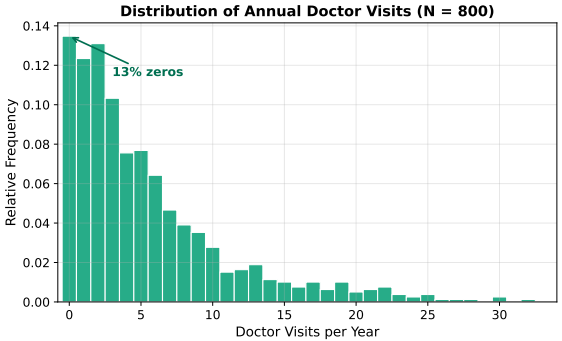
The Data

A health economist surveys **800 individuals** and records their **annual doctor visits**. Covariates include age, insurance status, and a health index (centered near 0; higher = healthier).



The Data

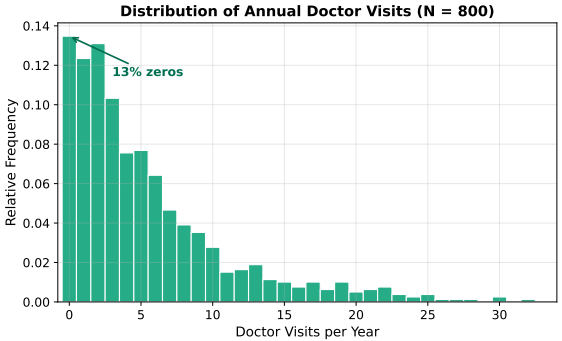
A health economist surveys **800 individuals** and records their **annual doctor visits**. Covariates include age, insurance status, and a health index (centered near 0; higher = healthier).



What do you notice about this distribution?

The Data

A health economist surveys **800 individuals** and records their **annual doctor visits**. Covariates include age, insurance status, and a health index (centered near 0; higher = healthier).



What do you notice about this distribution?

The outcome is a **count**: non-negative integers (0, 1, 2, . . .). Right-skewed with a spike at zero. Mean = 5.7, but 13% have zero visits.

OLS Predictions on Count Data

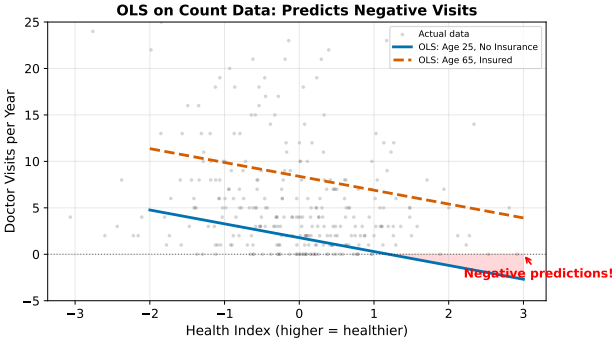
Treat doctor visits as a continuous variable and regress on covariates:

$$\text{Visits}_i = \beta_0 + \beta_1 \, \text{Age}_i + \beta_2 \, \text{Insurance}_i + \beta_3 \, \text{Health}_i + \varepsilon_i$$

OLS Predictions on Count Data

Treat doctor visits as a continuous variable and regress on covariates:

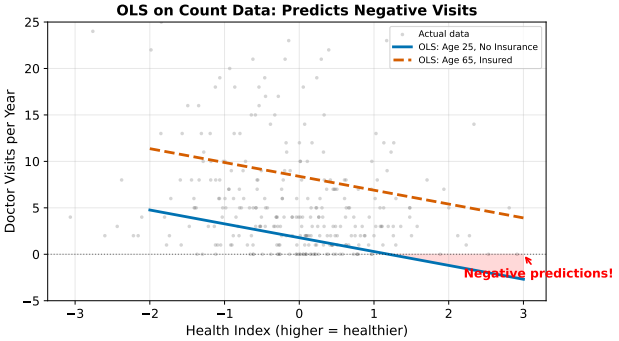
$$\text{Visits}_i = \beta_0 + \beta_1 \, \text{Age}_i + \beta_2 \, \text{Insurance}_i + \beta_3 \, \text{Health}_i + \varepsilon_i$$



OLS Predictions on Count Data

Treat doctor visits as a continuous variable and regress on covariates:

$$\text{Visits}_i = \beta_0 + \beta_1 \, \text{Age}_i + \beta_2 \, \text{Insurance}_i + \beta_3 \, \text{Health}_i + \varepsilon_i$$



For a 25-year-old without insurance, OLS predicts **negative visits** once the health index exceeds about 1.5. Doctor visits cannot be negative.

Three Failures of OLS on Counts

The plot reveals the first problem, but there are two more:

Three Failures of OLS on Counts

The plot reveals the first problem, but there are two more:

1. **Negative predictions.** OLS can predict -2.4 visits for a young, healthy, uninsured person

Three Failures of OLS on Counts

The plot reveals the first problem, but there are two more:

1. **Negative predictions.** OLS can predict -2.4 visits for a young, healthy, uninsured person
2. **Non-constant variance.** People who average 10 visits have far more variation than those who average 1

Three Failures of OLS on Counts

The plot reveals the first problem, but there are two more:

1. **Negative predictions.** OLS can predict -2.4 visits for a young, healthy, uninsured person
2. **Non-constant variance.** People who average 10 visits have far more variation than those who average 1
3. **Non-normal residuals.** Count data is right-skewed and discrete; OLS assumes symmetric, continuous errors

Three Failures of OLS on Counts

The plot reveals the first problem, but there are two more:

1. **Negative predictions.** OLS can predict -2.4 visits for a young, healthy, uninsured person
2. **Non-constant variance.** People who average 10 visits have far more variation than those who average 1
3. **Non-normal residuals.** Count data is right-skewed and discrete; OLS assumes symmetric, continuous errors

$\Longrightarrow$ We need a model built for count outcomes from the start.

What Would a Better Model Need?

A model for count outcomes should:

What Would a Better Model Need?

A model for count outcomes should:

1. **Guarantee non-negative predictions.** $\hat{y}_i \geq 0$ for all covariate values

What Would a Better Model Need?

A model for count outcomes should:

1. **Guarantee non-negative predictions.** $\hat{y}_i \geq 0$ for all covariate values
2. **Handle the variance-mean relationship.** Individuals with higher expected visits naturally have more spread

What Would a Better Model Need?

A model for count outcomes should:

1. **Guarantee non-negative predictions.** $\hat{y}_i \geq 0$ for all covariate values
2. **Handle the variance-mean relationship.** Individuals with higher expected visits naturally have more spread
3. **Accommodate the spike at zero.** Many people never visit the doctor; the model should not be surprised by this

What Would a Better Model Need?

A model for count outcomes should:

1. **Guarantee non-negative predictions.** $\hat{y}_i \geq 0$ for all covariate values
2. **Handle the variance-mean relationship.** Individuals with higher expected visits naturally have more spread
3. **Accommodate the spike at zero.** Many people never visit the doctor; the model should not be surprised by this

$\Longrightarrow$ Where can we find a probability distribution designed for non-negative integers?

The Poisson Distribution for Counts

You already know the binary case: we replaced OLS with logit/probit to keep predictions in $[0, 1]$.

The Poisson Distribution for Counts

You already know the binary case: we replaced OLS with logit/probit to keep predictions in $[0, 1]$.

Same logic here: we need a **distribution for counts** to replace the normal distribution.

The Poisson Distribution for Counts

You already know the binary case: we replaced OLS with logit/probit to keep predictions in $[0, 1]$.

Same logic here: we need a **distribution for counts** to replace the normal distribution.

The simplest count distribution is the **Poisson**: it assigns probabilities to 0, 1, 2, 3, ... and has one parameter that controls both the mean and the variance.

The Poisson Distribution for Counts

You already know the binary case: we replaced OLS with logit/probit to keep predictions in $[0, 1]$.

Same logic here: we need a **distribution for counts** to replace the normal distribution.

The simplest count distribution is the **Poisson**: it assigns probabilities to 0, 1, 2, 3, ... and has one parameter that controls both the mean and the variance.

$\Longrightarrow$ Let's build a regression model on top of the Poisson distribution, just as logit builds on the logistic distribution.

Outline

The Poisson Distribution

A random variable Y follows a Poisson distribution with parameter $\mu > 0$ if:

$$P(Y = k) = \frac{e^{-\mu}\,\mu^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

The Poisson Distribution

A random variable Y follows a Poisson distribution with parameter $\mu > 0$ if:

$$P(Y = k) = \frac{e^{-\mu}\,\mu^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

Properties:

- $\mathrm{E}[Y] = \mu$
- $\mathrm{Var}(Y) = \mu \quad \Longrightarrow$ **equidispersion**: the variance equals the mean
- As μ increases, the distribution shifts right and spreads out

The Poisson Distribution

A random variable Y follows a Poisson distribution with parameter $\mu > 0$ if:

$$P(Y = k) = \frac{e^{-\mu}\,\mu^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

Properties:

- $\mathrm{E}[Y] = \mu$
- $\mathrm{Var}(Y) = \mu \quad \Longrightarrow$ **equidispersion**: the variance equals the mean
- As μ increases, the distribution shifts right and spreads out

Example: if $\mu = 6$, then $P(Y = 0) = e^{-6} \approx 0.0025$ and $P(Y = 6) \approx 0.16$.

Poisson Regression: The Log Link

To build a regression, we let the Poisson parameter μ_i depend on covariates. But $\mu_i > 0$, so we need to keep predictions positive.

Poisson Regression: The Log Link

To build a regression, we let the Poisson parameter μ_i depend on covariates. But $\mu_i > 0$, so we need to keep predictions positive.

The log link: model the log of the conditional mean as a linear function:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

Poisson Regression: The Log Link

To build a regression, we let the Poisson parameter μ_i depend on covariates. But $\mu_i > 0$, so we need to keep predictions positive.

The log link: model the log of the conditional mean as a linear function:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

Equivalently:

$$\mu_i = \text{E}[\text{Visits}_i \mid \text{covariates}] = e^{\beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i}$$

Poisson Regression: The Log Link

To build a regression, we let the Poisson parameter μ_i depend on covariates. But $\mu_i > 0$, so we need to keep predictions positive.

The log link: model the log of the conditional mean as a linear function:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

Equivalently:

$$\mu_i = \text{E}[\text{Visits}_i \mid \text{covariates}] = e^{\beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i}$$

Since $e^{(\cdot)} > 0$ for any input, **predicted counts are always positive**. This solves the negative-prediction problem.

Estimation: Maximum Likelihood

Poisson regression is estimated by maximizing the log-likelihood:

$$\ell = \sum_{i=1}^{N} \Big[y_i \ln(\mu_i) - \mu_i - \ln(y_i!) \Big]$$

Estimation: Maximum Likelihood

Poisson regression is estimated by maximizing the log-likelihood:

$$\ell = \sum_{i=1}^{N} \Big[y_i \ln(\mu_i) - \mu_i - \ln(y_i!) \Big]$$

where $\mu_i = e^{\beta_0 + \beta_1 \text{ Age}_i + \beta_2 \text{ Insurance}_i + \beta_3 \text{ Health}_i}$.

Estimation: Maximum Likelihood

Poisson regression is estimated by maximizing the log-likelihood:

$$\ell = \sum_{i=1}^{N} \Big[y_i \ln(\mu_i) - \mu_i - \ln(y_i!) \Big]$$

where $\mu_i = e^{\beta_0 + \beta_1 \text{ Age}_i + \beta_2 \text{ Insurance}_i + \beta_3 \text{ Health}_i}$.

No closed-form solution $\implies$ solved numerically (same as logit). Software reports coefficients, standard errors, and predicted counts $\hat{\mu}_i$.

Estimation: Maximum Likelihood

Poisson regression is estimated by maximizing the log-likelihood:

$$\ell = \sum_{i=1}^{N} \Big[y_i \ln(\mu_i) - \mu_i - \ln(y_i!) \Big]$$

where $\mu_i = e^{\beta_0 + \beta_1 \text{ Age}_i + \beta_2 \text{ Insurance}_i + \beta_3 \text{ Health}_i}$.

No closed-form solution $\implies$ solved numerically (same as logit). Software reports coefficients, standard errors, and predicted counts $\hat{\mu}_i$.

$\implies$ The structure is identical to binary logit/probit MLE, just with a different distribution (Poisson instead of Bernoulli).

Interpreting Coefficients: Semi-Elasticities

Take the log-link equation:

$$\ln(\mu_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Insurance}_i + \beta_3 \text{Health}_i$$

Interpreting Coefficients: Semi-Elasticities

Take the log-link equation:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

Recall: a difference in logs approximates a percent change. A one-unit increase in Age_i, holding everything else fixed:

$$\ln(\mu_i^{\text{new}}) - \ln(\mu_i^{\text{old}}) = \beta_1 \quad \Longleftrightarrow \quad \frac{\mu_i^{\text{new}} - \mu_i^{\text{old}}}{\mu_i^{\text{old}}} \approx \beta_1$$

Interpreting Coefficients: Semi-Elasticities

Take the log-link equation:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

Recall: a difference in logs approximates a percent change. A one-unit increase in Age_i, holding everything else fixed:

$$\ln(\mu_i^{\text{new}}) - \ln(\mu_i^{\text{old}}) = \beta_1 \quad \Longleftrightarrow \quad \frac{\mu_i^{\text{new}} - \mu_i^{\text{old}}}{\mu_i^{\text{old}}} \approx \beta_1$$

$\Longrightarrow$ Each coefficient is a **semi-elasticity**: a one-unit increase in x_k changes the expected count by approximately $\beta_k \times 100\%$.

Interpreting Coefficients: Semi-Elasticities

Take the log-link equation:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

Recall: a difference in logs approximates a percent change. A one-unit increase in Age_i, holding everything else fixed:

$$\ln(\mu_i^{\text{new}}) - \ln(\mu_i^{\text{old}}) = \beta_1 \quad \Longleftrightarrow \quad \frac{\mu_i^{\text{new}} - \mu_i^{\text{old}}}{\mu_i^{\text{old}}} \approx \beta_1$$

$\Longrightarrow$ Each coefficient is a **semi-elasticity**: a one-unit increase in x_k changes the expected count by approximately $\beta_k \times 100\%$.

For small $|\beta_k|$ (say < 0.1), this approximation is accurate. For larger coefficients, use the exact formula: $100 \times (e^{\beta_k} - 1)\%$.

Interpreting Coefficients: Semi-Elasticities

Take the log-link equation:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

Recall: a difference in logs approximates a percent change. A one-unit increase in Age_i, holding everything else fixed:

$$\ln(\mu_i^{\text{new}}) - \ln(\mu_i^{\text{old}}) = \beta_1 \quad \Longleftrightarrow \quad \frac{\mu_i^{\text{new}} - \mu_i^{\text{old}}}{\mu_i^{\text{old}}} \approx \beta_1$$

$\Longrightarrow$ Each coefficient is a **semi-elasticity**: a one-unit increase in x_k changes the expected count by approximately $\beta_k \times 100\%$.

For small $|\beta_k|$ (say < 0.1), this approximation is accurate. For larger coefficients, use the exact formula: $100 \times (e^{\beta_k} - 1)\%$.

Example (Insurance, a dummy variable): if $\hat{\beta}_2 = 0.54$, then $e^{0.54} - 1 = 0.72$, so insured individuals have about 72% more visits.

Numeric Example: Predicted Visits

Suppose the Poisson estimates are $\hat{\beta}_0 = 0.50$, $\hat{\beta}_{\text{age}} = 0.017$, $\hat{\beta}_{\text{ins}} = 0.54$, $\hat{\beta}_{\text{health}} = -0.27$.

Numeric Example: Predicted Visits

Suppose the Poisson estimates are $\hat{\beta}_0 = 0.50$, $\hat{\beta}_{\text{age}} = 0.017$, $\hat{\beta}_{\text{ins}} = 0.54$, $\hat{\beta}_{\text{health}} = -0.27$.

Person A: 45 years old, insured, average health (Health $= 0$):

$$\ln(\hat{\mu}_A) = 0.50 + 0.017 \times 45 + 0.54 \times 1 + (-0.27) \times 0 = 1.805$$

$$\hat{\mu}_A = e^{1.805} \approx 6.1 \text{ visits per year}$$

Numeric Example: Predicted Visits

Suppose the Poisson estimates are $\hat{\beta}_0 = 0.50$, $\hat{\beta}_{\text{age}} = 0.017$, $\hat{\beta}_{\text{ins}} = 0.54$, $\hat{\beta}_{\text{health}} = -0.27$.

Person A: 45 years old, insured, average health (Health $= 0$):

$$\ln(\hat{\mu}_A) = 0.50 + 0.017 \times 45 + 0.54 \times 1 + (-0.27) \times 0 = 1.805$$

$$\hat{\mu}_A = e^{1.805} \approx 6.1 \text{ visits per year}$$

Person B: 25 years old, uninsured, healthy (Health $= 1.5$):

$$\ln(\hat{\mu}_B) = 0.50 + 0.017 \times 25 + 0.54 \times 0 + (-0.27) \times 1.5 = 0.520$$

$$\hat{\mu}_B = e^{0.520} \approx 1.7 \text{ visits per year}$$

Numeric Example: Predicted Visits

Suppose the Poisson estimates are $\hat{\beta}_0 = 0.50$, $\hat{\beta}_{\text{age}} = 0.017$, $\hat{\beta}_{\text{ins}} = 0.54$, $\hat{\beta}_{\text{health}} = -0.27$.

Person A: 45 years old, insured, average health (Health $= 0$):

$$\ln(\hat{\mu}_A) = 0.50 + 0.017 \times 45 + 0.54 \times 1 + (-0.27) \times 0 = 1.805$$

$$\hat{\mu}_A = e^{1.805} \approx 6.1 \text{ visits per year}$$

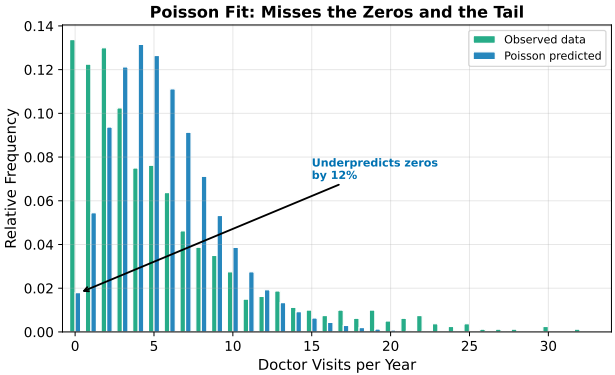
Person B: 25 years old, uninsured, healthy (Health $= 1.5$):

$$\ln(\hat{\mu}_B) = 0.50 + 0.017 \times 25 + 0.54 \times 0 + (-0.27) \times 1.5 = 0.520$$

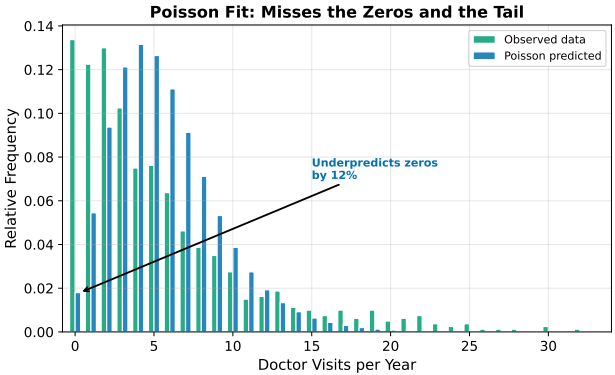
$$\hat{\mu}_B = e^{0.520} \approx 1.7 \text{ visits per year}$$

$\Longrightarrow$ Both predictions are positive. Compare to OLS, which predicted negative visits for Person B.

Poisson Fit to Our Data



Poisson Fit to Our Data



The Poisson predicts only 2% zeros; the data has 13%. It underpredicts zeros and underpredicts the right tail, concentrating too much mass in the middle. Why?

The Equidispersion Problem

Recall the Poisson assumption: $\text{Var}(Y_i) = \mu_i$. This restriction is called **equidispersion**: the variance must equal the mean.

The Equidispersion Problem

Recall the Poisson assumption: $\text{Var}(Y_i) = \mu_i$. This restriction is called **equidispersion**: the variance must equal the mean.

This means individuals with $\mu_i = 6$ expected visits should have variance $= 6$. But in our data:

The Equidispersion Problem

Recall the Poisson assumption: $\text{Var}(Y_i) = \mu_i$. This restriction is called **equidispersion**: the variance must equal the mean.

This means individuals with $\mu_i = 6$ expected visits should have variance $= 6$. But in our data:

	Mean visits	**Variance**
Full sample	5.7	43.8

The Equidispersion Problem

Recall the Poisson assumption: $\text{Var}(Y_i) = \mu_i$. This restriction is called **equidispersion**: the variance must equal the mean.

This means individuals with $\mu_i = 6$ expected visits should have variance $= 6$. But in our data:

	Mean visits	**Variance**
Full sample	5.7	43.8

The variance is **7.7 times** the mean. The Poisson model says these should be equal.

The Equidispersion Problem

Recall the Poisson assumption: $\text{Var}(Y_i) = \mu_i$. This restriction is called **equidispersion**: the variance must equal the mean.

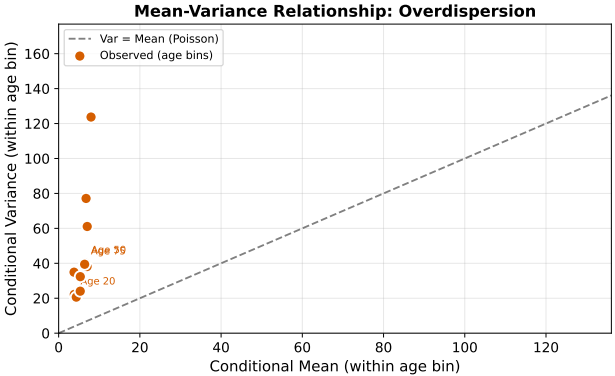
This means individuals with $\mu_i = 6$ expected visits should have variance $= 6$. But in our data:

	Mean visits	**Variance**
Full sample	5.7	43.8

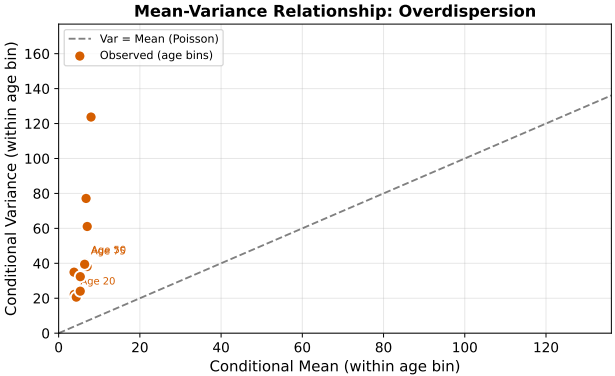
The variance is **7.7 times** the mean. The Poisson model says these should be equal.

This is called **overdispersion**: more variability in the data than the Poisson distribution allows. It is extremely common with count outcomes.

Visualizing Overdispersion



Visualizing Overdispersion



Every age bin lies **above** the 45-degree line. The variance grows faster than the mean, violating the Poisson assumption.

Consequence of Overdispersion: Standard Errors Are Wrong

What happens if we fit Poisson regression to overdispersed data?

Consequence of Overdispersion: Standard Errors Are Wrong

What happens if we fit Poisson regression to overdispersed data?

- The **coefficient estimates** are still consistent, as long as the conditional mean ($\mu_i = e^{\beta_0 + \beta_1 x_1 + \cdots}$) is correctly specified

Consequence of Overdispersion: Standard Errors Are Wrong

What happens if we fit Poisson regression to overdispersed data?

- The **coefficient estimates** are still consistent, as long as the conditional mean ($\mu_i = e^{\beta_0+\beta_1 x_1+\cdots}$) is correctly specified
- But the **model-based standard errors are too small** because they assume $\text{Var}(Y_i) = \mu_i$, while the true variance is larger

Consequence of Overdispersion: Standard Errors Are Wrong

What happens if we fit Poisson regression to overdispersed data?

- The **coefficient estimates** are still consistent, as long as the conditional mean ($\mu_i = e^{\beta_0 + \beta_1 x_1 + \cdots}$) is correctly specified
- But the **model-based standard errors are too small** because they assume $\text{Var}(Y_i) = \mu_i$, while the true variance is larger
- $\implies$ Confidence intervals are too narrow, p-values are too small, you reject the null too often

Consequence of Overdispersion: Standard Errors Are Wrong

What happens if we fit Poisson regression to overdispersed data?

- The **coefficient estimates** are still consistent, as long as the conditional mean ($\mu_i = e^{\beta_0 + \beta_1 x_1 + \cdots}$) is correctly specified
- But the **model-based standard errors are too small** because they assume $\text{Var}(Y_i) = \mu_i$, while the true variance is larger
- $\implies$ Confidence intervals are too narrow, p-values are too small, you reject the null too often

$\implies$ With overdispersion, Poisson regression gives you the right answer with the wrong confidence.

What Poisson Gets Right, and What It Gets Wrong

What Poisson gets right:

- Positive predictions for all covariate values (the log link)
- Coefficients are semi-elasticities, easy to interpret
- Consistent coefficient estimates (even with overdispersion)

What Poisson Gets Right, and What It Gets Wrong

What Poisson gets right:

- Positive predictions for all covariate values (the log link)
- Coefficients are semi-elasticities, easy to interpret
- Consistent coefficient estimates (even with overdispersion)

What Poisson gets wrong:

- Forces $\text{Var}(Y_i) = \mu_i$, but our data has variance $7.7\times$ the mean
- Standard errors are too small $\implies$ false confidence
- Predicted distribution misses the spike at zero and the long tail

What Poisson Gets Right, and What It Gets Wrong

What Poisson gets right:

- Positive predictions for all covariate values (the log link)
- Coefficients are semi-elasticities, easy to interpret
- Consistent coefficient estimates (even with overdispersion)

What Poisson gets wrong:

- Forces $\text{Var}(Y_i) = \mu_i$, but our data has variance $7.7\times$ the mean
- Standard errors are too small $\implies$ false confidence
- Predicted distribution misses the spike at zero and the long tail

Can we keep the Poisson's log link but relax the variance constraint?

Outline

Our Data Has Variance 7.7× the Mean

The Poisson forces $\text{Var}(Y_i) = \mu_i$. Our data violates this dramatically:

$$\frac{\text{Sample Variance}}{\text{Sample Mean}} = \frac{43.8}{5.7} = 7.7$$

Our Data Has Variance 7.7× the Mean

The Poisson forces $\text{Var}(Y_i) = \mu_i$. Our data violates this dramatically:

$$\frac{\text{Sample Variance}}{\text{Sample Mean}} = \frac{43.8}{5.7} = 7.7$$

We want a model that:

- Keeps the **same log link**: $\ln(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots$ (positive predictions, semi-elasticities)
- Adds a **free variance parameter** so the variance can exceed the mean

Our Data Has Variance 7.7× the Mean

The Poisson forces $\text{Var}(Y_i) = \mu_i$. Our data violates this dramatically:

$$\frac{\text{Sample Variance}}{\text{Sample Mean}} = \frac{43.8}{5.7} = 7.7$$

We want a model that:

- Keeps the **same log link**: $\ln(\mu_i) = \beta_0 + \beta_1 x_1 + \cdots$ (positive predictions, semi-elasticities)
- Adds a **free variance parameter** so the variance can exceed the mean

$\Longrightarrow$ The **Negative Binomial** does exactly this: it generalizes the Poisson by adding one parameter.

Adding an Overdispersion Parameter

The Poisson model forces $\text{Var}(Y_i) = \mu_i$. To allow overdispersion, we add a parameter $\alpha > 0$:

$$\text{Var}(Y_i) = \mu_i + \alpha \mu_i^2$$

Adding an Overdispersion Parameter

The Poisson model forces $\text{Var}(Y_i) = \mu_i$. To allow overdispersion, we add a parameter $\alpha > 0$:

$$\text{Var}(Y_i) = \mu_i + \alpha\,\mu_i^2$$

- The extra term $\alpha\,\mu_i^2$ lets the variance **exceed** the mean
- How much extra variance depends on α

Adding an Overdispersion Parameter

The Poisson model forces $\text{Var}(Y_i) = \mu_i$. To allow overdispersion, we add a parameter $\alpha > 0$:

$$\text{Var}(Y_i) = \mu_i + \alpha\,\mu_i^2$$

- The extra term $\alpha\,\mu_i^2$ lets the variance **exceed** the mean
- How much extra variance depends on α

Boundary condition: when $\alpha \to 0$, the extra term vanishes and we get $\text{Var}(Y_i) = \mu_i$. That is exactly Poisson.

Adding an Overdispersion Parameter

The Poisson model forces $\text{Var}(Y_i) = \mu_i$. To allow overdispersion, we add a parameter $\alpha > 0$:

$$\text{Var}(Y_i) = \mu_i + \alpha\,\mu_i^2$$

- The extra term $\alpha\,\mu_i^2$ lets the variance **exceed** the mean
- How much extra variance depends on α

Boundary condition: when $\alpha \to 0$, the extra term vanishes and we get $\text{Var}(Y_i) = \mu_i$. That is exactly Poisson.

$\Longrightarrow$ Poisson is a special case of the Negative Binomial with $\alpha = 0$. The NB nests the Poisson.

The Negative Binomial Model

The Negative Binomial regression model specifies:

The Negative Binomial Model

The Negative Binomial regression model specifies:

1. **Same log link** as Poisson:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

The Negative Binomial Model

The Negative Binomial regression model specifies:

1. **Same log link** as Poisson:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

2. **NB probability formula** instead of Poisson. It uses a different formula to assign probabilities to each count value (software handles it)

The Negative Binomial Model

The Negative Binomial regression model specifies:

1. **Same log link** as Poisson:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

2. **NB probability formula** instead of Poisson. It uses a different formula to assign probabilities to each count value (software handles it)
3. **Variance:** $\text{Var}(Y_i) = \mu_i + \alpha\,\mu_i^2$, where α is estimated from the data

The Negative Binomial Model

The Negative Binomial regression model specifies:

1. **Same log link** as Poisson:

$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

2. **NB probability formula** instead of Poisson. It uses a different formula to assign probabilities to each count value (software handles it)
3. **Variance:** $\text{Var}(Y_i) = \mu_i + \alpha\,\mu_i^2$, where α is estimated from the data

We estimate $(\beta_0, \beta_1, \beta_2, \beta_3)$ and α jointly by MLE.

The Negative Binomial Model

The Negative Binomial regression model specifies:

1. **Same log link** as Poisson:

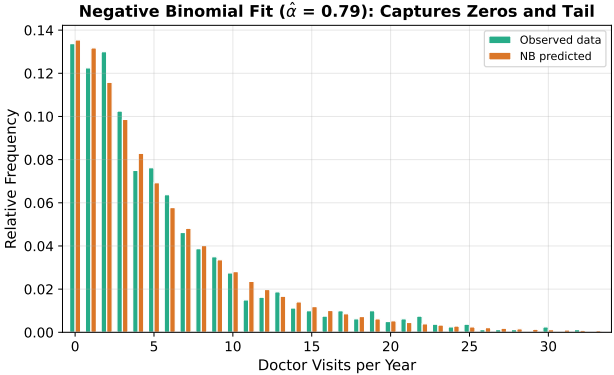
$$\ln(\mu_i) = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,\text{Insurance}_i + \beta_3 \,\text{Health}_i$$

2. **NB probability formula** instead of Poisson. It uses a different formula to assign probabilities to each count value (software handles it)
3. **Variance:** $\text{Var}(Y_i) = \mu_i + \alpha\,\mu_i^2$, where α is estimated from the data

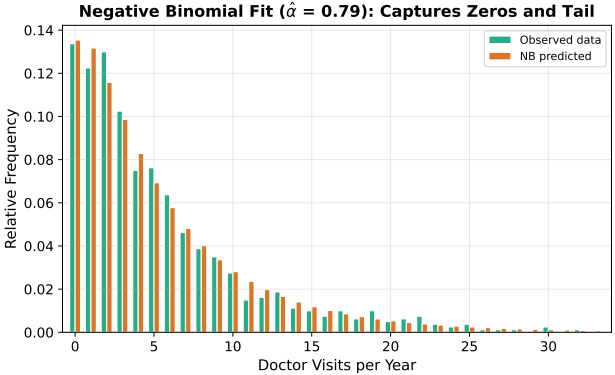
We estimate $(\beta_0, \beta_1, \beta_2, \beta_3)$ and α jointly by MLE.

$\Longrightarrow$ Coefficients have the **same semi-elasticity interpretation** as Poisson. The only change is allowing more variance.

NB Fit to Our Data

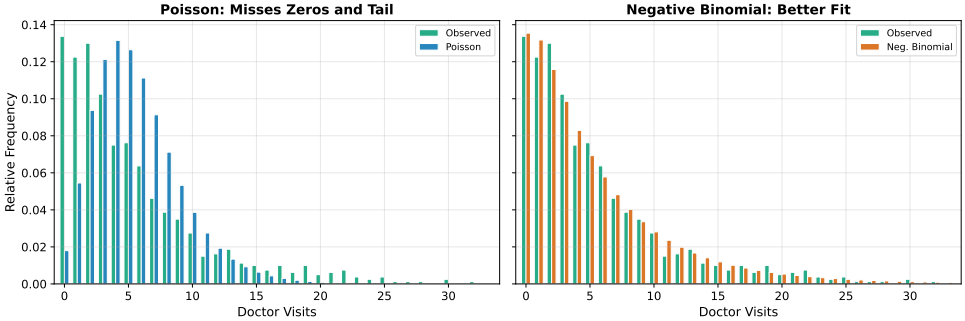


NB Fit to Our Data

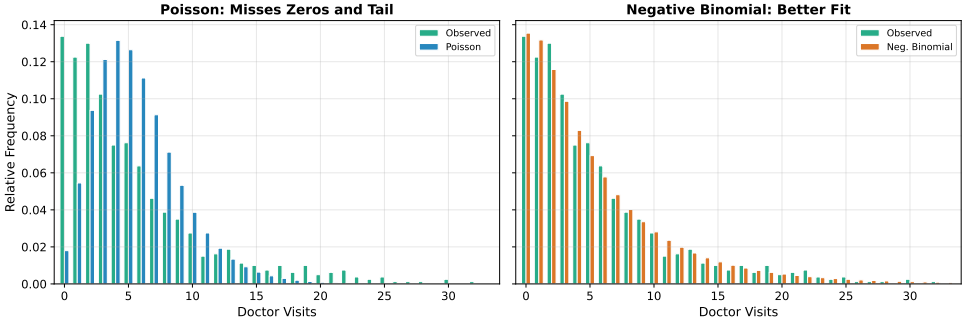


With an estimated $\hat{\alpha} = 0.79$, the Negative Binomial captures the spike at zero and the long right tail that Poisson missed.

Side-by-Side: Poisson vs. Negative Binomial



Side-by-Side: Poisson vs. Negative Binomial



The Poisson (left) squeezes too much mass into the middle. The NB (right) spreads it out to match the data.

Testing for Overdispersion

Since Poisson is nested inside NB (H_0: $\alpha = 0$), we can test directly.

Testing for Overdispersion

Since Poisson is nested inside NB (H_0: $\alpha = 0$), we can test directly.

Method 1: Cameron–Trivedi regression test.
Regress $(y_i - \hat{\mu}_i)^2 - y_i$ on $\hat{\mu}_i^2$ (no intercept). If the slope $\hat{\alpha}$ is significantly positive $\implies$ overdispersion.

Testing for Overdispersion

Since Poisson is nested inside NB (H_0: $\alpha = 0$), we can test directly.

Method 1: Cameron–Trivedi regression test.
Regress $(y_i - \hat{\mu}_i)^2 - y_i$ on $\hat{\mu}_i^2$ (no intercept). If the slope $\hat{\alpha}$ is significantly positive $\implies$ overdispersion.

Intuition: under the Poisson, $(y_i - \mu_i)^2 - y_i$ should average to zero. If it is systematically positive, there is extra variance beyond what Poisson allows.

Testing for Overdispersion

Since Poisson is nested inside NB (H_0: $\alpha = 0$), we can test directly.

Method 1: Cameron–Trivedi regression test.
Regress $(y_i - \hat{\mu}_i)^2 - y_i$ on $\hat{\mu}_i^2$ (no intercept). If the slope $\hat{\alpha}$ is significantly positive $\implies$ overdispersion.

Intuition: under the Poisson, $(y_i - \mu_i)^2 - y_i$ should average to zero. If it is systematically positive, there is extra variance beyond what Poisson allows.

Method 2: Likelihood ratio test.
$LR = 2\left[\ell_{\text{NB}} - \ell_{\text{Poisson}}\right] \sim \chi^2_1$ under H_0: $\alpha = 0$ (conservative, since $\alpha = 0$ is on the boundary of the parameter space).

Testing for Overdispersion

Since Poisson is nested inside NB (H_0: $\alpha = 0$), we can test directly.

Method 1: Cameron–Trivedi regression test.
Regress $(y_i - \hat{\mu}_i)^2 - y_i$ on $\hat{\mu}_i^2$ (no intercept). If the slope $\hat{\alpha}$ is significantly positive $\implies$ overdispersion.

Intuition: under the Poisson, $(y_i - \mu_i)^2 - y_i$ should average to zero. If it is systematically positive, there is extra variance beyond what Poisson allows.

Method 2: Likelihood ratio test.
$LR = 2\left[\ell_{\text{NB}} - \ell_{\text{Poisson}}\right] \sim \chi^2_1$ under H_0: $\alpha = 0$ (conservative, since $\alpha = 0$ is on the boundary of the parameter space).

In our data: $\hat{\alpha} = 0.79$ with $p < 0.001$.

Testing for Overdispersion

Since Poisson is nested inside NB (H_0: $\alpha = 0$), we can test directly.

Method 1: Cameron–Trivedi regression test.
Regress $(y_i - \hat{\mu}_i)^2 - y_i$ on $\hat{\mu}_i^2$ (no intercept). If the slope $\hat{\alpha}$ is significantly positive $\implies$ overdispersion.

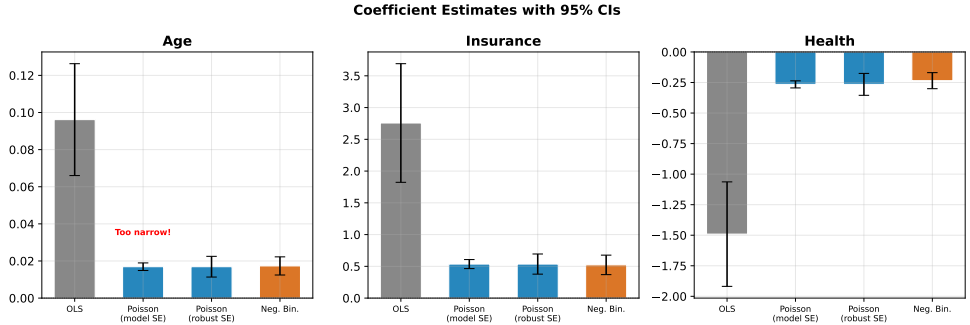
Intuition: under the Poisson, $(y_i - \mu_i)^2 - y_i$ should average to zero. If it is systematically positive, there is extra variance beyond what Poisson allows.

Method 2: Likelihood ratio test.
$LR = 2\left[\ell_{\text{NB}} - \ell_{\text{Poisson}}\right] \sim \chi^2_1$ under H_0: $\alpha = 0$ (conservative, since $\alpha = 0$ is on the boundary of the parameter space).

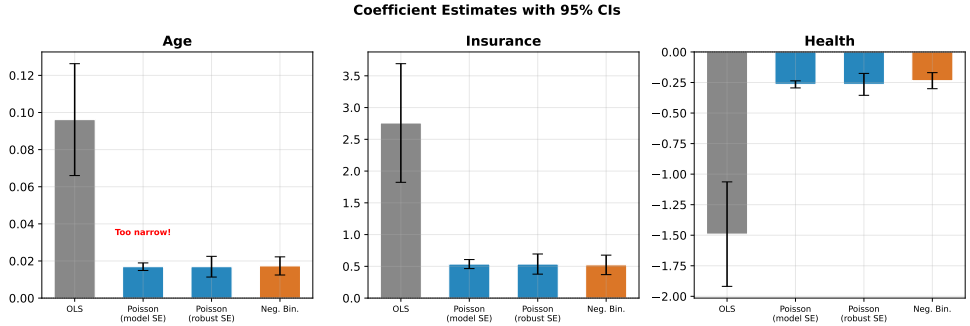
In our data: $\hat{\alpha} = 0.79$ with $p < 0.001$.

$\implies$ Strong evidence of overdispersion. The Poisson model is rejected in favor of NB.

Coefficient Estimates: OLS vs. Poisson vs. NB



Coefficient Estimates: OLS vs. Poisson vs. NB



Poisson and NB give similar coefficient estimates, but Poisson model-based SEs are far too narrow. The NB SEs properly account for overdispersion.

Why Poisson SEs Are Too Small

	Poisson (model SE)	**Poisson (robust SE)**	**NB**
Age	0.001	0.003	0.003
Insurance	0.036	0.081	0.078
Health	0.015	0.046	0.034

Why Poisson SEs Are Too Small

	Poisson (model SE)	Poisson (robust SE)	NB
Age	0.001	0.003	0.003
Insurance	0.036	0.081	0.078
Health	0.015	0.046	0.034

Poisson model SEs assume $\text{Var}(Y_i) = \mu_i$. Since the true variance is much larger, these SEs are roughly 2–3 times too small.

Why Poisson SEs Are Too Small

	Poisson (model SE)	Poisson (robust SE)	NB
Age	0.001	0.003	0.003
Insurance	0.036	0.081	0.078
Health	0.015	0.046	0.034

Poisson model SEs assume $\text{Var}(Y_i) = \mu_i$. Since the true variance is much larger, these SEs are roughly 2–3 times too small.

Two fixes:

1. **Robust (sandwich) SEs:** keep the Poisson model but correct the SEs
2. **Negative Binomial:** model the extra variance directly

Why Poisson SEs Are Too Small

	Poisson (model SE)	Poisson (robust SE)	NB
Age	0.001	0.003	0.003
Insurance	0.036	0.081	0.078
Health	0.015	0.046	0.034

Poisson model SEs assume $\text{Var}(Y_i) = \mu_i$. Since the true variance is much larger, these SEs are roughly 2–3 times too small.

Two fixes:

1. **Robust (sandwich) SEs:** keep the Poisson model but correct the SEs
2. **Negative Binomial:** model the extra variance directly

$\Longrightarrow$ Both give CIs based on a consistent estimator of the true sampling variance, so coverage is correct asymptotically (typically wider than the under-dispersion-assuming Poisson CI).

Three-Model Comparison: OLS vs. Poisson vs. NB

	OLS	**Poisson**	**Neg. Binomial**
Predicted range	$(-\infty, +\infty)$	$(0, +\infty)$	$(0, +\infty)$
Variance assumption	constant	$\text{Var} = \mu$	$\text{Var} = \mu + \alpha\mu^2$
SE reliability (model-based)	heteroskedasticity biased	too small if overdispersed	correct if α well-estimated
Coefficient interpretation	level change (Δy per unit Δx)	semi-elasticity ($\approx \%\Delta y$)	semi-elasticity ($\approx \%\Delta y$)

Three-Model Comparison: OLS vs. Poisson vs. NB

	OLS	**Poisson**	**Neg. Binomial**
Predicted range	$(-\infty, +\infty)$	$(0, +\infty)$	$(0, +\infty)$
Variance assumption	constant	$\text{Var} = \mu$	$\text{Var} = \mu + \alpha\mu^2$
SE reliability (model-based)	heteroskedasticity biased	too small if overdispersed	correct if α well-estimated
Coefficient interpretation	level change (Δy per unit Δx)	semi-elasticity ($\approx \%\Delta y$)	semi-elasticity ($\approx \%\Delta y$)

$\Longrightarrow$ Moving from OLS to Poisson solves the boundary problem; moving from Poisson to NB solves the variance problem.

Outline

Quasi-Poisson: A Quick SE Correction

Sometimes you want to keep the Poisson model structure but fix the SEs. The **Quasi-Poisson** approach:

Quasi-Poisson: A Quick SE Correction

Sometimes you want to keep the Poisson model structure but fix the SEs. The **Quasi-Poisson** approach:

- Estimates the same coefficients as Poisson MLE
- Introduces a dispersion parameter ϕ: $\text{Var}(Y_i) = \phi\,\mu_i$
- Multiplies all Poisson SEs by $\sqrt{\hat{\phi}}$, where $\hat{\phi}$ is estimated from the model residuals

Quasi-Poisson: A Quick SE Correction

Sometimes you want to keep the Poisson model structure but fix the SEs. The **Quasi-Poisson** approach:

- Estimates the same coefficients as Poisson MLE
- Introduces a dispersion parameter ϕ: $\text{Var}(Y_i) = \phi\,\mu_i$
- Multiplies all Poisson SEs by $\sqrt{\hat{\phi}}$, where $\hat{\phi}$ is estimated from the model residuals

In our data, Quasi-Poisson SEs are roughly 2–3 times larger than Poisson model SEs.

Quasi-Poisson: A Quick SE Correction

Sometimes you want to keep the Poisson model structure but fix the SEs. The **Quasi-Poisson** approach:

- Estimates the same coefficients as Poisson MLE
- Introduces a dispersion parameter ϕ: $\text{Var}(Y_i) = \phi\,\mu_i$
- Multiplies all Poisson SEs by $\sqrt{\hat{\phi}}$, where $\hat{\phi}$ is estimated from the model residuals

In our data, Quasi-Poisson SEs are roughly 2–3 times larger than Poisson model SEs.

Quasi-Poisson vs. robust SEs: Quasi-Poisson assumes $\text{Var} = \phi\mu$ (overdispersion is a linear scaling of the mean). Robust SEs make no assumption about the variance form.

Quasi-Poisson: A Quick SE Correction

Sometimes you want to keep the Poisson model structure but fix the SEs. The **Quasi-Poisson** approach:

- Estimates the same coefficients as Poisson MLE
- Introduces a dispersion parameter ϕ: $\text{Var}(Y_i) = \phi\,\mu_i$
- Multiplies all Poisson SEs by $\sqrt{\hat{\phi}}$, where $\hat{\phi}$ is estimated from the model residuals

In our data, Quasi-Poisson SEs are roughly 2–3 times larger than Poisson model SEs.

Quasi-Poisson vs. robust SEs: Quasi-Poisson assumes $\text{Var} = \phi\mu$ (overdispersion is a linear scaling of the mean). Robust SEs make no assumption about the variance form.

Approach	**Variance structure**	**When to use**
Poisson	$\text{Var} = \mu$	Mild or no overdispersion
Quasi-Poisson	$\text{Var} = \phi\,\mu$	Quick SE correction; no full likelihood
Neg. Binomial	$\text{Var} = \mu + \alpha\mu^2$	Full model; predictions, LR tests, AIC

Excess Zeros: When to Consider Zero-Inflated Models

Sometimes overdispersion comes from **excess zeros**: more zeros than even the NB can accommodate.

Excess Zeros: When to Consider Zero-Inflated Models

Sometimes overdispersion comes from **excess zeros**: more zeros than even the NB can accommodate.

Example: doctor visits. Some people *never* go (they avoid doctors entirely), while others go based on their health needs. Two different processes generate the zeros.

Excess Zeros: When to Consider Zero-Inflated Models

Sometimes overdispersion comes from **excess zeros**: more zeros than even the NB can accommodate.

Example: doctor visits. Some people *never* go (they avoid doctors entirely), while others go based on their health needs. Two different processes generate the zeros.

Zero-inflated models combine:

1. A binary model (logit) for whether someone is a "certain zero" vs. a potential visitor
2. A count model (Poisson or NB) for potential visitors

Excess Zeros: When to Consider Zero-Inflated Models

Sometimes overdispersion comes from **excess zeros**: more zeros than even the NB can accommodate.

Example: doctor visits. Some people *never* go (they avoid doctors entirely), while others go based on their health needs. Two different processes generate the zeros.

Zero-inflated models combine:

1. A binary model (logit) for whether someone is a "certain zero" vs. a potential visitor
2. A count model (Poisson or NB) for potential visitors

How to tell if you need one:

- Compare observed zero proportion to the predicted zero proportion from your NB model
- If NB already fits the zeros well, zero-inflation is unnecessary

Excess Zeros: When to Consider Zero-Inflated Models

Sometimes overdispersion comes from **excess zeros**: more zeros than even the NB can accommodate.

Example: doctor visits. Some people *never* go (they avoid doctors entirely), while others go based on their health needs. Two different processes generate the zeros.

Zero-inflated models combine:

1. A binary model (logit) for whether someone is a "certain zero" vs. a potential visitor
2. A count model (Poisson or NB) for potential visitors

How to tell if you need one:

- Compare observed zero proportion to the predicted zero proportion from your NB model
- If NB already fits the zeros well, zero-inflation is unnecessary

$\Longrightarrow$ In our data, NB captures the 13% zeros adequately. Zero-inflation would be needed if, say, 40% of the sample had zero visits.

Decision Framework: Which Count Model to Use

1. **Start with Poisson.** It is the simplest count model and gives consistent coefficient estimates even under overdispersion

Decision Framework: Which Count Model to Use

1. **Start with Poisson.** It is the simplest count model and gives consistent coefficient estimates even under overdispersion
2. **Test for overdispersion.** Cameron–Trivedi test or LR test (H_0: $\alpha = 0$)

Decision Framework: Which Count Model to Use

1. **Start with Poisson.** It is the simplest count model and gives consistent coefficient estimates even under overdispersion
2. **Test for overdispersion.** Cameron–Trivedi test or LR test (H_0: $\alpha = 0$)
3. **If overdispersion is detected:**
   - **Minimum fix:** use robust (sandwich) SEs with the Poisson model
   - **Better fix:** switch to Negative Binomial regression

Decision Framework: Which Count Model to Use

1. **Start with Poisson.** It is the simplest count model and gives consistent coefficient estimates even under overdispersion
2. **Test for overdispersion.** Cameron–Trivedi test or LR test (H_0: $\alpha = 0$)
3. **If overdispersion is detected:**
   - **Minimum fix:** use robust (sandwich) SEs with the Poisson model
   - **Better fix:** switch to Negative Binomial regression
4. **If excess zeros remain:** consider a zero-inflated Poisson (ZIP) or zero-inflated NB (ZINB)

Decision Framework: Which Count Model to Use

1. **Start with Poisson.** It is the simplest count model and gives consistent coefficient estimates even under overdispersion
2. **Test for overdispersion.** Cameron–Trivedi test or LR test (H_0: $\alpha = 0$)
3. **If overdispersion is detected:**
   - **Minimum fix:** use robust (sandwich) SEs with the Poisson model
   - **Better fix:** switch to Negative Binomial regression
4. **If excess zeros remain:** consider a zero-inflated Poisson (ZIP) or zero-inflated NB (ZINB)
5. **If the outcome has a known upper bound** (e.g., number correct out of 10):

   $\Longrightarrow$ This is not a count model problem; consider binomial regression instead

Summary: Back to Doctor Visits

1. **OLS on counts fails:** it predicted negative visits for young, healthy, uninsured individuals

Summary: Back to Doctor Visits

1. **OLS on counts fails:** it predicted negative visits for young, healthy, uninsured individuals
2. **Poisson regression** uses a log link ($\ln \mu_i = \beta_0 + \beta_1 x_1 + \cdots$) to guarantee positive predictions. Coefficients are semi-elasticities

Summary: Back to Doctor Visits

1. **OLS on counts fails:** it predicted negative visits for young, healthy, uninsured individuals
2. **Poisson regression** uses a log link ($\ln \mu_i = \beta_0 + \beta_1 x_1 + \cdots$) to guarantee positive predictions. Coefficients are semi-elasticities
3. **Equidispersion** ($\text{Var} = \mu$) almost never holds in practice. Our doctor visits data had variance 7.7$\times$ the mean, so Poisson SEs were 2–3$\times$ too small

Summary: Back to Doctor Visits

1. **OLS on counts fails:** it predicted negative visits for young, healthy, uninsured individuals
2. **Poisson regression** uses a log link ($\ln \mu_i = \beta_0 + \beta_1 x_1 + \cdots$) to guarantee positive predictions. Coefficients are semi-elasticities
3. **Equidispersion** ($\text{Var} = \mu$) almost never holds in practice. Our doctor visits data had variance 7.7$\times$ the mean, so Poisson SEs were 2–3$\times$ too small
4. **Negative Binomial** adds one parameter (α) that allows $\text{Var} = \mu + \alpha\mu^2$. It captured the spike at zero and the long tail that Poisson missed

Summary: Back to Doctor Visits

1. **OLS on counts fails:** it predicted negative visits for young, healthy, uninsured individuals
2. **Poisson regression** uses a log link ($\ln \mu_i = \beta_0 + \beta_1 x_1 + \cdots$) to guarantee positive predictions. Coefficients are semi-elasticities
3. **Equidispersion** ($\text{Var} = \mu$) almost never holds in practice. Our doctor visits data had variance 7.7$\times$ the mean, so Poisson SEs were 2–3$\times$ too small
4. **Negative Binomial** adds one parameter (α) that allows $\text{Var} = \mu + \alpha\mu^2$. It captured the spike at zero and the long tail that Poisson missed
5. **Test for overdispersion** before reporting Poisson results. Use the Cameron–Trivedi test or a likelihood ratio test

Summary: Back to Doctor Visits

1. **OLS on counts fails:** it predicted negative visits for young, healthy, uninsured individuals
2. **Poisson regression** uses a log link ($\ln \mu_i = \beta_0 + \beta_1 x_1 + \cdots$) to guarantee positive predictions. Coefficients are semi-elasticities
3. **Equidispersion** ($\text{Var} = \mu$) almost never holds in practice. Our doctor visits data had variance 7.7$\times$ the mean, so Poisson SEs were 2–3$\times$ too small
4. **Negative Binomial** adds one parameter (α) that allows $\text{Var} = \mu + \alpha\mu^2$. It captured the spike at zero and the long tail that Poisson missed
5. **Test for overdispersion** before reporting Poisson results. Use the Cameron–Trivedi test or a likelihood ratio test
6. **Zero-inflated models** are a further extension when excess zeros come from a separate process

Summary: Back to Doctor Visits

1. **OLS on counts fails:** it predicted negative visits for young, healthy, uninsured individuals
2. **Poisson regression** uses a log link ($\ln \mu_i = \beta_0 + \beta_1 x_1 + \cdots$) to guarantee positive predictions. Coefficients are semi-elasticities
3. **Equidispersion** ($\text{Var} = \mu$) almost never holds in practice. Our doctor visits data had variance 7.7$\times$ the mean, so Poisson SEs were 2–3$\times$ too small
4. **Negative Binomial** adds one parameter (α) that allows $\text{Var} = \mu + \alpha\mu^2$. It captured the spike at zero and the long tail that Poisson missed
5. **Test for overdispersion** before reporting Poisson results. Use the Cameron–Trivedi test or a likelihood ratio test
6. **Zero-inflated models** are a further extension when excess zeros come from a separate process

$\Longrightarrow$ Always start with Poisson, test for overdispersion, and upgrade to NB or robust SEs as needed.

Thank you!
jakeanderson@g.ucla.edu

The Tobit Model (Censored Regression)

When 40% of Your Data Is Piled Up at Zero

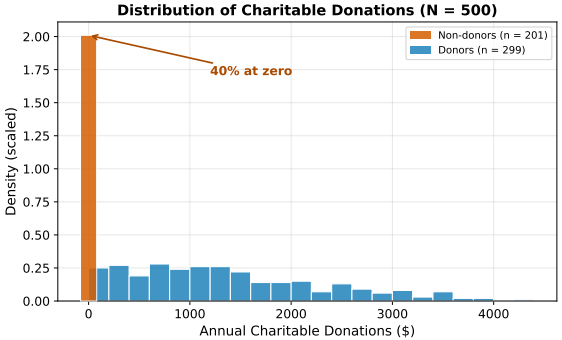
Jake Anderson

May 16, 2026

Outline

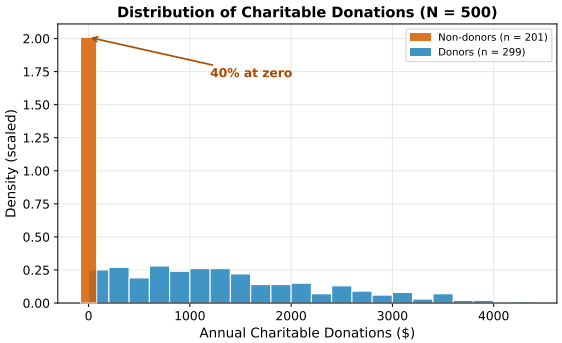
The Data: Charitable Donations

A researcher surveys **500 households** and records **annual charitable donations** ($). Covariates include household income ($1000s), years of education, and number of children.



The Data: Charitable Donations

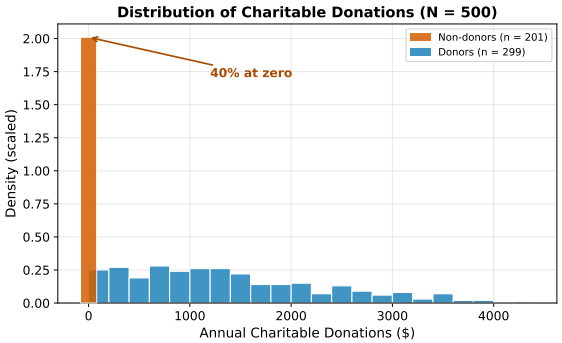
A researcher surveys **500 households** and records **annual charitable donations** ($). Covariates include household income ($1000s), years of education, and number of children.



What do you notice about this distribution?

The Data: Charitable Donations

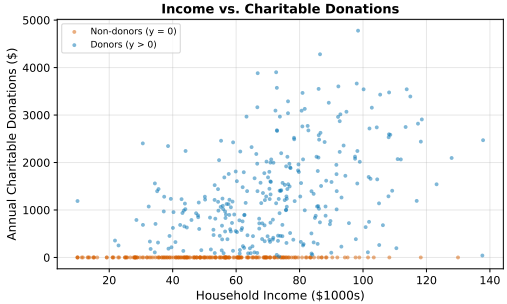
A researcher surveys **500 households** and records **annual charitable donations** ($). Covariates include household income ($1000s), years of education, and number of children.



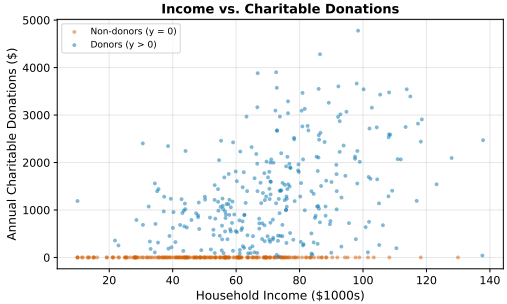
What do you notice about this distribution?

40% donate nothing; the rest are continuous and right-skewed. This is a **corner solution outcome**: a spike at zero plus a continuous positive tail.

Income and Donations



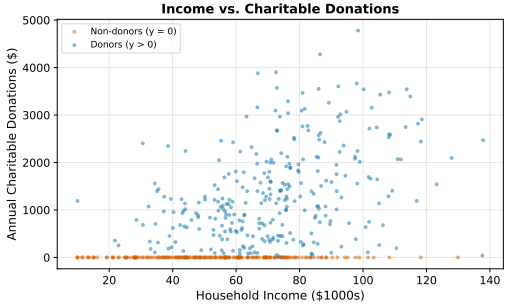
Income and Donations



Two features stand out:

- Orange points **piled up along** $y = 0$, mostly at lower incomes.
- Among donors (blue), a **positive relationship** between income and donations.

Income and Donations

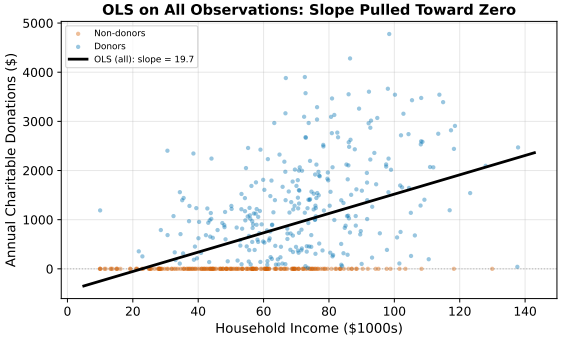


Two features stand out:

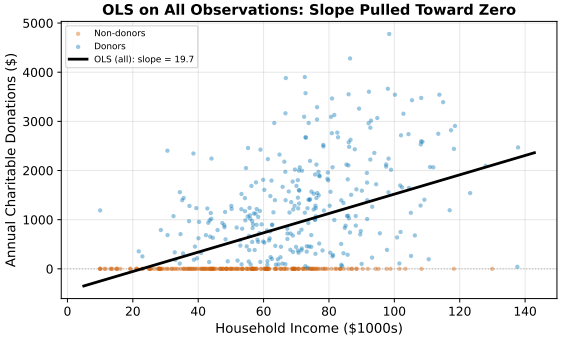
- Orange points **piled up along** $y = 0$, mostly at lower incomes.
- Among donors (blue), a **positive relationship** between income and donations.

What happens if we run OLS on this data?

OLS on All Observations: Slope Pulled Down

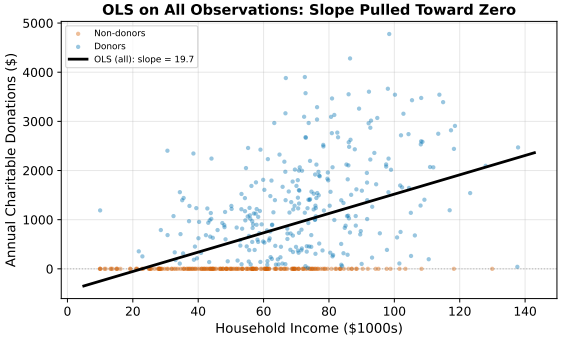


OLS on All Observations: Slope Pulled Down



OLS slope = 19.7 dollars per \$1000 income. But the true effect in the underlying model is **30 dollars per \$1000 income**.

OLS on All Observations: Slope Pulled Down

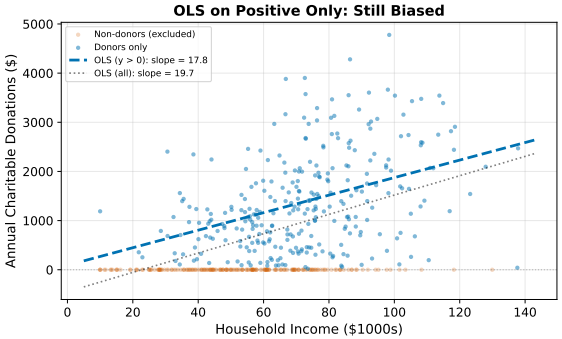


OLS slope = 19.7 dollars per \$1000 income. But the true effect in the underlying model is **30 dollars per \$1000 income**.

Why is OLS attenuated? The 200 non-donors all sit at $y = 0$ regardless of their income. OLS treats these as real zeros and tilts the line **toward the pile-up**, underestimating the income effect by about a third.

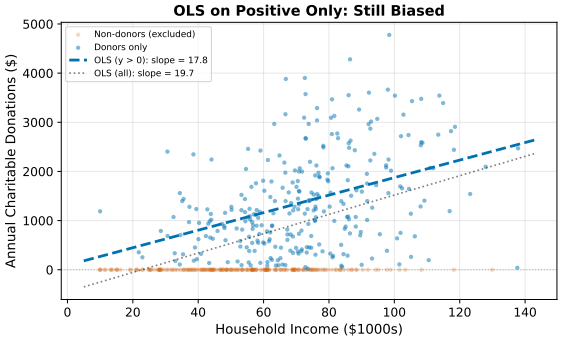
OLS on Positive Observations Only: Still Biased

Maybe we should drop the zeros and run OLS on donors only?



OLS on Positive Observations Only: Still Biased

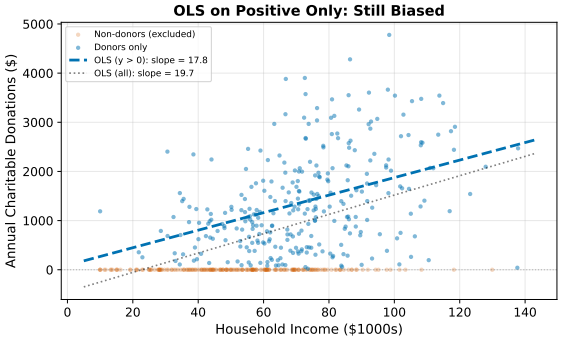
Maybe we should drop the zeros and run OLS on donors only?



OLS on donors only: slope $= 17.8$ (true $= 30$). Even worse!

OLS on Positive Observations Only: Still Biased

Maybe we should drop the zeros and run OLS on donors only?



OLS on donors only: slope $= 17.8$ (true $= 30$). Even worse!

The problem: by conditioning on $y > 0$, we have selected a non-random subsample. Among low-income households, the only donors are those with unusually large positive shocks. This **sample selection** distorts the income-donation relationship among the survivors.

Outline

Neither OLS Works: Where Does That Leave Us?

Approach	**Slope estimate**	**True slope**
OLS on all observations	19.7	30
OLS on positive only	17.8	30

Neither OLS Works: Where Does That Leave Us?

Approach	Slope estimate	True slope
OLS on all observations	19.7	30
OLS on positive only	17.8	30

OLS on all: treats the zeros as legitimate data points, pulling the slope toward zero.

Neither OLS Works: Where Does That Leave Us?

Approach	Slope estimate	True slope
OLS on all observations	19.7	30
OLS on positive only	17.8	30

OLS on all: treats the zeros as legitimate data points, pulling the slope toward zero.

OLS on positives: throws away 40% of the data and introduces sample selection bias.

Neither OLS Works: Where Does That Leave Us?

Approach	**Slope estimate**	**True slope**
OLS on all observations	19.7	30
OLS on positive only	17.8	30

OLS on all: treats the zeros as legitimate data points, pulling the slope toward zero.

OLS on positives: throws away 40% of the data and introduces sample selection bias.

$\Longrightarrow$ Both approaches ignore the **mechanism** that generates the zeros. We need a model that understands *why* some households donate zero.

What Would a Better Model Need?

A model for corner solution outcomes should:

What Would a Better Model Need?

A model for corner solution outcomes should:

1. **Explain the zeros:** some households *would* donate if they could, but their desired amount is negative (they are constrained to zero). This is what creates the pile-up that **attenuates the OLS-on-all slope** from 30 down to 19.7

What Would a Better Model Need?

A model for corner solution outcomes should:

1. **Explain the zeros:** some households *would* donate if they could, but their desired amount is negative (they are constrained to zero). This is what creates the pile-up that **attenuates the OLS-on-all slope** from 30 down to 19.7
2. **Use all the data:** both zeros and positives carry information about the income effect. Dropping the zeros introduces the **sample selection bias** that made OLS-on-positives even worse (17.8)

What Would a Better Model Need?

A model for corner solution outcomes should:

1. **Explain the zeros:** some households *would* donate if they could, but their desired amount is negative (they are constrained to zero). This is what creates the pile-up that **attenuates the OLS-on-all slope** from 30 down to 19.7
2. **Use all the data:** both zeros and positives carry information about the income effect. Dropping the zeros introduces the **sample selection bias** that made OLS-on-positives even worse (17.8)
3. **Recover the true slope:** the underlying relationship between income and desired donations, not the censored version OLS estimates

What Would a Better Model Need?

A model for corner solution outcomes should:

1. **Explain the zeros:** some households *would* donate if they could, but their desired amount is negative (they are constrained to zero). This is what creates the pile-up that **attenuates the OLS-on-all slope** from 30 down to 19.7
2. **Use all the data:** both zeros and positives carry information about the income effect. Dropping the zeros introduces the **sample selection bias** that made OLS-on-positives even worse (17.8)
3. **Recover the true slope:** the underlying relationship between income and desired donations, not the censored version OLS estimates

$\Longrightarrow$ We need a model that distinguishes between the **desired** outcome and the **observed** outcome. This is the latent variable idea you already know from logit/probit.

Outline

The Idea: Desired vs. Observed Donations

We saw two problems: the pile-up at zero attenuates the OLS slope, and dropping zeros introduces selection. Both stem from the same source: **we observe donations, not desires**.

The Idea: Desired vs. Observed Donations

We saw two problems: the pile-up at zero attenuates the OLS slope, and dropping zeros introduces selection. Both stem from the same source: **we observe donations, not desires**.

Imagine each household has a *desired* donation that could be positive or negative. A household with low income and no particular charitable inclination might "want" to donate −\$500 (they would take donations back if they could). But donations cannot be negative, so these households are constrained to zero.

The Idea: Desired vs. Observed Donations

We saw two problems: the pile-up at zero attenuates the OLS slope, and dropping zeros introduces selection. Both stem from the same source: **we observe donations, not desires**.

Imagine each household has a *desired* donation that could be positive or negative. A household with low income and no particular charitable inclination might "want" to donate −\$500 (they would take donations back if they could). But donations cannot be negative, so these households are constrained to zero.

If we could observe the desires directly, OLS would work perfectly: there would be no pile-up, no selection, just a clean linear relationship.

The Idea: Desired vs. Observed Donations

We saw two problems: the pile-up at zero attenuates the OLS slope, and dropping zeros introduces selection. Both stem from the same source: **we observe donations, not desires**.

Imagine each household has a *desired* donation that could be positive or negative. A household with low income and no particular charitable inclination might "want" to donate −\$500 (they would take donations back if they could). But donations cannot be negative, so these households are constrained to zero.

If we could observe the desires directly, OLS would work perfectly: there would be no pile-up, no selection, just a clean linear relationship.

$\Longrightarrow$ The Tobit model reconstructs those unobserved desires. It posits a **latent variable** for what each household *wants* to give, and an **observation rule** that censors negative desires to zero.

Latent Variable Framework

Define the **latent (unobserved)** variable y_i^* as each household's *desired* donations:

$$y_i^* = \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

Latent Variable Framework

Define the **latent (unobserved)** variable y_i^* as each household's *desired* donations:

$$y_i^* = \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

y_i^* can be **positive or negative**. A household with low income might have $y_i^* = -\$500$: they would "take back" donations if they could.

Latent Variable Framework

Define the **latent (unobserved)** variable y_i^* as each household's *desired* donations:

$$y_i^* = \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

y_i^* can be **positive or negative**. A household with low income might have $y_i^* = -\$500$: they would "take back" donations if they could.

But donations cannot be negative. The **observation rule** censors the latent variable:

$$y_i = \max(0,\ y_i^*) = \begin{cases} y_i^* & \text{if } y_i^* > 0 \quad \text{(donor)} \\ 0 & \text{if } y_i^* \leq 0 \quad \text{(non-donor, censored)} \end{cases}$$

Latent Variable Framework

Define the **latent (unobserved)** variable y_i^* as each household's *desired* donations:

$$y_i^* = \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

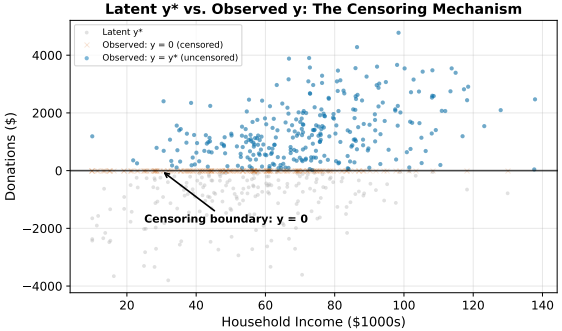
y_i^* can be **positive or negative**. A household with low income might have $y_i^* = -\$500$: they would "take back" donations if they could.

But donations cannot be negative. The **observation rule** censors the latent variable:

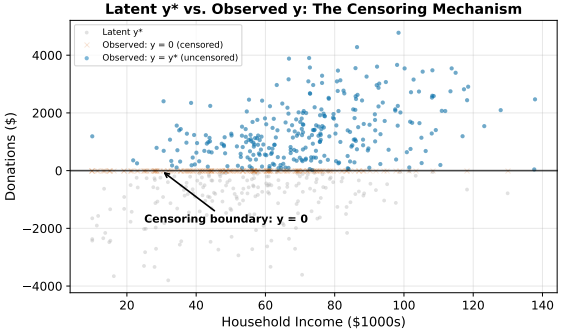
$$y_i = \max(0,\ y_i^*) = \begin{cases} y_i^* & \text{if } y_i^* > 0 \quad \text{(donor)} \\ 0 & \text{if } y_i^* \leq 0 \quad \text{(non-donor, censored)} \end{cases}$$

$\Longrightarrow$ The zeros are not real zeros. They are **censored observations** where the true desired donation is negative.

Seeing the Censoring Mechanism

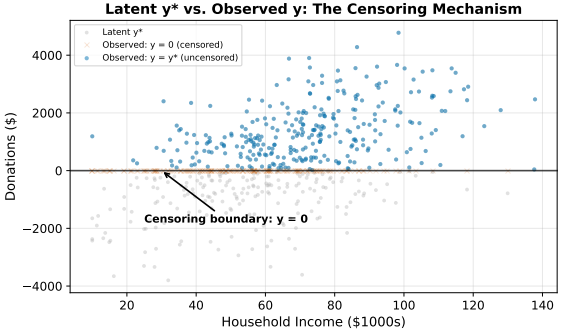


Seeing the Censoring Mechanism



Gray points: latent y_i^* (including negative values). Blue points: observed $y_i = y_i^*$ for donors. Orange crosses: observed $y_i = 0$ for non-donors.

Seeing the Censoring Mechanism



Gray points: latent y_i^* (including negative values). Blue points: observed $y_i = y_i^*$ for donors. Orange crosses: observed $y_i = 0$ for non-donors.

The censoring "folds" all negative latent values onto zero. This is what creates the pile-up, and this is what OLS cannot handle.

The Tobit Likelihood: Two Pieces

Since y_i comes from two different processes, the likelihood has two pieces.

The Tobit Likelihood: Two Pieces

Since y_i comes from two different processes, the likelihood has two pieces.

Recall: $\phi(\cdot)$ is the standard normal PDF and $\Phi(\cdot)$ is the standard normal CDF.

The Tobit Likelihood: Two Pieces

Since y_i comes from two different processes, the likelihood has two pieces.

Recall: $\phi(\cdot)$ is the standard normal PDF and $\Phi(\cdot)$ is the standard normal CDF.

Uncensored observations ($y_i > 0$): we observe the actual value, so the contribution is the normal density:

$$f(y_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - \beta_0 - \beta_1 \text{Income}_i - \beta_2 \text{Educ}_i - \beta_3 \text{Children}_i}{\sigma}\right)$$

The Tobit Likelihood: Two Pieces

Since y_i comes from two different processes, the likelihood has two pieces.

Recall: $\phi(\cdot)$ is the standard normal PDF and $\Phi(\cdot)$ is the standard normal CDF.

Uncensored observations ($y_i > 0$): we observe the actual value, so the contribution is the normal density:

$$f(y_i) = \frac{1}{\sigma}\,\phi\left(\frac{y_i - \beta_0 - \beta_1 \text{Income}_i - \beta_2 \text{Educ}_i - \beta_3 \text{Children}_i}{\sigma}\right)$$

Censored observations ($y_i = 0$): we only know $y_i^* \leq 0$, so the contribution is the probability of censoring:

$$P(y_i = 0) = P(y_i^* \leq 0) = \Phi\left(\frac{-(\beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Educ}_i + \beta_3 \text{Children}_i)}{\sigma}\right)$$

The Tobit Likelihood: Two Pieces

Since y_i comes from two different processes, the likelihood has two pieces.

Recall: $\phi(\cdot)$ is the standard normal PDF and $\Phi(\cdot)$ is the standard normal CDF.

Uncensored observations ($y_i > 0$): we observe the actual value, so the contribution is the normal density:

$$f(y_i) = \frac{1}{\sigma}\,\phi\left(\frac{y_i - \beta_0 - \beta_1 \text{Income}_i - \beta_2 \text{Educ}_i - \beta_3 \text{Children}_i}{\sigma}\right)$$

Censored observations ($y_i = 0$): we only know $y_i^* \leq 0$, so the contribution is the probability of censoring:

$$P(y_i = 0) = P(y_i^* \leq 0) = \Phi\left(\frac{-(\beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Educ}_i + \beta_3 \text{Children}_i)}{\sigma}\right)$$

Why the negative sign inside Φ? We need $P(\varepsilon_i \leq -\text{XB}_i)$. Flipping the sign converts $y_i^* \leq 0$ into a standard CDF evaluation.

The Log-Likelihood

Define the shorthand $\text{XB}_i \equiv \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i$ for the linear index.

The Log-Likelihood

Define the shorthand $\text{XB}_i \equiv \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i$ for the linear index.

The log-likelihood sums over two groups: first the uncensored observations (donors), then the censored observations (non-donors):

$$\ell = \sum_{i:\, y_i > 0} \left[\ln \phi\left(\frac{y_i - \text{XB}_i}{\sigma} \right) - \ln \sigma \right] + \sum_{i:\, y_i = 0} \ln \Phi\left(\frac{-\text{XB}_i}{\sigma} \right)$$

The Log-Likelihood

Define the shorthand $\text{XB}_i \equiv \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i$ for the linear index.

The log-likelihood sums over two groups: first the uncensored observations (donors), then the censored observations (non-donors):

$$\ell = \sum_{i:\, y_i > 0} \left[\ln \phi\left(\frac{y_i - \text{XB}_i}{\sigma} \right) - \ln \sigma \right] + \sum_{i:\, y_i = 0} \ln \Phi\left(\frac{-\text{XB}_i}{\sigma} \right)$$

This combines **two types of contributions**: the first sum handles the continuous part (like OLS with normal errors), and the second sum handles the discrete part (like probit). Unlike a true mixture model, we observe which group each household belongs to.

The Log-Likelihood

Define the shorthand $\text{XB}_i \equiv \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i$ for the linear index.

The log-likelihood sums over two groups: first the uncensored observations (donors), then the censored observations (non-donors):

$$\ell = \sum_{i:\, y_i > 0} \left[\ln \phi\left(\frac{y_i - \text{XB}_i}{\sigma} \right) - \ln \sigma \right] + \sum_{i:\, y_i = 0} \ln \Phi\left(\frac{-\text{XB}_i}{\sigma} \right)$$

This combines **two types of contributions**: the first sum handles the continuous part (like OLS with normal errors), and the second sum handles the discrete part (like probit). Unlike a true mixture model, we observe which group each household belongs to.

Software maximizes ℓ over $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma)$ numerically, just as in logit/probit.

The Log-Likelihood

Define the shorthand $\text{XB}_i \equiv \beta_0 + \beta_1 \,\text{Income}_i + \beta_2 \,\text{Educ}_i + \beta_3 \,\text{Children}_i$ for the linear index.

The log-likelihood sums over two groups: first the uncensored observations (donors), then the censored observations (non-donors):

$$\ell = \sum_{i:\ y_i > 0} \left[\ln \phi\left(\frac{y_i - \text{XB}_i}{\sigma} \right) - \ln \sigma \right] + \sum_{i:\ y_i = 0} \ln \Phi\left(\frac{-\text{XB}_i}{\sigma} \right)$$

This combines **two types of contributions**: the first sum handles the continuous part (like OLS with normal errors), and the second sum handles the discrete part (like probit). Unlike a true mixture model, we observe which group each household belongs to.

Software maximizes ℓ over $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma)$ numerically, just as in logit/probit.

$\Longrightarrow$ The Tobit likelihood combines a regression component and a binary component into a single model.

Numeric Example: Our Donation Data

Tobit MLE on the 500-household sample (true parameters in parentheses):

Numeric Example: Our Donation Data

Tobit MLE on the 500-household sample (true parameters in parentheses):

Parameter	**Tobit estimate**	**True value**
$\hat{\beta}_0$ (intercept)	-3324	-3500
$\hat{\beta}_1$ (Income, per \$1000)	32.3	30
$\hat{\beta}_2$ (Education, per year)	119	150
$\hat{\beta}_3$ (Children, per child)	-16	-100
$\hat{\sigma}$	1249	1200

Numeric Example: Our Donation Data

Tobit MLE on the 500-household sample (true parameters in parentheses):

Parameter	**Tobit estimate**	**True value**
$\hat{\beta}_0$ (intercept)	-3324	-3500
$\hat{\beta}_1$ (Income, per \$1000)	32.3	30
$\hat{\beta}_2$ (Education, per year)	119	150
$\hat{\beta}_3$ (Children, per child)	-16	-100
$\hat{\sigma}$	1249	1200

The income coefficient recovers **32.3** (true 30), much closer than OLS on all (19.7) or OLS on positives (17.8).

Numeric Example: Our Donation Data

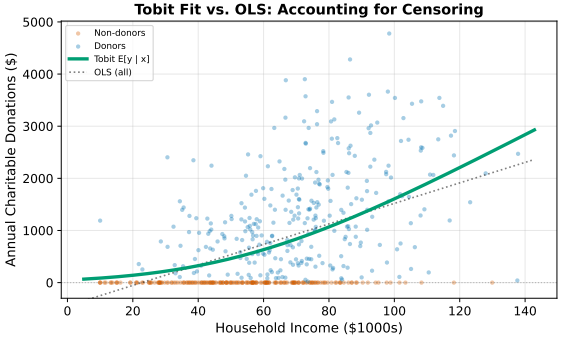
Tobit MLE on the 500-household sample (true parameters in parentheses):

Parameter	**Tobit estimate**	**True value**
$\hat{\beta}_0$ (intercept)	-3324	-3500
$\hat{\beta}_1$ (Income, per \$1000)	32.3	30
$\hat{\beta}_2$ (Education, per year)	119	150
$\hat{\beta}_3$ (Children, per child)	-16	-100
$\hat{\sigma}$	1249	1200

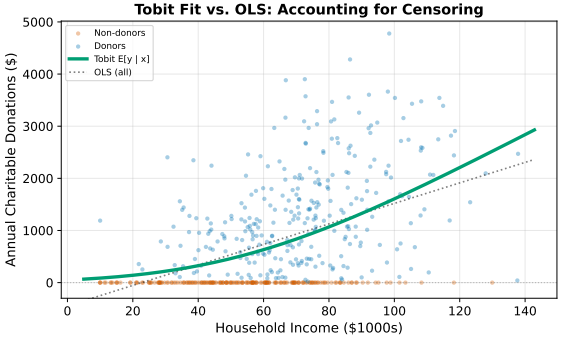
The income coefficient recovers **32.3** (true 30), much closer than OLS on all (19.7) or OLS on positives (17.8).

$\Longrightarrow$ By modeling the censoring mechanism explicitly, Tobit recovers the **latent** slope that OLS cannot.

Tobit Fit on the Data

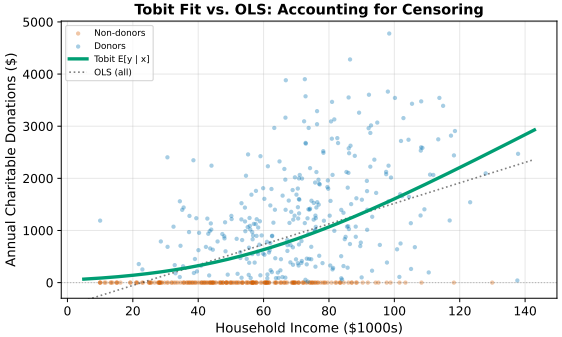


Tobit Fit on the Data



The green Tobit curve shows **unconditional** $\mathrm{E}[y_i \mid \mathrm{Income}_i]$ (including zeros), holding education and children at their means. OLS (dotted) misses the nonlinear shape entirely.

Tobit Fit on the Data



The green Tobit curve shows **unconditional** $\mathrm{E}[y_i \mid \mathrm{Income}_i]$ (including zeros), holding education and children at their means. OLS (dotted) misses the nonlinear shape entirely.

Notice the curve is flat near zero for low incomes (most households are censored, so additional income mainly shifts the *probability* of donating), then rises steeply through the transition region, and becomes approximately linear for high incomes (where nearly everyone donates).

Three Questions, Three Marginal Effects

In OLS, $\hat{\beta}_1$ is the marginal effect, full stop. In Tobit, $\hat{\beta}_1 = 32.3$ is the effect on the **latent** variable y^*. But we observe $y = \max(0, y^*)$, so different research questions call for different marginal effects:

Three Questions, Three Marginal Effects

In OLS, $\hat{\beta}_1$ is the marginal effect, full stop. In Tobit, $\hat{\beta}_1 = 32.3$ is the effect on the **latent** variable y^*. But we observe $y = \max(0, y^*)$, so different research questions call for different marginal effects:

1. **"What does the household want to give?"** The effect on latent desired donations y^*, including negative desires. Use this when you want the structural parameter of the underlying model

Three Questions, Three Marginal Effects

In OLS, $\hat{\beta}_1$ is the marginal effect, full stop. In Tobit, $\hat{\beta}_1 = 32.3$ is the effect on the **latent** variable y^*. But we observe $y = \max(0, y^*)$, so different research questions call for different marginal effects:

1. **"What does the household want to give?"** The effect on latent desired donations y^*, including negative desires. Use this when you want the structural parameter of the underlying model
2. **"Does an extra dollar of income bring new households into the donor pool?"** The effect on $P(y > 0)$, the extensive margin. Use this when the policy goal is to increase *participation*

Three Questions, Three Marginal Effects

In OLS, $\hat{\beta}_1$ is the marginal effect, full stop. In Tobit, $\hat{\beta}_1 = 32.3$ is the effect on the **latent** variable y^*. But we observe $y = \max(0, y^*)$, so different research questions call for different marginal effects:

1. **"What does the household want to give?"** The effect on latent desired donations y^*, including negative desires. Use this when you want the structural parameter of the underlying model
2. **"Does an extra dollar of income bring new households into the donor pool?"** The effect on $P(y > 0)$, the extensive margin. Use this when the policy goal is to increase *participation*
3. **"How much more does the average household actually give, including those stuck at zero?"** The effect on unconditional observed $\mathrm{E}[y]$. Use this when you want the overall impact on total giving

Three Questions, Three Marginal Effects

In OLS, $\hat{\beta}_1$ is the marginal effect, full stop. In Tobit, $\hat{\beta}_1 = 32.3$ is the effect on the **latent** variable y^*. But we observe $y = \max(0, y^*)$, so different research questions call for different marginal effects:

1. **"What does the household want to give?"** The effect on latent desired donations y^*, including negative desires. Use this when you want the structural parameter of the underlying model
2. **"Does an extra dollar of income bring new households into the donor pool?"** The effect on $P(y > 0)$, the extensive margin. Use this when the policy goal is to increase *participation*
3. **"How much more does the average household actually give, including those stuck at zero?"** The effect on unconditional observed $\mathrm{E}[y]$. Use this when you want the overall impact on total giving

$\Longrightarrow$ Each question has its own formula. Let's see them.

Marginal Effect 1: Latent Desired Donations

Question: What does the household *want* to give?

Marginal Effect 1: Latent Desired Donations

Question: What does the household *want* to give?

$$\frac{\partial \ \mathrm{E}[y_i^*]}{\partial \ \mathrm{Income}_i} = \beta_1 = 32.3$$

Marginal Effect 1: Latent Desired Donations

Question: What does the household *want* to give?

$$\frac{\partial \, \mathrm{E}[y_i^*]}{\partial \, \mathrm{Income}_i} = \beta_1 = 32.3$$

This is the simplest: the coefficient itself. It tells us the effect on the latent outcome, as if nobody were constrained.

Marginal Effect 1: Latent Desired Donations

Question: What does the household *want* to give?

$$\frac{\partial \, \mathrm{E}[y_i^*]}{\partial \, \mathrm{Income}_i} = \beta_1 = 32.3$$

This is the simplest: the coefficient itself. It tells us the effect on the latent outcome, as if nobody were constrained.

Use this when you are interested in the structural relationship between income and the desire to donate, ignoring the censoring constraint. This is the parameter OLS was trying (and failing) to estimate.

Marginal Effects 2 and 3: Probability and Observed Amount

Question 2: Does an extra dollar of income bring new donors?

Marginal Effects 2 and 3: Probability and Observed Amount

Question 2: Does an extra dollar of income bring new donors?

$$\frac{\partial\, P(y_i > 0)}{\partial\, \text{Income}_i} = \phi\left(\frac{\text{XB}_i}{\sigma}\right) \cdot \frac{\beta_1}{\sigma}$$

Use this when the policy goal is to expand the donor pool (extensive margin).

Marginal Effects 2 and 3: Probability and Observed Amount

Question 2: Does an extra dollar of income bring new donors?

$$\frac{\partial\, P(y_i > 0)}{\partial\, \text{Income}_i} = \phi\left(\frac{\text{XB}_i}{\sigma}\right) \cdot \frac{\beta_1}{\sigma}$$

Use this when the policy goal is to expand the donor pool (extensive margin).

Question 3: How much more does the average household actually give?

Marginal Effects 2 and 3: Probability and Observed Amount

Question 2: Does an extra dollar of income bring new donors?

$$\frac{\partial\, P(y_i > 0)}{\partial\, \text{Income}_i} = \phi\left(\frac{\text{XB}_i}{\sigma}\right) \cdot \frac{\beta_1}{\sigma}$$

Use this when the policy goal is to expand the donor pool (extensive margin).

Question 3: How much more does the average household actually give?

$$\frac{\partial\ \text{E}[y_i]}{\partial\, \text{Income}_i} = \Phi\left(\frac{\text{XB}_i}{\sigma}\right) \cdot \beta_1$$

Use this when you want the total impact on giving, combining both channels.

Marginal Effects 2 and 3: Probability and Observed Amount

Question 2: Does an extra dollar of income bring new donors?

$$\frac{\partial\, P(y_i > 0)}{\partial\, \text{Income}_i} = \phi\left(\frac{\text{XB}_i}{\sigma}\right) \cdot \frac{\beta_1}{\sigma}$$

Use this when the policy goal is to expand the donor pool (extensive margin).

Question 3: How much more does the average household actually give?

$$\frac{\partial\ \text{E}[y_i]}{\partial\, \text{Income}_i} = \Phi\left(\frac{\text{XB}_i}{\sigma}\right) \cdot \beta_1$$

Use this when you want the total impact on giving, combining both channels.

$\Longrightarrow$ Effects 2 and 3 depend on *where* you evaluate them (which household). A household far from the censoring threshold has different marginal effects than one near it.

Numeric Marginal Effects: Low-Income Household

Using $\hat{\beta}_1 = 32.3$ and $\hat{\sigma} = 1249$:

Numeric Marginal Effects: Low-Income Household

Using $\hat{\beta}_1 = 32.3$ and $\hat{\sigma} = 1249$:

Low-income household (Income $= 30$, Educ $= 14$, Children $= 2$):

- $\text{XB} = -3324 + 32.3 \times 30 + 119 \times 14 - 16 \times 2 = -721$
- $\text{XB}/\sigma = -721/1249 = -0.58$

Numeric Marginal Effects: Low-Income Household

Using $\hat{\beta}_1 = 32.3$ and $\hat{\sigma} = 1249$:

Low-income household (Income $= 30$, Educ $= 14$, Children $= 2$):

- $\text{XB} = -3324 + 32.3 \times 30 + 119 \times 14 - 16 \times 2 = -721$
- $\text{XB}/\sigma = -721/1249 = -0.58$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income

Numeric Marginal Effects: Low-Income Household

Using $\hat{\beta}_1 = 32.3$ and $\hat{\sigma} = 1249$:

Low-income household (Income $= 30$, Educ $= 14$, Children $= 2$):

- $\text{XB} = -3324 + 32.3 \times 30 + 119 \times 14 - 16 \times 2 = -721$
- $\text{XB}/\sigma = -721/1249 = -0.58$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income
- ME on $P(y > 0)$: $\phi(-0.58) \times 32.3/1249 = \mathbf{0.009}$ (0.9 percentage points per \$1000)

Numeric Marginal Effects: Low-Income Household

Using $\hat{\beta}_1 = 32.3$ and $\hat{\sigma} = 1249$:

Low-income household (Income $= 30$, Educ $= 14$, Children $= 2$):

- $XB = -3324 + 32.3 \times 30 + 119 \times 14 - 16 \times 2 = -721$
- $XB/\sigma = -721/1249 = -0.58$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income
- ME on $P(y > 0)$: $\phi(-0.58) \times 32.3/1249 = \mathbf{0.009}$ (0.9 percentage points per \$1000)
- ME on E$[y]$: $\Phi(-0.58) \times 32.3 = 0.28 \times 32.3 = \mathbf{\$9.1}$ per \$1000 income

Numeric Marginal Effects: Low-Income Household

Using $\hat{\beta}_1 = 32.3$ and $\hat{\sigma} = 1249$:

Low-income household (Income $= 30$, Educ $= 14$, Children $= 2$):

- $XB = -3324 + 32.3 \times 30 + 119 \times 14 - 16 \times 2 = -721$
- $XB/\sigma = -721/1249 = -0.58$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income
- ME on $P(y > 0)$: $\phi(-0.58) \times 32.3/1249 = \mathbf{0.009}$ (0.9 percentage points per \$1000)
- ME on E$[y]$: $\Phi(-0.58) \times 32.3 = 0.28 \times 32.3 = \mathbf{\$9.1}$ per \$1000 income

$\Longrightarrow$ For this household (only 28% chance of donating), most of the income effect is "absorbed" by the probability margin. The unconditional effect (\$9.1) is far below the latent effect (\$32.3).

Numeric Marginal Effects: High-Income Household

High-income household (Income $= 100$, Educ $= 16$, Children $= 1$):

- $\text{XB} = -3324 + 32.3 \times 100 + 119 \times 16 - 16 \times 1 = 1794$
- $\text{XB}/\sigma = 1794/1249 = 1.44$

Numeric Marginal Effects: High-Income Household

High-income household (Income $= 100$, Educ $= 16$, Children $= 1$):

- $\text{XB} = -3324 + 32.3 \times 100 + 119 \times 16 - 16 \times 1 = 1794$
- $\text{XB}/\sigma = 1794/1249 = 1.44$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income

Numeric Marginal Effects: High-Income Household

High-income household (Income $= 100$, Educ $= 16$, Children $= 1$):

- $\text{XB} = -3324 + 32.3 \times 100 + 119 \times 16 - 16 \times 1 = 1794$
- $\text{XB}/\sigma = 1794/1249 = 1.44$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income
- ME on $P(y > 0)$: $\phi(1.44) \times 32.3/1249 = \mathbf{0.004}$ (0.4 pp per \$1000)

Numeric Marginal Effects: High-Income Household

High-income household (Income $= 100$, Educ $= 16$, Children $= 1$):

- $\text{XB} = -3324 + 32.3 \times 100 + 119 \times 16 - 16 \times 1 = 1794$
- $\text{XB}/\sigma = 1794/1249 = 1.44$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income
- ME on $P(y > 0)$: $\phi(1.44) \times 32.3/1249 = \mathbf{0.004}$ (0.4 pp per \$1000)
- ME on E$[y]$: $\Phi(1.44) \times 32.3 = 0.925 \times 32.3 = \mathbf{\$29.9}$ per \$1000 income

Numeric Marginal Effects: High-Income Household

High-income household (Income $= 100$, Educ $= 16$, Children $= 1$):

- $\text{XB} = -3324 + 32.3 \times 100 + 119 \times 16 - 16 \times 1 = 1794$
- $\text{XB}/\sigma = 1794/1249 = 1.44$
- ME on latent y^*: $\beta_1 = \mathbf{32.3}$ dollars per \$1000 income
- ME on $P(y > 0)$: $\phi(1.44) \times 32.3/1249 = \mathbf{0.004}$ (0.4 pp per \$1000)
- ME on E$[y]$: $\Phi(1.44) \times 32.3 = 0.925 \times 32.3 = \$\mathbf{29.9}$ per \$1000 income

$\Longrightarrow$ This household is nearly certain to donate (92%), so almost all of the latent effect passes through to the observed outcome. The unconditional ME (\$29.9) is close to the raw coefficient (\$32.3).

Comparing Marginal Effects Across Households

	Latent y^*	$P(y > 0)$	E$[y]$
Low-income (XB$/\sigma = -0.58$)	\$32.3	0.9 pp	\$9.1
High-income (XB$/\sigma = 1.44$)	\$32.3	0.4 pp	\$29.9

Comparing Marginal Effects Across Households

	Latent y^*	$P(y > 0)$	E$[y]$
Low-income (XB$/\sigma = -0.58$)	\$32.3	0.9 pp	\$9.1
High-income (XB$/\sigma = 1.44$)	\$32.3	0.4 pp	\$29.9

The latent effect is always the same (\$32.3). But the observed effects differ dramatically:

Comparing Marginal Effects Across Households

	Latent y^*	$P(y > 0)$	E[y]
Low-income ($XB/\sigma = -0.58$)	\$32.3	0.9 pp	\$9.1
High-income ($XB/\sigma = 1.44$)	\$32.3	0.4 pp	\$29.9

The latent effect is always the same (\$32.3). But the observed effects differ dramatically:

- For the low-income household, income mainly affects *whether* they donate (extensive margin)
- For the high-income household, income mainly affects *how much* they donate (intensive margin)

Comparing Marginal Effects Across Households

	Latent y^*	$P(y > 0)$	E[y]
Low-income ($XB/\sigma = -0.58$)	\$32.3	0.9 pp	\$9.1
High-income ($XB/\sigma = 1.44$)	\$32.3	0.4 pp	\$29.9

The latent effect is always the same (\$32.3). But the observed effects differ dramatically:

- For the low-income household, income mainly affects *whether* they donate (extensive margin)
- For the high-income household, income mainly affects *how much* they donate (intensive margin)

$\Longrightarrow$ Where a household sits relative to the censoring threshold determines which margin dominates.

Boundary Condition: When Tobit Reduces to OLS

Look at the unconditional marginal effect formula again:

$$\frac{\partial\, \mathrm{E}[y_i]}{\partial\, x_k} = \Phi\left(\frac{\mathrm{XB}_i}{\sigma}\right) \cdot \beta_k$$

Boundary Condition: When Tobit Reduces to OLS

Look at the unconditional marginal effect formula again:

$$\frac{\partial\, \mathrm{E}[y_i]}{\partial\, x_k} = \Phi\left(\frac{\mathrm{XB}_i}{\sigma}\right) \cdot \beta_k$$

When $\mathbf{XB}_i/\sigma$ is very large (almost everyone donates):

- $\Phi(\mathrm{XB}_i/\sigma) \to 1$
- The marginal effect $\to \beta_k$
- Tobit behaves like OLS

Boundary Condition: When Tobit Reduces to OLS

Look at the unconditional marginal effect formula again:

$$\frac{\partial\, \mathrm{E}[y_i]}{\partial\, x_k} = \Phi\left(\frac{\mathrm{XB}_i}{\sigma}\right) \cdot \beta_k$$

When XB_i/σ is very large (almost everyone donates):

- $\Phi(\mathrm{XB}_i/\sigma) \to 1$
- The marginal effect $\to \beta_k$
- Tobit behaves like OLS

When XB_i/σ is very negative (almost no one donates):

- $\Phi(\mathrm{XB}_i/\sigma) \to 0$
- The marginal effect $\to 0$
- Income increases mostly change the *probability* of donating, not the amount

Boundary Condition: When Tobit Reduces to OLS

Look at the unconditional marginal effect formula again:

$$\frac{\partial\, \mathrm{E}[y_i]}{\partial\, x_k} = \Phi\left(\frac{\mathrm{XB}_i}{\sigma}\right) \cdot \beta_k$$

When XB_i/σ is very large (almost everyone donates):

- $\Phi(\mathrm{XB}_i/\sigma) \to 1$
- The marginal effect $\to \beta_k$
- Tobit behaves like OLS

When XB_i/σ is very negative (almost no one donates):

- $\Phi(\mathrm{XB}_i/\sigma) \to 0$
- The marginal effect $\to 0$
- Income increases mostly change the *probability* of donating, not the amount

$\Longrightarrow$ If censoring is rare (few zeros), Tobit and OLS give similar answers. The more censoring, the more Tobit differs from OLS.

The McDonald-Moffitt Decomposition

A policymaker wants to increase total charitable giving. Should they target existing donors to give more (intensive margin), or convert non-donors into donors (extensive margin)? The McDonald-Moffitt decomposition answers exactly this.

The McDonald-Moffitt Decomposition

A policymaker wants to increase total charitable giving. Should they target existing donors to give more (intensive margin), or convert non-donors into donors (extensive margin)? The McDonald-Moffitt decomposition answers exactly this.

McDonald and Moffitt (1980) showed that the unconditional marginal effect (Effect #3 from two slides ago) can be split into two channels via the product rule:

The McDonald-Moffitt Decomposition

A policymaker wants to increase total charitable giving. Should they target existing donors to give more (intensive margin), or convert non-donors into donors (extensive margin)? The McDonald-Moffitt decomposition answers exactly this.

McDonald and Moffitt (1980) showed that the unconditional marginal effect (Effect #3 from two slides ago) can be split into two channels via the product rule:

$$\underbrace{\frac{\partial\, \mathrm{E}[y]}{\partial\, x_k}}_{\text{total effect}} = \underbrace{P(y>0) \cdot \frac{\partial\, \mathrm{E}[y \mid y>0]}{\partial\, x_k}}_{\text{how much more do donors give?}} + \underbrace{\mathrm{E}[y \mid y>0] \cdot \frac{\partial\, P(y>0)}{\partial\, x_k}}_{\text{how many new donors?}}$$

The McDonald-Moffitt Decomposition

A policymaker wants to increase total charitable giving. Should they target existing donors to give more (intensive margin), or convert non-donors into donors (extensive margin)? The McDonald-Moffitt decomposition answers exactly this.

McDonald and Moffitt (1980) showed that the unconditional marginal effect (Effect #3 from two slides ago) can be split into two channels via the product rule:

$$\underbrace{\frac{\partial\, \mathrm{E}[y]}{\partial\, x_k}}_{\text{total effect}} = \underbrace{P(y>0) \cdot \frac{\partial\, \mathrm{E}[y \mid y>0]}{\partial\, x_k}}_{\text{how much more do donors give?}} + \underbrace{\mathrm{E}[y \mid y>0] \cdot \frac{\partial\, P(y>0)}{\partial\, x_k}}_{\text{how many new donors?}}$$

Channel 1 (intensive): Among current donors, income raises donations by \$X.
Channel 2 (extensive): Higher income brings new households into the donor pool, each contributing $\mathrm{E}[y \mid y>0]$.

The McDonald-Moffitt Decomposition

A policymaker wants to increase total charitable giving. Should they target existing donors to give more (intensive margin), or convert non-donors into donors (extensive margin)? The McDonald-Moffitt decomposition answers exactly this.

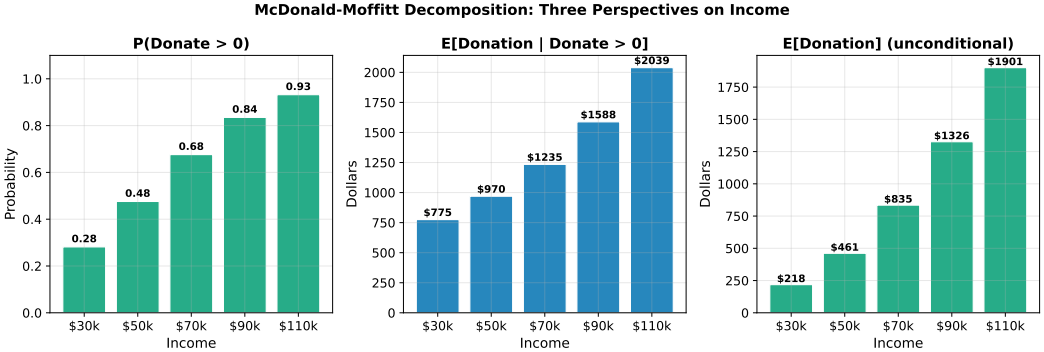
McDonald and Moffitt (1980) showed that the unconditional marginal effect (Effect #3 from two slides ago) can be split into two channels via the product rule:

$$\underbrace{\frac{\partial\, \mathrm{E}[y]}{\partial\, x_k}}_{\text{total effect}} = \underbrace{P(y>0) \cdot \frac{\partial\, \mathrm{E}[y \mid y>0]}{\partial\, x_k}}_{\text{how much more do donors give?}} + \underbrace{\mathrm{E}[y \mid y>0] \cdot \frac{\partial\, P(y>0)}{\partial\, x_k}}_{\text{how many new donors?}}$$

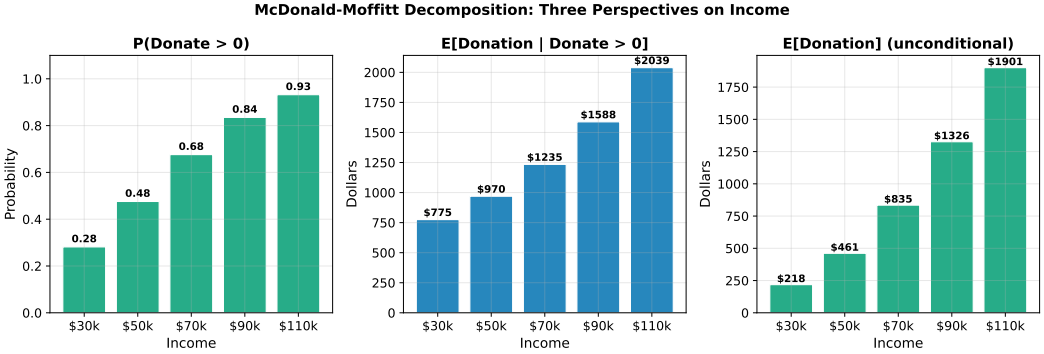
Channel 1 (intensive): Among current donors, income raises donations by \$X.
Channel 2 (extensive): Higher income brings new households into the donor pool, each contributing $\mathrm{E}[y \mid y>0]$.

$\Longrightarrow$ A single coefficient β_1 drives *both* the probability of participation and the level of giving. This is the defining feature (and restriction) of the Tobit model.

Visualizing the Decomposition



Visualizing the Decomposition



At Income = \$30k, only 28% donate; at \$110k, 93% donate. Among donors, average giving rises from \$775 to \$2,039. Both channels contribute to unconditional E[y] rising from \$218 to \$1,901.

Outline

Censoring vs. Truncation: They Sound Similar but Differ

These two terms describe different data situations. Consider the same donation study:

Censoring vs. Truncation: They Sound Similar but Differ

These two terms describe different data situations. Consider the same donation study:

Censoring (Tobit):

- The researcher surveys *all* 500 households
- For non-donors, she records $y_i = 0$ along with their income, education, and children
- She knows the non-donors exist and can count them

Censoring vs. Truncation: They Sound Similar but Differ

These two terms describe different data situations. Consider the same donation study:

Censoring (Tobit):

- The researcher surveys *all* 500 households
- For non-donors, she records $y_i = 0$ along with their income, education, and children
- She knows the non-donors exist and can count them

Truncation:

- The researcher only has records from a charity's donor database
- She observes the 299 donors ($y_i > 0$) and their characteristics
- Non-donors are *completely absent*: she does not even know how many there are

Censoring vs. Truncation: They Sound Similar but Differ

These two terms describe different data situations. Consider the same donation study:

Censoring (Tobit):

- The researcher surveys *all* 500 households
- For non-donors, she records $y_i = 0$ along with their income, education, and children
- She knows the non-donors exist and can count them

Truncation:

- The researcher only has records from a charity's donor database
- She observes the 299 donors ($y_i > 0$) and their characteristics
- Non-donors are *completely absent*: she does not even know how many there are

$\Longrightarrow$ With censoring, the zeros are **in the data**. With truncation, the zeros are **missing entirely**.

When Do You Face Each Situation?

Censored examples (use Tobit):

- Charitable donations: non-donors report \$0, all households surveyed
- Hours worked: non-workers report 0 hours, all individuals in the sample
- Expenditure on durable goods: many households spend \$0 on new cars

When Do You Face Each Situation?

Censored examples (use Tobit):

- Charitable donations: non-donors report \$0, all households surveyed
- Hours worked: non-workers report 0 hours, all individuals in the sample
- Expenditure on durable goods: many households spend \$0 on new cars

Truncated examples (use truncated regression):

- Firm profits: only firms that survived to be surveyed appear; failed firms are gone
- Scholarship amounts: only recipients are in the data; rejected applicants are absent
- Wages: only observed for employed workers; non-workers absent from the data. (In practice, wages are usually better handled by the **Heckman selection model**, since the decision to work is a separate process from wage determination.)

When Do You Face Each Situation?

Censored examples (use Tobit):

- Charitable donations: non-donors report \$0, all households surveyed
- Hours worked: non-workers report 0 hours, all individuals in the sample
- Expenditure on durable goods: many households spend \$0 on new cars

Truncated examples (use truncated regression):

- Firm profits: only firms that survived to be surveyed appear; failed firms are gone
- Scholarship amounts: only recipients are in the data; rejected applicants are absent
- Wages: only observed for employed workers; non-workers absent from the data. (In practice, wages are usually better handled by the **Heckman selection model**, since the decision to work is a separate process from wage determination.)

$\Longrightarrow$ If the zeros are in your data, it is censoring. If you only see the positive values and do not know how many zeros were excluded, it is truncation.

Outline

The Tobit Model Rests on Two Assumptions

Assumption 1: Normal, homoskedastic errors.

$$\varepsilon_i \sim N(0, \sigma^2), \quad \text{independent of covariates}$$

The Tobit Model Rests on Two Assumptions

Assumption 1: Normal, homoskedastic errors.

$$\varepsilon_i \sim N(0, \sigma^2), \quad \text{independent of covariates}$$

Two ways this can fail:

- **Non-normality:** skewed or heavy-tailed errors
- **Heteroskedasticity:** $\text{Var}(\varepsilon_i)$ depends on covariates. Unlike OLS, where heteroskedasticity only affects standard errors, in Tobit it makes the **coefficient estimates themselves inconsistent**

The Tobit Model Rests on Two Assumptions

Assumption 1: Normal, homoskedastic errors.

$$\varepsilon_i \sim N(0, \sigma^2), \quad \text{independent of covariates}$$

Two ways this can fail:

- **Non-normality:** skewed or heavy-tailed errors
- **Heteroskedasticity:** $\text{Var}(\varepsilon_i)$ depends on covariates. Unlike OLS, where heteroskedasticity only affects standard errors, in Tobit it makes the **coefficient estimates themselves inconsistent**

Assumption 2: Same mechanism governs participation and amount.
The same β_1 determines both *whether* a household donates and *how much*. This is a strong restriction.

The Tobit Model Rests on Two Assumptions

Assumption 1: Normal, homoskedastic errors.

$$\varepsilon_i \sim N(0, \sigma^2), \quad \text{independent of covariates}$$

Two ways this can fail:

- **Non-normality:** skewed or heavy-tailed errors
- **Heteroskedasticity:** $\text{Var}(\varepsilon_i)$ depends on covariates. Unlike OLS, where heteroskedasticity only affects standard errors, in Tobit it makes the **coefficient estimates themselves inconsistent**

Assumption 2: Same mechanism governs participation and amount.
The same β_1 determines both *whether* a household donates and *how much*. This is a strong restriction.

Example where this fails: a very wealthy household may decide whether to donate based on social pressure ($\beta_1^{\text{participate}}$), but the amount depends on tax incentives ($\beta_1^{\text{amount}} \neq \beta_1^{\text{participate}}$).

The Tobit Model Rests on Two Assumptions

Assumption 1: Normal, homoskedastic errors.

$$\varepsilon_i \sim N(0, \sigma^2), \quad \text{independent of covariates}$$

Two ways this can fail:

- **Non-normality:** skewed or heavy-tailed errors
- **Heteroskedasticity:** $\text{Var}(\varepsilon_i)$ depends on covariates. Unlike OLS, where heteroskedasticity only affects standard errors, in Tobit it makes the **coefficient estimates themselves inconsistent**

Assumption 2: Same mechanism governs participation and amount.
The same β_1 determines both *whether* a household donates and *how much*. This is a strong restriction.

Example where this fails: a very wealthy household may decide whether to donate based on social pressure ($\beta_1^{\text{participate}}$), but the amount depends on tax incentives ($\beta_1^{\text{amount}} \neq \beta_1^{\text{participate}}$).

$\Longrightarrow$ When Assumption 2 fails, the Tobit model forces a single coefficient to represent two distinct processes. What alternatives exist?

When Assumption 2 Fails: Alternatives

Two-Part Model (also called the "hurdle" model):

- Part 1: Probit or logit for $P(y > 0)$ with coefficients γ
- Part 2: OLS or truncated regression for $\mathrm{E}[y \mid y > 0]$ with coefficients δ
- γ and δ are estimated **separately**, so the participation and amount decisions can differ

When Assumption 2 Fails: Alternatives

Two-Part Model (also called the "hurdle" model):

- Part 1: Probit or logit for $P(y > 0)$ with coefficients γ
- Part 2: OLS or truncated regression for $\mathrm{E}[y \mid y > 0]$ with coefficients δ
- γ and δ are estimated **separately**, so the participation and amount decisions can differ

Heckman Selection Model:

- For situations where the zeros are not corner solutions but **sample selection**
- Example: wages are only observed for people who choose to work; the decision to work is a separate equation
- Adds a selection correction (inverse Mills ratio) to the outcome equation

When Assumption 2 Fails: Alternatives

Two-Part Model (also called the "hurdle" model):

- Part 1: Probit or logit for $P(y > 0)$ with coefficients γ
- Part 2: OLS or truncated regression for $\mathrm{E}[y \mid y > 0]$ with coefficients δ
- γ and δ are estimated **separately**, so the participation and amount decisions can differ

Heckman Selection Model:

- For situations where the zeros are not corner solutions but **sample selection**
- Example: wages are only observed for people who choose to work; the decision to work is a separate equation
- Adds a selection correction (inverse Mills ratio) to the outcome equation

$\Longrightarrow$ Tobit assumes one mechanism. The two-part model relaxes that. The Heckman model is for a fundamentally different problem: selection, not censoring.

Decision Flowchart: Choosing the Right Model

1. Is your outcome non-negative with a pile-up at zero?

Decision Flowchart: Choosing the Right Model

1. Is your outcome non-negative with a pile-up at zero?
2. **Are the zeros corner solutions?** (The person *wants* a negative value but is constrained to zero.)
   - Yes, and you believe the same β governs both participation and amount $\implies$ **Tobit**

Decision Flowchart: Choosing the Right Model

1. Is your outcome non-negative with a pile-up at zero?
2. **Are the zeros corner solutions?** (The person *wants* a negative value but is constrained to zero.)
   - Yes, and you believe the same β governs both participation and amount $\implies$ **Tobit**
   - Yes, but participation and amount may have different drivers $\implies$ **Two-Part Model**

Decision Flowchart: Choosing the Right Model

1. Is your outcome non-negative with a pile-up at zero?
2. **Are the zeros corner solutions?** (The person *wants* a negative value but is constrained to zero.)
   - Yes, and you believe the same β governs both participation and amount $\implies$ **Tobit**
   - Yes, but participation and amount may have different drivers $\implies$ **Two-Part Model**
3. **Are the zeros from sample selection?** (The outcome *exists* but you do not observe it.)
   - A worker has a wage, but you only observe it if they participate in the labor market $\implies$ **Heckman Selection Model**

Decision Flowchart: Choosing the Right Model

1. Is your outcome non-negative with a pile-up at zero?
2. **Are the zeros corner solutions?** (The person *wants* a negative value but is constrained to zero.)
   - Yes, and you believe the same β governs both participation and amount $\implies$ **Tobit**
   - Yes, but participation and amount may have different drivers $\implies$ **Two-Part Model**
3. **Are the zeros from sample selection?** (The outcome *exists* but you do not observe it.)
   - A worker has a wage, but you only observe it if they participate in the labor market $\implies$ **Heckman Selection Model**
4. **Is your outcome a count?** (Non-negative integers: 0, 1, 2, ...)
   - $\implies$ **Poisson / Negative Binomial**, not Tobit

Outline

Summary: Back to Charitable Donations

1. **The data problem:** 40% of households donate $0, creating a spike at zero with a continuous positive tail. This is a **corner solution**

Summary: Back to Charitable Donations

1. **The data problem:** 40% of households donate \$0, creating a spike at zero with a continuous positive tail. This is a **corner solution**
2. **OLS fails twice:** OLS on all data gives slope $= 19.7$ (attenuated by the pile-up). OLS on positives gives slope $= 17.8$ (distorted by sample selection). True latent slope $= 30$

Summary: Back to Charitable Donations

1. **The data problem:** 40% of households donate \$0, creating a spike at zero with a continuous positive tail. This is a **corner solution**
2. **OLS fails twice:** OLS on all data gives slope $= 19.7$ (attenuated by the pile-up). OLS on positives gives slope $= 17.8$ (distorted by sample selection). True latent slope $= 30$
3. **The Tobit model** uses a latent variable $y^* = \beta_0 + \beta_1 \text{Income} + \cdots + \varepsilon$ with $y = \max(0, y^*)$. Its likelihood combines a probit component (zeros) and a regression component (positives)

Summary: Back to Charitable Donations

1. **The data problem:** 40% of households donate \$0, creating a spike at zero with a continuous positive tail. This is a **corner solution**
2. **OLS fails twice:** OLS on all data gives slope $= 19.7$ (attenuated by the pile-up). OLS on positives gives slope $= 17.8$ (distorted by sample selection). True latent slope $= 30$
3. **The Tobit model** uses a latent variable $y^* = \beta_0 + \beta_1 \text{Income} + \cdots + \varepsilon$ with $y = \max(0, y^*)$. Its likelihood combines a probit component (zeros) and a regression component (positives)
4. **Tobit recovers the latent slope:** $\hat{\beta}_1 = 32.3 \approx 30$. Three types of marginal effects tell you the effect on latent y^*, on $P(y > 0)$, and on $\text{E}[y]$

Summary: Back to Charitable Donations

1. **The data problem:** 40% of households donate \$0, creating a spike at zero with a continuous positive tail. This is a **corner solution**
2. **OLS fails twice:** OLS on all data gives slope = 19.7 (attenuated by the pile-up). OLS on positives gives slope = 17.8 (distorted by sample selection). True latent slope = 30
3. **The Tobit model** uses a latent variable $y^* = \beta_0 + \beta_1 \text{Income} + \cdots + \varepsilon$ with $y = \max(0, y^*)$. Its likelihood combines a probit component (zeros) and a regression component (positives)
4. **Tobit recovers the latent slope:** $\hat{\beta}_1 = 32.3 \approx 30$. Three types of marginal effects tell you the effect on latent y^*, on $P(y > 0)$, and on $\text{E}[y]$
5. **The McDonald-Moffitt decomposition** splits the total effect into a participation channel (new donors) and an amount channel (existing donors give more)

Summary: Back to Charitable Donations

1. **The data problem:** 40% of households donate \$0, creating a spike at zero with a continuous positive tail. This is a **corner solution**
2. **OLS fails twice:** OLS on all data gives slope = 19.7 (attenuated by the pile-up). OLS on positives gives slope = 17.8 (distorted by sample selection). True latent slope = 30
3. **The Tobit model** uses a latent variable $y^* = \beta_0 + \beta_1 \text{Income} + \cdots + \varepsilon$ with $y = \max(0, y^*)$. Its likelihood combines a probit component (zeros) and a regression component (positives)
4. **Tobit recovers the latent slope:** $\hat{\beta}_1 = 32.3 \approx 30$. Three types of marginal effects tell you the effect on latent y^*, on $P(y > 0)$, and on $\text{E}[y]$
5. **The McDonald-Moffitt decomposition** splits the total effect into a participation channel (new donors) and an amount channel (existing donors give more)
6. **Tobit assumes one mechanism** for participation and amount. When that fails, use a two-part model. When zeros come from selection rather than censoring, use Heckman

Comparison Table: Tobit vs. Alternatives

	Tobit	**Two-Part**	**Heckman**
Zero mechanism	corner solution	corner solution	selection
Same β for participation & amount?	yes	no (separate γ, δ)	no (separate equations)
Normality required?	yes	only if Part 2 uses truncated regression	yes (or semi-parametric)
Typical example	donations, hours worked	health expenditures	wages (if employed)

Comparison Table: Tobit vs. Alternatives

	Tobit	**Two-Part**	**Heckman**
Zero mechanism	corner solution	corner solution	selection
Same β for participation & amount?	yes	no (separate γ, δ)	no (separate equations)
Normality required?	yes	only if Part 2 uses truncated regression	yes (or semi-parametric)
Typical example	donations, hours worked	health expenditures	wages (if employed)

$\Longrightarrow$ The correct choice depends on the economic mechanism generating the zeros: censoring from a corner solution, or selection from a missing-data process.

Thank you!
jakeanderson@g.ucla.edu

The Heckman Selection Model

When Your Data Only Includes People Who Showed Up

Jake Anderson

May 16, 2026

Outline

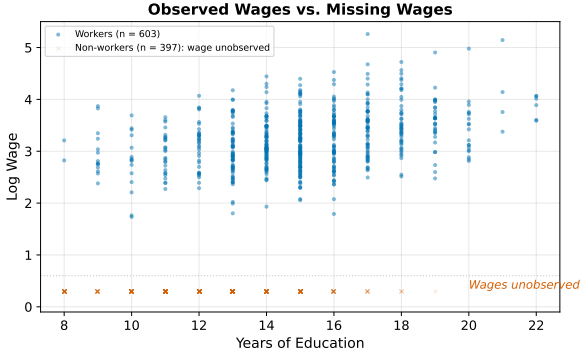
The Data: Wages and Education

A labor economist surveys **1,000 adults** and records their hourly wage, years of education, work experience, number of children, and spouse's income.

The Data: Wages and Education

A labor economist surveys **1,000 adults** and records their hourly wage, years of education, work experience, number of children, and spouse's income.

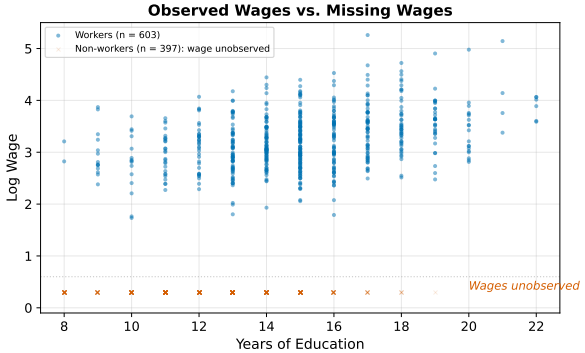
She wants to estimate the **return to education**: how much does one more year of schooling raise log wages? There is a catch: **only 603 people work**. The other 397 have no wage to observe.



The Data: Wages and Education

A labor economist surveys **1,000 adults** and records their hourly wage, years of education, work experience, number of children, and spouse's income.

She wants to estimate the **return to education**: how much does one more year of schooling raise log wages? There is a catch: **only 603 people work**. The other 397 have no wage to observe.



Blue dots: workers with observed wages. Orange crosses: non-workers with *no wage data*.

Who Are the Non-Workers?

The non-workers are not a random sample of the population:

Who Are the Non-Workers?

The non-workers are not a random sample of the population:

	Workers (n = 603)	**Non-Workers** (n = 397)
Mean education (years)	15.0	12.3
Mean children	0.9	1.6
Mean spouse income ($1000s)	36.2	44.7

Who Are the Non-Workers?

The non-workers are not a random sample of the population:

	Workers (n = 603)	**Non-Workers** (n = 397)
Mean education (years)	15.0	12.3
Mean children	0.9	1.6
Mean spouse income ($1000s)	36.2	44.7

Non-workers have **less education**, **more children**, and **higher-earning spouses**. The people with missing wages are systematically different from those with observed wages.

Who Are the Non-Workers?

The non-workers are not a random sample of the population:

	Workers (n = 603)	**Non-Workers** (n = 397)
Mean education (years)	15.0	12.3
Mean children	0.9	1.6
Mean spouse income ($1000s)	36.2	44.7

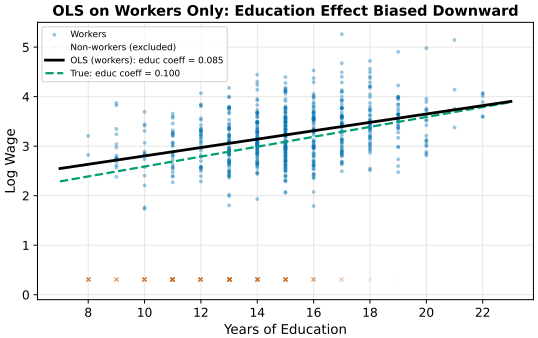
Non-workers have **less education**, **more children**, and **higher-earning spouses**. The people with missing wages are systematically different from those with observed wages.

$\Longrightarrow$ This is **sample selection**: the decision to work is not random, and it correlates with the outcome we care about (wages).

OLS on Workers Only: Biased

Ignoring the missing data and running OLS on the 603 workers:

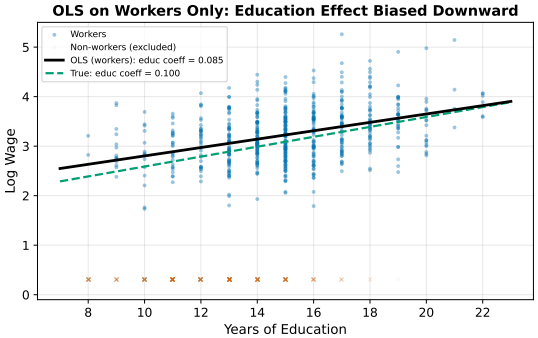
$$\widehat{\log(\text{wage})}_i = 1.373 + \underbrace{0.085}_{\substack{\text{biased downward} \\ \text{in this setting}}} \cdot \text{educ}_i + 0.040 \cdot \text{exper}_i$$



OLS on Workers Only: Biased

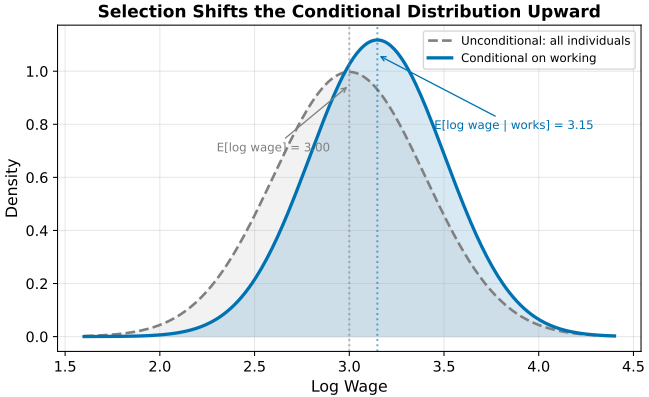
Ignoring the missing data and running OLS on the 603 workers:

$$\widehat{\log(\text{wage})}_i = 1.373 + \underbrace{0.085}_{\substack{\text{biased downward} \\ \text{in this setting}}} \cdot \text{educ}_i + 0.040 \cdot \text{exper}_i$$

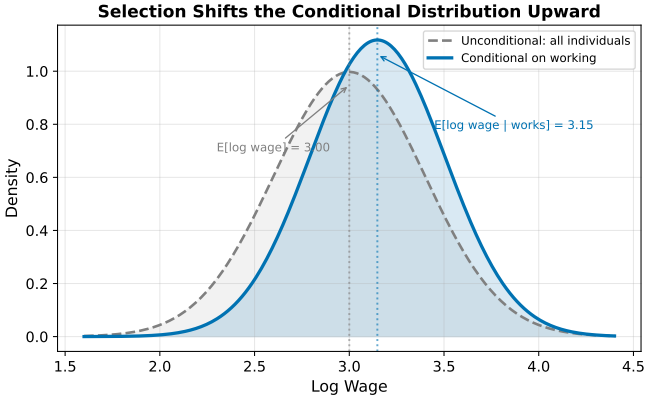


The OLS education coefficient is **0.085** (true = 0.100). That is a 15% underestimate. Why?

What OLS Misses: The Conditional Distribution

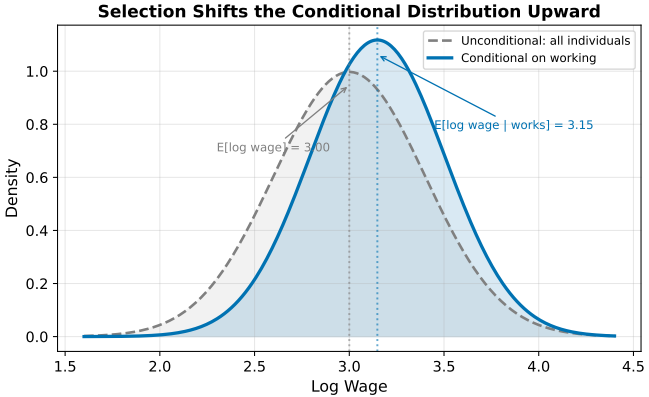


What OLS Misses: The Conditional Distribution



Gray: wage distribution for *all* individuals. Blue: *conditional on working*. Selection shifts the distribution to the right.

What OLS Misses: The Conditional Distribution



Gray: wage distribution for *all* individuals. Blue: *conditional on working*. Selection shifts the distribution to the right.

OLS on workers fits a line through the blue distribution, but the true regression line passes through the gray one. The shift gets absorbed into the intercept and correlated coefficients, biasing them.

What If We Include Non-Workers?

One natural idea: assign wage $= 0$ (or $\log(\text{wage}) = 0$) to non-workers and run OLS on all 1,000 people.

What If We Include Non-Workers?

One natural idea: assign wage $= 0$ (or $\log(\text{wage}) = 0$) to non-workers and run OLS on all 1,000 people.

This is also wrong:

- Non-workers do not have a wage of zero. They have a **missing** wage. Imputing zero creates artificial data points that pull the regression line toward zero for people with low education

What If We Include Non-Workers?

One natural idea: assign wage $= 0$ (or $\log(\text{wage}) = 0$) to non-workers and run OLS on all 1,000 people.

This is also wrong:

- Non-workers do not have a wage of zero. They have a **missing** wage. Imputing zero creates artificial data points that pull the regression line toward zero for people with low education
- The resulting slope reflects a mix of two relationships: the true effect of education on wages *and* the effect of education on whether someone works at all

What If We Include Non-Workers?

One natural idea: assign wage $= 0$ (or $\log(\text{wage}) = 0$) to non-workers and run OLS on all 1,000 people.

This is also wrong:

- Non-workers do not have a wage of zero. They have a **missing** wage. Imputing zero creates artificial data points that pull the regression line toward zero for people with low education
- The resulting slope reflects a mix of two relationships: the true effect of education on wages *and* the effect of education on whether someone works at all
- If you set missing wages to any fixed number, you change the distribution of the dependent variable in a way that depends on the selection process

What If We Include Non-Workers?

One natural idea: assign wage $= 0$ (or $\log(\text{wage}) = 0$) to non-workers and run OLS on all 1,000 people.

This is also wrong:

- Non-workers do not have a wage of zero. They have a **missing** wage. Imputing zero creates artificial data points that pull the regression line toward zero for people with low education
- The resulting slope reflects a mix of two relationships: the true effect of education on wages *and* the effect of education on whether someone works at all
- If you set missing wages to any fixed number, you change the distribution of the dependent variable in a way that depends on the selection process

$\Longrightarrow$ Neither dropping non-workers (OLS on workers) nor imputing values fixes the problem. We need a model that **explicitly accounts for the selection process**.

The Basketball Player Analogy

Suppose you want to estimate the effect of **height on free-throw accuracy** in the general population.

The Basketball Player Analogy

Suppose you want to estimate the effect of **height on free-throw accuracy** in the general population.

But you only observe people who **play basketball**. Who plays? Mostly tall people. And among tall people, the ones who play are not randomly selected: they are the ones with basketball talent, which also helps free throws.

The Basketball Player Analogy

Suppose you want to estimate the effect of **height on free-throw accuracy** in the general population.

But you only observe people who **play basketball**. Who plays? Mostly tall people. And among tall people, the ones who play are not randomly selected: they are the ones with basketball talent, which also helps free throws.

The result: among basketball players, height looks less important for free throws, because everyone is already tall and talented. The height effect is attenuated by selection into the sample.

The Basketball Player Analogy

Suppose you want to estimate the effect of **height on free-throw accuracy** in the general population.

But you only observe people who **play basketball**. Who plays? Mostly tall people. And among tall people, the ones who play are not randomly selected: they are the ones with basketball talent, which also helps free throws.

The result: among basketball players, height looks less important for free throws, because everyone is already tall and talented. The height effect is attenuated by selection into the sample.

$\Longrightarrow$ Our wage data has the same problem. Workers are not a random draw: they are the people whose unobserved characteristics (motivation, ability) pushed them into the labor force. These same characteristics also affect wages.

What Would a Better Model Need?

OLS on workers fails because it ignores the selection process. A better approach should:

What Would a Better Model Need?

OLS on workers fails because it ignores the selection process. A better approach should:

1. **Model the selection:** why some people work and others do not. The non-workers carry information about the selection process, even though they have no wages

What Would a Better Model Need?

OLS on workers fails because it ignores the selection process. A better approach should:

1. **Model the selection:** why some people work and others do not. The non-workers carry information about the selection process, even though they have no wages
2. **Correct the wage equation:** the workers we observe are not representative. Their unobserved characteristics are systematically different from the population average

What Would a Better Model Need?

OLS on workers fails because it ignores the selection process. A better approach should:

1. **Model the selection:** why some people work and others do not. The non-workers carry information about the selection process, even though they have no wages
2. **Correct the wage equation:** the workers we observe are not representative. Their unobserved characteristics are systematically different from the population average
3. **Recover the true return to education:** the structural relationship between education and wages, free from selection bias

What Would a Better Model Need?

OLS on workers fails because it ignores the selection process. A better approach should:

1. **Model the selection:** why some people work and others do not. The non-workers carry information about the selection process, even though they have no wages
2. **Correct the wage equation:** the workers we observe are not representative. Their unobserved characteristics are systematically different from the population average
3. **Recover the true return to education:** the structural relationship between education and wages, free from selection bias

$\Longrightarrow$ We need to jointly model the wage process and the selection process. This is different from Tobit (censoring): here the **decision to work is a separate equation** from the wage itself.

Outline

Two Equations, Two Error Terms

The Heckman model has two equations, each with its own error term.

Two Equations, Two Error Terms

The Heckman model has two equations, each with its own error term.

Wage equation (outcome, only observed for workers):

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \, \text{educ}_i + \beta_2 \, \text{exper}_i + u_i$$

$$\beta_0 = 1.0, \quad \beta_1 = 0.10, \quad \beta_2 = 0.04$$

Two Equations, Two Error Terms

The Heckman model has two equations, each with its own error term.

Wage equation (outcome, only observed for workers):

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \, \text{educ}_i + \beta_2 \, \text{exper}_i + u_i$$

$$\beta_0 = 1.0, \quad \beta_1 = 0.10, \quad \beta_2 = 0.04$$

Selection equation (determines who works):

$$\text{work}_i^* = \gamma_0 + \gamma_1 \, \text{educ}_i + \gamma_2 \, \text{kids}_i + \gamma_3 \, \text{spouse\_inc}_i + v_i$$

$$\gamma_0 = -2.5, \quad \gamma_1 = 0.30, \quad \gamma_2 = -0.50, \quad \gamma_3 = -0.02$$

Two Equations, Two Error Terms

The Heckman model has two equations, each with its own error term.

Wage equation (outcome, only observed for workers):

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + u_i$$

$$\beta_0 = 1.0, \quad \beta_1 = 0.10, \quad \beta_2 = 0.04$$

Selection equation (determines who works):

$$\text{work}_i^* = \gamma_0 + \gamma_1 \,\text{educ}_i + \gamma_2 \,\text{kids}_i + \gamma_3 \,\text{spouse\_inc}_i + v_i$$

$$\gamma_0 = -2.5, \quad \gamma_1 = 0.30, \quad \gamma_2 = -0.50, \quad \gamma_3 = -0.02$$

work_i^* is an **unobserved latent variable**: person i works if $\text{work}_i^* > 0$, and we observe $\log(\text{wage}_i)$ only when that happens.

Two Equations, Two Error Terms

The Heckman model has two equations, each with its own error term.

Wage equation (outcome, only observed for workers):

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + u_i$$

$$\beta_0 = 1.0, \quad \beta_1 = 0.10, \quad \beta_2 = 0.04$$

Selection equation (determines who works):

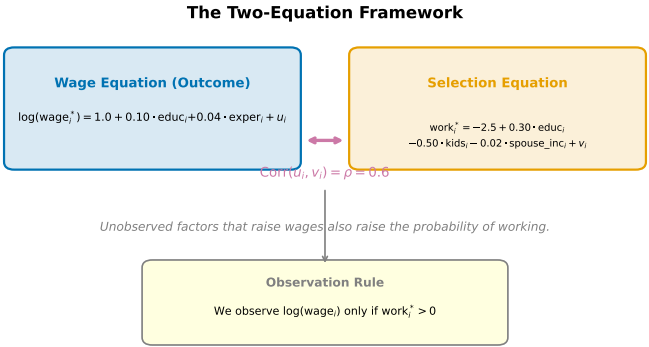
$$\text{work}_i^* = \gamma_0 + \gamma_1 \,\text{educ}_i + \gamma_2 \,\text{kids}_i + \gamma_3 \,\text{spouse\_inc}_i + v_i$$

$$\gamma_0 = -2.5, \quad \gamma_1 = 0.30, \quad \gamma_2 = -0.50, \quad \gamma_3 = -0.02$$

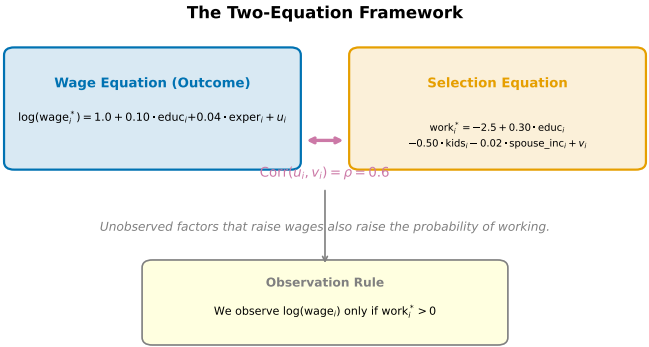
work_i^* is an **unobserved latent variable**: person i works if $\text{work}_i^* > 0$, and we observe $\log(\text{wage}_i)$ only when that happens.

$\Longrightarrow$ If u_i and v_i are correlated ($\rho \neq 0$), workers are a **selected** subsample and OLS on workers is biased.

Where the Bias Comes From: Correlated Errors

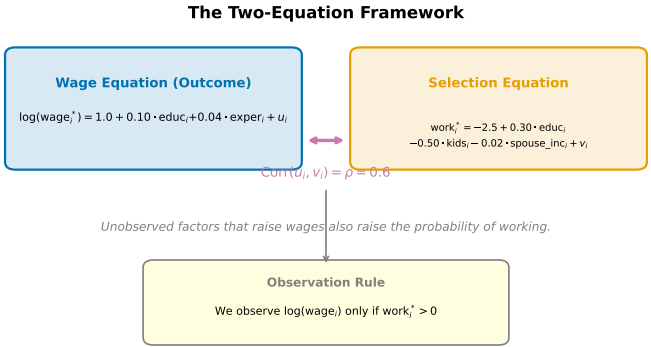


Where the Bias Comes From: Correlated Errors



In our DGP, $\rho = \text{Corr}(u_i, v_i) = 0.6 > 0$. Unobserved factors that raise wages (ability, motivation) *also* make a person more likely to work. Among workers, the average u_i is **positive**, not zero.

Where the Bias Comes From: Correlated Errors



In our DGP, $\rho = \text{Corr}(u_i, v_i) = 0.6 > 0$. Unobserved factors that raise wages (ability, motivation) *also* make a person more likely to work. Among workers, the average u_i is **positive**, not zero.

$\implies$ OLS assumes $\text{E}[u_i \mid \text{works}] = 0$, but in reality $\text{E}[u_i \mid \text{works}] > 0$. This violates the zero conditional mean assumption.

The Conditional Expectation: What Does Selection Do?

What is the expected wage of a worker, given that she chose to work?

The Conditional Expectation: What Does Selection Do?

What is the expected wage of a worker, given that she chose to work?

Population regression (without conditioning on selection):

$$\mathrm{E}[\log(\mathrm{wage}_i) \mid X_i] = \beta_0 + \beta_1 \,\mathrm{educ}_i + \beta_2 \,\mathrm{exper}_i$$

The Conditional Expectation: What Does Selection Do?

What is the expected wage of a worker, given that she chose to work?

Population regression (without conditioning on selection):
$\mathsf{E}[\log(\text{wage}_i) \mid X_i] = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i$

Selection means v_i was large enough for the person to work. Conditioning on this:

$$\mathsf{E}[\log(\text{wage}_i) \mid \text{works}] = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + \underbrace{\mathsf{E}[u_i \mid v_i > -\gamma_0 - \gamma_1 \text{educ}_i - \cdots]}_{\text{selection bias term}}$$

The Conditional Expectation: What Does Selection Do?

What is the expected wage of a worker, given that she chose to work?

Population regression (without conditioning on selection):
$\mathrm{E}[\log(\text{wage}_i) \mid X_i] = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i$

Selection means v_i was large enough for the person to work. Conditioning on this:

$$\mathrm{E}[\log(\text{wage}_i) \mid \text{works}] = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + \underbrace{\mathrm{E}[u_i \mid v_i > -\gamma_0 - \gamma_1 \text{educ}_i - \cdots]}_{\text{selection bias term}}$$

Because u_i and v_i are correlated, knowing that someone works (v_i is large enough) tells us something about their wage error (u_i). This "something" is the selection bias term.

The Conditional Expectation: What Does Selection Do?

What is the expected wage of a worker, given that she chose to work?

Population regression (without conditioning on selection):
$\mathrm{E}[\log(\text{wage}_i) \mid X_i] = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i$

Selection means v_i was large enough for the person to work. Conditioning on this:

$$\mathrm{E}[\log(\text{wage}_i) \mid \text{works}] = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + \underbrace{\mathrm{E}[u_i \mid v_i > -\gamma_0 - \gamma_1 \text{educ}_i - \cdots]}_{\text{selection bias term}}$$

Because u_i and v_i are correlated, knowing that someone works (v_i is large enough) tells us something about their wage error (u_i). This "something" is the selection bias term.

$\implies$ If we can calculate this term and include it in our regression, we remove the bias.

The Selection Index and Normal Distribution Notation

Before deriving the correction formula, we need two pieces of notation.

The Selection Index and Normal Distribution Notation

Before deriving the correction formula, we need two pieces of notation.

The selection index. Define the shorthand:

$$\text{GZ}_i \equiv \gamma_0 + \gamma_1 \,\text{educ}_i + \gamma_2 \,\text{kids}_i + \gamma_3 \,\text{spouse\_inc}_i$$

This is the linear combination from the selection equation (everything except the error v_i). A larger GZ_i means person i is more likely to work.

The Selection Index and Normal Distribution Notation

Before deriving the correction formula, we need two pieces of notation.

The selection index. Define the shorthand:

$$\text{GZ}_i \equiv \gamma_0 + \gamma_1 \,\text{educ}_i + \gamma_2 \,\text{kids}_i + \gamma_3 \,\text{spouse\_inc}_i$$

This is the linear combination from the selection equation (everything except the error v_i). A larger GZ_i means person i is more likely to work.

Standard normal PDF and CDF. Recall from your statistics courses:

- $\phi(z)$: the standard normal **density** (PDF) evaluated at z. Bell-curve height
- $\Phi(z)$: the standard normal **cumulative distribution** (CDF) evaluated at z. Area to the left of z

The Selection Index and Normal Distribution Notation

Before deriving the correction formula, we need two pieces of notation.

The selection index. Define the shorthand:

$$\text{GZ}_i \equiv \gamma_0 + \gamma_1 \,\text{educ}_i + \gamma_2 \,\text{kids}_i + \gamma_3 \,\text{spouse\_inc}_i$$

This is the linear combination from the selection equation (everything except the error v_i). A larger GZ_i means person i is more likely to work.

Standard normal PDF and CDF. Recall from your statistics courses:

- $\phi(z)$: the standard normal **density** (PDF) evaluated at z. Bell-curve height
- $\Phi(z)$: the standard normal **cumulative distribution** (CDF) evaluated at z. Area to the left of z

$\Longrightarrow$ In the probit model, $P(\text{work}_i = 1) = \Phi(\text{GZ}_i)$. Both ϕ and Φ will appear in the selection correction formula.

The Inverse Mills Ratio

For bivariate normal errors (u_i, v_i), the selection bias term has a closed form:

$$\mathrm{E}[u_i \mid \text{works}] = \rho\,\sigma_u \cdot \underbrace{\frac{\phi(\mathrm{GZ}_i)}{\Phi(\mathrm{GZ}_i)}}_{\equiv\ \lambda_i\ \text{(inverse Mills ratio)}}$$

The Inverse Mills Ratio

For bivariate normal errors (u_i, v_i), the selection bias term has a closed form:

$$\mathrm{E}[u_i \mid \text{works}] = \rho\,\sigma_u \cdot \underbrace{\frac{\phi(\mathrm{GZ}_i)}{\Phi(\mathrm{GZ}_i)}}_{\equiv\ \lambda_i\ \text{(inverse Mills ratio)}}$$

The IMR $\lambda_i = \phi(\mathrm{GZ}_i)/\Phi(\mathrm{GZ}_i)$ measures how strongly selection affects each individual. It depends only on the selection index, not on the wage equation.

The Inverse Mills Ratio

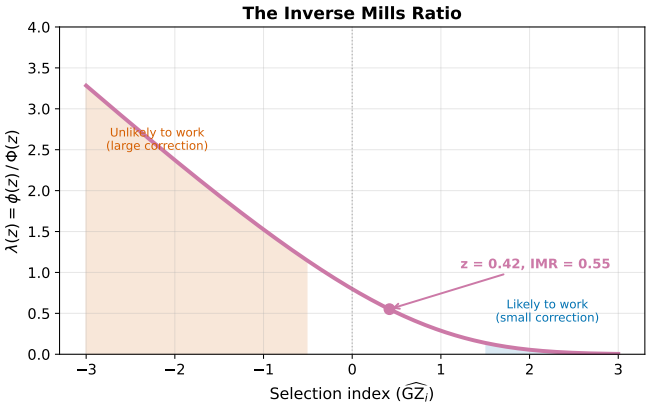
For bivariate normal errors (u_i, v_i), the selection bias term has a closed form:

$$\mathrm{E}[u_i \mid \text{works}] = \rho\,\sigma_u \cdot \underbrace{\frac{\phi(\mathrm{GZ}_i)}{\Phi(\mathrm{GZ}_i)}}_{\equiv\ \lambda_i\ \text{(inverse Mills ratio)}}$$

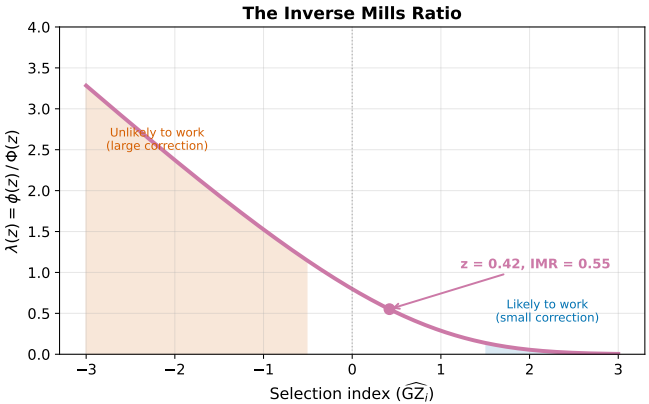
The IMR $\lambda_i = \phi(\mathrm{GZ}_i)/\Phi(\mathrm{GZ}_i)$ measures how strongly selection affects each individual. It depends only on the selection index, not on the wage equation.

$\Longrightarrow$ Connection to Tobit: you saw the IMR in the conditional expectation $\mathrm{E}[y \mid y > 0]$. Same mathematical object, different context. In Tobit it corrects for censoring; here it corrects for selection.

How the IMR Works

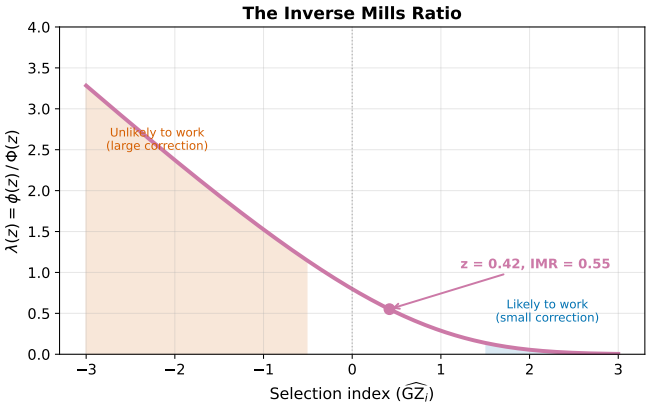


How the IMR Works



When GZ_i is small (unlikely to work), the IMR is **large**: if she works despite low predicted probability, her unobserved characteristics must be unusually favorable.

How the IMR Works



When GZ_i is small (unlikely to work), the IMR is **large**: if she works despite low predicted probability, her unobserved characteristics must be unusually favorable.

When GZ_i is large (very likely to work), the IMR is **small**: working tells us little about her unobservables. The correction is minimal.

The Corrected Wage Equation

Substituting the IMR into the conditional expectation:

$$\mathsf{E}[\log(\mathsf{wage}_i) \mid \mathsf{works}] = \beta_0 + \beta_1 \,\mathsf{educ}_i + \beta_2 \,\mathsf{exper}_i + \underbrace{\rho\sigma_u}_{\equiv\,\delta} \cdot \lambda_i$$

The Corrected Wage Equation

Substituting the IMR into the conditional expectation:

$$\mathrm{E}[\log(\text{wage}_i) \mid \text{works}] = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + \underbrace{\rho\sigma_u}_{\equiv\,\delta} \cdot \lambda_i$$

We combine ρ and σ_u into a single parameter $\delta = \rho\sigma_u$ because the two-step procedure cannot separately identify them: the second-stage OLS only estimates the product, not the individual components. Then:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + \delta \cdot \lambda_i + \text{error}$$

The Corrected Wage Equation

Substituting the IMR into the conditional expectation:

$$\mathrm{E}[\log(\mathrm{wage}_i) \mid \mathrm{works}] = \beta_0 + \beta_1\,\mathrm{educ}_i + \beta_2\,\mathrm{exper}_i + \underbrace{\rho\sigma_u}_{\equiv\,\delta} \cdot \lambda_i$$

We combine ρ and σ_u into a single parameter $\delta = \rho\sigma_u$ because the two-step procedure cannot separately identify them: the second-stage OLS only estimates the product, not the individual components. Then:

$$\log(\mathrm{wage}_i) = \beta_0 + \beta_1\,\mathrm{educ}_i + \beta_2\,\mathrm{exper}_i + \delta \cdot \lambda_i + \mathrm{error}$$

The error in this equation is heteroskedastic (its variance depends on GZ_i). This is why the Step 2 standard errors need correction.

The Corrected Wage Equation

Substituting the IMR into the conditional expectation:

$$\mathrm{E}[\log(\text{wage}_i) \mid \text{works}] = \beta_0 + \beta_1\,\text{educ}_i + \beta_2\,\text{exper}_i + \underbrace{\rho\sigma_u}_{\equiv\,\delta} \cdot \lambda_i$$

We combine ρ and σ_u into a single parameter $\delta = \rho\sigma_u$ because the two-step procedure cannot separately identify them: the second-stage OLS only estimates the product, not the individual components. Then:

$$\log(\text{wage}_i) = \beta_0 + \beta_1\,\text{educ}_i + \beta_2\,\text{exper}_i + \delta \cdot \lambda_i + \text{error}$$

The error in this equation is heteroskedastic (its variance depends on GZ_i). This is why the Step 2 standard errors need correction.

This is an OLS regression with one additional variable: λ_i. If we knew the γ coefficients (from the selection equation), we could compute λ_i for each worker and run this regression.

The Corrected Wage Equation

Substituting the IMR into the conditional expectation:

$$\mathrm{E}[\log(\text{wage}_i) \mid \text{works}] = \beta_0 + \beta_1\,\text{educ}_i + \beta_2\,\text{exper}_i + \underbrace{\rho\sigma_u}_{\equiv\,\delta} \cdot \lambda_i$$

We combine ρ and σ_u into a single parameter $\delta = \rho\sigma_u$ because the two-step procedure cannot separately identify them: the second-stage OLS only estimates the product, not the individual components. Then:

$$\log(\text{wage}_i) = \beta_0 + \beta_1\,\text{educ}_i + \beta_2\,\text{exper}_i + \delta \cdot \lambda_i + \text{error}$$

The error in this equation is heteroskedastic (its variance depends on GZ_i). This is why the Step 2 standard errors need correction.

This is an OLS regression with one additional variable: λ_i. If we knew the γ coefficients (from the selection equation), we could compute λ_i for each worker and run this regression.

$\Longrightarrow$ This is the logic behind the Heckman two-step procedure.

Estimation Roadmap

Where we stand:

Estimation Roadmap

Where we stand:

1. We showed that OLS on workers is biased because $\mathrm{E}[u_i \mid \text{works}] \neq 0$

Estimation Roadmap

Where we stand:

1. We showed that OLS on workers is biased because $\mathrm{E}[u_i \mid \text{works}] \neq 0$
2. We derived that the bias equals $\rho\sigma_u \cdot \lambda_i$, where λ_i is the inverse Mills ratio

Estimation Roadmap

Where we stand:

1. We showed that OLS on workers is biased because $\mathrm{E}[u_i \mid \text{works}] \neq 0$
2. We derived that the bias equals $\rho\sigma_u \cdot \lambda_i$, where λ_i is the inverse Mills ratio
3. We showed that if we add λ_i to the wage regression, we can recover the true β coefficients

Estimation Roadmap

Where we stand:

1. We showed that OLS on workers is biased because $\mathrm{E}[u_i \mid \text{works}] \neq 0$
2. We derived that the bias equals $\rho\sigma_u \cdot \lambda_i$, where λ_i is the inverse Mills ratio
3. We showed that if we add λ_i to the wage regression, we can recover the true β coefficients

The remaining problem: computing λ_i requires the selection coefficients $\gamma_0, \gamma_1, \gamma_2, \gamma_3$, which we do not know.

Estimation Roadmap

Where we stand:

1. We showed that OLS on workers is biased because $\mathrm{E}[u_i \mid \text{works}] \neq 0$
2. We derived that the bias equals $\rho\sigma_u \cdot \lambda_i$, where λ_i is the inverse Mills ratio
3. We showed that if we add λ_i to the wage regression, we can recover the true β coefficients

The remaining problem: computing λ_i requires the selection coefficients $\gamma_0, \gamma_1, \gamma_2, \gamma_3$, which we do not know.

$\Longrightarrow$ We need to estimate the γ coefficients first. This suggests a **two-step procedure**: (1) estimate the selection equation, (2) use the estimated $\hat{\lambda}_i$ in the wage regression.

Outline

Step 1: Estimate the Selection Equation (Probit)

Run a **probit** on all 1,000 observations (workers and non-workers):

$$P(\text{work}_i = 1) = \Phi(\gamma_0 + \gamma_1\,\text{educ}_i + \gamma_2\,\text{kids}_i + \gamma_3\,\text{spouse\_inc}_i)$$

Step 1: Estimate the Selection Equation (Probit)

Run a **probit** on all 1,000 observations (workers and non-workers):

$$P(\text{work}_i = 1) = \Phi(\gamma_0 + \gamma_1\,\text{educ}_i + \gamma_2\,\text{kids}_i + \gamma_3\,\text{spouse\_inc}_i)$$

Parameter	**Probit estimate**	**True value**
$\hat{\gamma}_0$ (intercept)	-2.629	-2.5
$\hat{\gamma}_1$ (education)	0.314	0.3
$\hat{\gamma}_2$ (children)	-0.512	-0.5
$\hat{\gamma}_3$ (spouse income)	-0.019	-0.02

Step 1: Estimate the Selection Equation (Probit)

Run a **probit** on all 1,000 observations (workers and non-workers):

$$P(\text{work}_i = 1) = \Phi(\gamma_0 + \gamma_1\,\text{educ}_i + \gamma_2\,\text{kids}_i + \gamma_3\,\text{spouse\_inc}_i)$$

Parameter	**Probit estimate**	**True value**
$\hat{\gamma}_0$ (intercept)	-2.629	-2.5
$\hat{\gamma}_1$ (education)	0.314	0.3
$\hat{\gamma}_2$ (children)	-0.512	-0.5
$\hat{\gamma}_3$ (spouse income)	-0.019	-0.02

More education $\Longrightarrow$ more likely to work. More children, higher spouse income $\Longrightarrow$ less likely to work.

Step 1: Estimate the Selection Equation (Probit)

Run a **probit** on all 1,000 observations (workers and non-workers):

$$P(\text{work}_i = 1) = \Phi(\gamma_0 + \gamma_1\,\text{educ}_i + \gamma_2\,\text{kids}_i + \gamma_3\,\text{spouse\_inc}_i)$$

Parameter	**Probit estimate**	**True value**
$\hat{\gamma}_0$ (intercept)	-2.629	-2.5
$\hat{\gamma}_1$ (education)	0.314	0.3
$\hat{\gamma}_2$ (children)	-0.512	-0.5
$\hat{\gamma}_3$ (spouse income)	-0.019	-0.02

More education $\Longrightarrow$ more likely to work. More children, higher spouse income $\Longrightarrow$ less likely to work.

From these estimates, compute the estimated selection index $\widehat{\text{GZ}}_i$ for every individual.

Step 1 Continued: Compute the IMR

For each worker i, compute the inverse Mills ratio using the estimated selection index $\widehat{\mathrm{GZ}}_i$:

$$\hat{\lambda}_i = \frac{\phi(\widehat{\mathrm{GZ}}_i)}{\Phi(\widehat{\mathrm{GZ}}_i)}$$

Step 1 Continued: Compute the IMR

For each worker i, compute the inverse Mills ratio using the estimated selection index $\widehat{\mathrm{GZ}}_i$:

$$\hat{\lambda}_i = \frac{\phi(\widehat{\mathrm{GZ}}_i)}{\Phi(\widehat{\mathrm{GZ}}_i)}$$

Numeric example: a woman with 16 years of education, 2 children, spouse earning \$50k:

$$\begin{aligned}\widehat{\mathrm{GZ}} &= -2.629 + 0.314 \times 16 + (-0.512) \times 2 + (-0.019) \times 50 \\ &= -2.629 + 5.024 - 1.024 - 0.950 = 0.421\end{aligned}$$

Step 1 Continued: Compute the IMR

For each worker i, compute the inverse Mills ratio using the estimated selection index $\widehat{\text{GZ}}_i$:

$$\hat{\lambda}_i = \frac{\phi(\widehat{\text{GZ}}_i)}{\Phi(\widehat{\text{GZ}}_i)}$$

Numeric example: a woman with 16 years of education, 2 children, spouse earning \$50k:

$$\begin{aligned}\widehat{\text{GZ}} &= -2.629 + 0.314 \times 16 + (-0.512) \times 2 + (-0.019) \times 50 \\ &= -2.629 + 5.024 - 1.024 - 0.950 = 0.421\end{aligned}$$

- $P(\text{works}) = \Phi(0.421) = 0.663$ (66% chance of working)
- $\hat{\lambda} = \phi(0.421)/\Phi(0.421) = 0.365/0.663 = 0.551$

Step 1 Continued: Compute the IMR

For each worker i, compute the inverse Mills ratio using the estimated selection index $\widehat{\text{GZ}}_i$:

$$\hat{\lambda}_i = \frac{\phi(\widehat{\text{GZ}}_i)}{\Phi(\widehat{\text{GZ}}_i)}$$

Numeric example: a woman with 16 years of education, 2 children, spouse earning \$50k:

$$\begin{aligned}\widehat{\text{GZ}} &= -2.629 + 0.314 \times 16 + (-0.512) \times 2 + (-0.019) \times 50 \\ &= -2.629 + 5.024 - 1.024 - 0.950 = 0.421\end{aligned}$$

- $P(\text{works}) = \Phi(0.421) = 0.663$ (66% chance of working)
- $\hat{\lambda} = \phi(0.421)/\Phi(0.421) = 0.365/0.663 = 0.551$

$\Longrightarrow$ This worker has a moderate selection correction. If she were nearly certain to work, $\hat{\lambda}$ would be close to zero.

Step 2: OLS with the IMR as an Extra Regressor

Run OLS on the **603 workers**, adding $\hat{\lambda}_i$ as an additional regressor:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + \delta \,\hat{\lambda}_i + \text{error}$$

Step 2: OLS with the IMR as an Extra Regressor

Run OLS on the **603 workers**, adding $\hat{\lambda}_i$ as an additional regressor:

$$\log(\text{wage}_i) = \beta_0 + \beta_1\,\text{educ}_i + \beta_2\,\text{exper}_i + \delta\,\hat{\lambda}_i + \text{error}$$

Parameter	**OLS (workers)**	**Heckman**	**True**
$\hat{\beta}_0$ (intercept)	1.373	0.874	1.0
$\hat{\beta}_1$ (education)	0.085	**0.111**	0.100
$\hat{\beta}_2$ (experience)	0.040	0.039	0.040
$\hat{\delta}$ (IMR)	–	0.260	0.240

Step 2: OLS with the IMR as an Extra Regressor

Run OLS on the **603 workers**, adding $\hat{\lambda}_i$ as an additional regressor:

$$\log(\text{wage}_i) = \beta_0 + \beta_1\,\text{educ}_i + \beta_2\,\text{exper}_i + \delta\,\hat{\lambda}_i + \text{error}$$

Parameter	OLS (workers)	Heckman	True
$\hat{\beta}_0$ (intercept)	1.373	0.874	1.0
$\hat{\beta}_1$ (education)	0.085	**0.111**	0.100
$\hat{\beta}_2$ (experience)	0.040	0.039	0.040
$\hat{\delta}$ (IMR)	–	0.260	0.240

The Heckman education coefficient (**0.111**) is much closer to the true value (0.100) than OLS on workers (0.085).

Step 2: OLS with the IMR as an Extra Regressor

Run OLS on the **603 workers**, adding $\hat{\lambda}_i$ as an additional regressor:

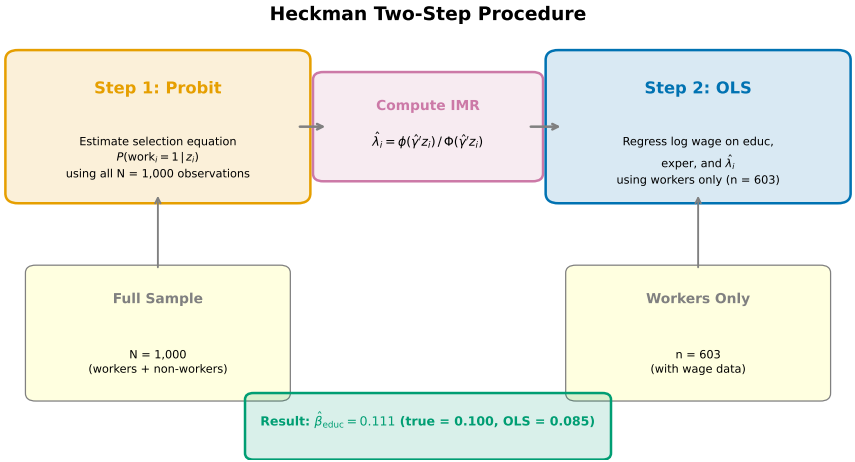
$$\log(\text{wage}_i) = \beta_0 + \beta_1 \,\text{educ}_i + \beta_2 \,\text{exper}_i + \delta\,\hat{\lambda}_i + \text{error}$$

Parameter	**OLS (workers)**	**Heckman**	**True**
$\hat{\beta}_0$ (intercept)	1.373	0.874	1.0
$\hat{\beta}_1$ (education)	0.085	**0.111**	0.100
$\hat{\beta}_2$ (experience)	0.040	0.039	0.040
$\hat{\delta}$ (IMR)	–	0.260	0.240

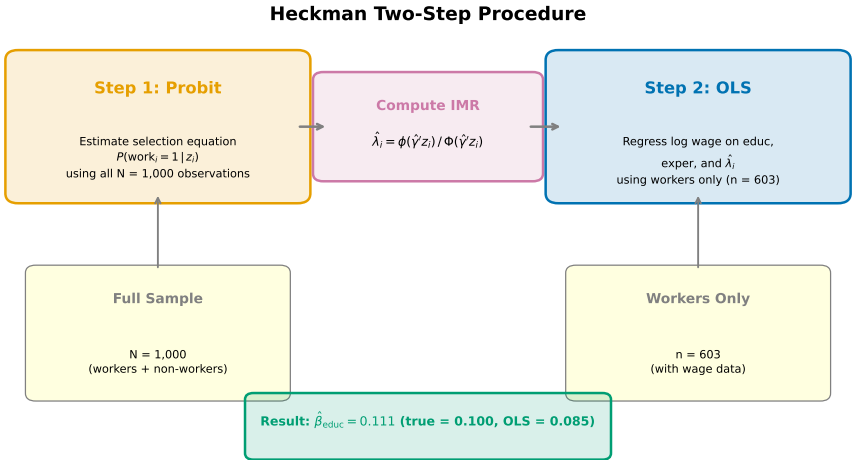
The Heckman education coefficient (**0.111**) is much closer to the true value (0.100) than OLS on workers (0.085).

$\Longrightarrow$ The selection correction removes the downward bias in the education coefficient.

The Procedure in One Picture

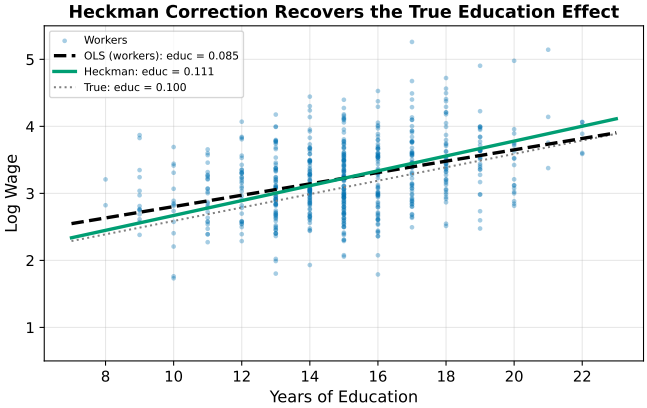


The Procedure in One Picture

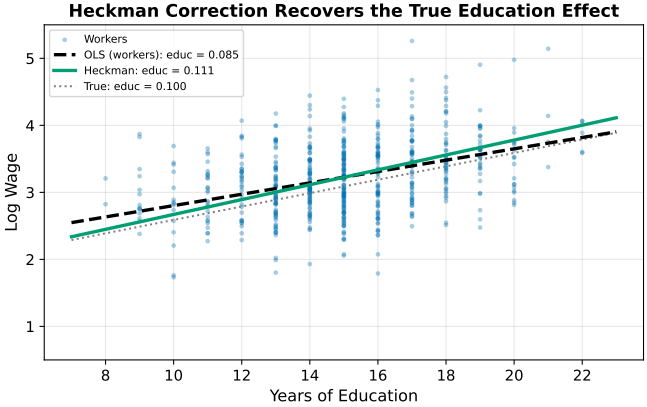


Step 1 uses **everyone** to learn who works. Step 2 uses **workers only** to learn how wages relate to education, after accounting for selection.

The Correction on the Scatter Plot



The Correction on the Scatter Plot



The dashed OLS line is too flat. The solid Heckman line is steeper, matching the true slope (gray dotted). Adding the IMR absorbs the selection effect that was biasing the education coefficient downward.

Interpreting the IMR Coefficient ($\hat{\delta}$)

The estimated IMR coefficient is $\hat{\delta} = 0.260$. What does it tell us?

Interpreting the IMR Coefficient ($\hat{\delta}$)

The estimated IMR coefficient is $\hat{\delta} = 0.260$. What does it tell us?

Recall: $\delta = \rho \cdot \sigma_u$, where $\rho = \text{Corr}(u, v)$ and σ_u is the wage error SD.

Interpreting the IMR Coefficient ($\hat{\delta}$)

The estimated IMR coefficient is $\hat{\delta} = 0.260$. What does it tell us?

Recall: $\delta = \rho \cdot \sigma_u$, where $\rho = \text{Corr}(u, v)$ and σ_u is the wage error SD.

- $\hat{\delta} > 0 \implies \hat{\rho} > 0$: unobserved factors that raise wages also raise the probability of working. Workers have higher-than-average wage errors

Interpreting the IMR Coefficient ($\hat{\delta}$)

The estimated IMR coefficient is $\hat{\delta} = 0.260$. What does it tell us?

Recall: $\delta = \rho \cdot \sigma_u$, where $\rho = \text{Corr}(u, v)$ and σ_u is the wage error SD.

- $\hat{\delta} > 0 \implies \hat{\rho} > 0$: unobserved factors that raise wages also raise the probability of working. Workers have higher-than-average wage errors
- $\hat{\delta} < 0$ would mean workers have *lower*-than-average unobserved wage determinants (e.g., people with high non-labor income stay home, regardless of their potential wage)

Interpreting the IMR Coefficient ($\hat{\delta}$)

The estimated IMR coefficient is $\hat{\delta} = 0.260$. What does it tell us?

Recall: $\delta = \rho \cdot \sigma_u$, where $\rho = \text{Corr}(u, v)$ and σ_u is the wage error SD.

- $\hat{\delta} > 0 \implies \hat{\rho} > 0$: unobserved factors that raise wages also raise the probability of working. Workers have higher-than-average wage errors
- $\hat{\delta} < 0$ would mean workers have *lower*-than-average unobserved wage determinants (e.g., people with high non-labor income stay home, regardless of their potential wage)
- $\hat{\delta} = 0$ would mean no selection bias: who works is unrelated to unobserved wage factors, and OLS on workers would be consistent

Interpreting the IMR Coefficient ($\hat{\delta}$)

The estimated IMR coefficient is $\hat{\delta} = 0.260$. What does it tell us?

Recall: $\delta = \rho \cdot \sigma_u$, where $\rho = \text{Corr}(u, v)$ and σ_u is the wage error SD.

- $\hat{\delta} > 0 \implies \hat{\rho} > 0$: unobserved factors that raise wages also raise the probability of working. Workers have higher-than-average wage errors
- $\hat{\delta} < 0$ would mean workers have *lower*-than-average unobserved wage determinants (e.g., people with high non-labor income stay home, regardless of their potential wage)
- $\hat{\delta} = 0$ would mean no selection bias: who works is unrelated to unobserved wage factors, and OLS on workers would be consistent

$\implies$ In our data, $\hat{\delta} = 0.260 > 0$: positive selection. The non-random sample of workers overrepresents high-ability individuals.

Where Are We? A Recap

1. **Problem:** OLS on workers underestimates the return to education (0.085 vs. 0.100) because high-ability people self-select into work

Where Are We? A Recap

1. **Problem:** OLS on workers underestimates the return to education (0.085 vs. 0.100) because high-ability people self-select into work
2. **Fix:** the Heckman two-step adds the inverse Mills ratio $\hat{\lambda}_i$ to the wage regression. This absorbs the selection effect and recovers a coefficient (0.111) much closer to the truth

Where Are We? A Recap

1. **Problem:** OLS on workers underestimates the return to education (0.085 vs. 0.100) because high-ability people self-select into work
2. **Fix:** the Heckman two-step adds the inverse Mills ratio $\hat{\lambda}_i$ to the wage regression. This absorbs the selection effect and recovers a coefficient (0.111) much closer to the truth
3. **The IMR coefficient** $\hat{\delta} = 0.260 > 0$ confirms positive selection: workers have above-average unobserved wage determinants

Where Are We? A Recap

1. **Problem:** OLS on workers underestimates the return to education (0.085 vs. 0.100) because high-ability people self-select into work
2. **Fix:** the Heckman two-step adds the inverse Mills ratio $\hat{\lambda}_i$ to the wage regression. This absorbs the selection effect and recovers a coefficient (0.111) much closer to the truth
3. **The IMR coefficient** $\hat{\delta} = 0.260 > 0$ confirms positive selection: workers have above-average unobserved wage determinants

What remains:

- What makes the Heckman model **credible**? (The exclusion restriction)
- How do we **test** whether selection bias is present?
- When does the model **fail**?

Outline

The Exclusion Restriction: Why Kids and Spouse Income

Notice that **kids** and **spouse income** appear in the selection equation but *not* in the wage equation.

The Exclusion Restriction: Why Kids and Spouse Income

Notice that **kids** and **spouse income** appear in the selection equation but *not* in the wage equation.

This is the **exclusion restriction**: we need at least one variable that affects *whether* someone works but does not directly affect their wage:

- **Number of children:** affects labor force participation (childcare responsibilities) but does not directly determine hourly productivity
- **Spouse income:** affects the financial need to work (reservation wage) but does not directly affect the wage an employer offers

The Exclusion Restriction: Why Kids and Spouse Income

Notice that **kids** and **spouse income** appear in the selection equation but *not* in the wage equation.

This is the **exclusion restriction**: we need at least one variable that affects *whether* someone works but does not directly affect their wage:

- **Number of children:** affects labor force participation (childcare responsibilities) but does not directly determine hourly productivity
- **Spouse income:** affects the financial need to work (reservation wage) but does not directly affect the wage an employer offers

Without an exclusion restriction, the IMR $\hat{\lambda}_i$ would be a nonlinear function of the same variables in the wage equation. The model is "technically identified" by functional form alone, but estimates become very unstable.

The Exclusion Restriction: Why Kids and Spouse Income

Notice that **kids** and **spouse income** appear in the selection equation but *not* in the wage equation.

This is the **exclusion restriction**: we need at least one variable that affects *whether* someone works but does not directly affect their wage:

- **Number of children:** affects labor force participation (childcare responsibilities) but does not directly determine hourly productivity
- **Spouse income:** affects the financial need to work (reservation wage) but does not directly affect the wage an employer offers

Without an exclusion restriction, the IMR $\hat{\lambda}_i$ would be a nonlinear function of the same variables in the wage equation. The model is "technically identified" by functional form alone, but estimates become very unstable.

$\implies$ A credible exclusion restriction is what separates a convincing Heckman model from one that is essentially relying on the normality assumption.

Full MLE: An Alternative to Two-Step

The two-step procedure is intuitive and easy to implement, but there is an alternative: estimate both equations simultaneously by **maximum likelihood**.

Full MLE: An Alternative to Two-Step

The two-step procedure is intuitive and easy to implement, but there is an alternative: estimate both equations simultaneously by **maximum likelihood**.

The full MLE maximizes the joint likelihood of the wage data (for workers) and the selection data (for everyone) together, estimating β, γ, σ_u, and ρ in one step.

Full MLE: An Alternative to Two-Step

The two-step procedure is intuitive and easy to implement, but there is an alternative: estimate both equations simultaneously by **maximum likelihood**.

The full MLE maximizes the joint likelihood of the wage data (for workers) and the selection data (for everyone) together, estimating β, γ, σ_u, and ρ in one step.

Advantages of full MLE:

- More efficient (smaller standard errors) than two-step
- Directly estimates ρ and σ_u separately

Full MLE: An Alternative to Two-Step

The two-step procedure is intuitive and easy to implement, but there is an alternative: estimate both equations simultaneously by **maximum likelihood**.

The full MLE maximizes the joint likelihood of the wage data (for workers) and the selection data (for everyone) together, estimating β, γ, σ_u, and ρ in one step.

Advantages of full MLE:

- More efficient (smaller standard errors) than two-step
- Directly estimates ρ and σ_u separately

Advantages of two-step:

- Less sensitive to distributional misspecification: the two-step does not impose the full joint likelihood structure, so it degrades more gracefully when normality is approximate
- Easier to diagnose: you can examine the probit and OLS stages separately
- The IMR coefficient test (next slide) provides a simple check for selection bias

Full MLE: An Alternative to Two-Step

The two-step procedure is intuitive and easy to implement, but there is an alternative: estimate both equations simultaneously by **maximum likelihood**.

The full MLE maximizes the joint likelihood of the wage data (for workers) and the selection data (for everyone) together, estimating β, γ, σ_u, and ρ in one step.

Advantages of full MLE:

- More efficient (smaller standard errors) than two-step
- Directly estimates ρ and σ_u separately

Advantages of two-step:

- Less sensitive to distributional misspecification: the two-step does not impose the full joint likelihood structure, so it degrades more gracefully when normality is approximate
- Easier to diagnose: you can examine the probit and OLS stages separately
- The IMR coefficient test (next slide) provides a simple check for selection bias

$\Longrightarrow$ In practice, many researchers run both and compare results. If they agree, the findings are more credible.

Testing for Selection Bias: Is the IMR Significant?

A simple test for selection bias: in the Heckman second stage, check whether $\hat{\delta}$ is statistically significant.

Testing for Selection Bias: Is the IMR Significant?

A simple test for selection bias: in the Heckman second stage, check whether $\hat{\delta}$ is statistically significant.

$$H_0\colon\ \delta = 0 \quad \Longleftrightarrow \quad \text{no selection bias (OLS on workers is consistent)}$$

$$H_1\colon\ \delta \neq 0 \quad \Longleftrightarrow \quad \text{selection bias present}$$

Testing for Selection Bias: Is the IMR Significant?

A simple test for selection bias: in the Heckman second stage, check whether $\hat{\delta}$ is statistically significant.

$$H_0\colon\ \delta = 0 \quad \Longleftrightarrow \quad \text{no selection bias (OLS on workers is consistent)}$$

$$H_1\colon\ \delta \neq 0 \quad \Longleftrightarrow \quad \text{selection bias present}$$

If you cannot reject H_0, selection bias may not be a problem, and OLS on workers is adequate.

Testing for Selection Bias: Is the IMR Significant?

A simple test for selection bias: in the Heckman second stage, check whether $\hat{\delta}$ is statistically significant.

$$H_0\colon\ \delta = 0 \quad \Longleftrightarrow \quad \text{no selection bias (OLS on workers is consistent)}$$

$$H_1\colon\ \delta \neq 0 \quad \Longleftrightarrow \quad \text{selection bias present}$$

If you cannot reject H_0, selection bias may not be a problem, and OLS on workers is adequate.

If you reject H_0, the Heckman correction is needed.

Testing for Selection Bias: Is the IMR Significant?

A simple test for selection bias: in the Heckman second stage, check whether $\hat{\delta}$ is statistically significant.

$$H_0\text{: } \delta = 0 \quad \Longleftrightarrow \quad \text{no selection bias (OLS on workers is consistent)}$$

$$H_1\text{: } \delta \neq 0 \quad \Longleftrightarrow \quad \text{selection bias present}$$

If you cannot reject H_0, selection bias may not be a problem, and OLS on workers is adequate.

If you reject H_0, the Heckman correction is needed.

Caveat: the standard errors from the naive Step 2 OLS are *incorrect* because $\hat{\lambda}_i$ is a generated regressor (estimated in Step 1). Software adjusts for this automatically; if computing by hand, you need a correction.

Testing for Selection Bias: Is the IMR Significant?

A simple test for selection bias: in the Heckman second stage, check whether $\hat{\delta}$ is statistically significant.

$$H_0\text{: } \delta = 0 \quad \Longleftrightarrow \quad \text{no selection bias (OLS on workers is consistent)}$$

$$H_1\text{: } \delta \neq 0 \quad \Longleftrightarrow \quad \text{selection bias present}$$

If you cannot reject H_0, selection bias may not be a problem, and OLS on workers is adequate.

If you reject H_0, the Heckman correction is needed.

Caveat: the standard errors from the naive Step 2 OLS are *incorrect* because $\hat{\lambda}_i$ is a generated regressor (estimated in Step 1). Software adjusts for this automatically; if computing by hand, you need a correction.

$\Longrightarrow$ In our data, $\hat{\delta} = 0.260$ is positive and statistically significant, confirming that selection bias is present and the correction is needed.

When the Heckman Model Fails

The Heckman model relies on several assumptions. It can fail when:

When the Heckman Model Fails

The Heckman model relies on several assumptions. It can fail when:

1. **No valid exclusion restriction:** if every variable that affects selection also affects wages, the IMR is identified only by functional form. Small deviations from normality produce large changes in estimates

When the Heckman Model Fails

The Heckman model relies on several assumptions. It can fail when:

1. **No valid exclusion restriction:** if every variable that affects selection also affects wages, the IMR is identified only by functional form. Small deviations from normality produce large changes in estimates
2. **Non-normal errors:** both the probit in Step 1 and the IMR formula assume joint normality of (u_i, v_i). Heavy tails or skewness invalidate the correction

When the Heckman Model Fails

The Heckman model relies on several assumptions. It can fail when:

1. **No valid exclusion restriction:** if every variable that affects selection also affects wages, the IMR is identified only by functional form. Small deviations from normality produce large changes in estimates
2. **Non-normal errors:** both the probit in Step 1 and the IMR formula assume joint normality of (u_i, v_i). Heavy tails or skewness invalidate the correction
3. **Misspecified selection equation:** if the probit model is wrong (missing variables, wrong functional form), the estimated IMR is wrong, and the correction introduces bias rather than removing it

When the Heckman Model Fails

The Heckman model relies on several assumptions. It can fail when:

1. **No valid exclusion restriction:** if every variable that affects selection also affects wages, the IMR is identified only by functional form. Small deviations from normality produce large changes in estimates
2. **Non-normal errors:** both the probit in Step 1 and the IMR formula assume joint normality of (u_i, v_i). Heavy tails or skewness invalidate the correction
3. **Misspecified selection equation:** if the probit model is wrong (missing variables, wrong functional form), the estimated IMR is wrong, and the correction introduces bias rather than removing it
4. **Weak selection:** if almost everyone works (or almost no one does), the IMR has very little variation across individuals, making it hard to identify δ

When the Heckman Model Fails

The Heckman model relies on several assumptions. It can fail when:

1. **No valid exclusion restriction:** if every variable that affects selection also affects wages, the IMR is identified only by functional form. Small deviations from normality produce large changes in estimates
2. **Non-normal errors:** both the probit in Step 1 and the IMR formula assume joint normality of (u_i, v_i). Heavy tails or skewness invalidate the correction
3. **Misspecified selection equation:** if the probit model is wrong (missing variables, wrong functional form), the estimated IMR is wrong, and the correction introduces bias rather than removing it
4. **Weak selection:** if almost everyone works (or almost no one does), the IMR has very little variation across individuals, making it hard to identify δ

$\Longrightarrow$ The Heckman model is powerful but not a magic fix. A credible exclusion restriction and reasonable normality are essential.

Decision Flowchart: Heckman vs. Tobit vs. OLS

1. Is your outcome missing for a non-random subset of observations?

Decision Flowchart: Heckman vs. Tobit vs. OLS

1. Is your outcome missing for a non-random subset of observations?
   - **Yes:** the missing values come from a *separate selection process* (e.g., wages unobserved because the person does not work)
     - Do you have a valid exclusion restriction? $\Longrightarrow$ **Heckman Selection Model**
     - No exclusion restriction? $\Longrightarrow$ Consider **bounds** or **sensitivity analysis**

Decision Flowchart: Heckman vs. Tobit vs. OLS

1. Is your outcome missing for a non-random subset of observations?
   - **Yes:** the missing values come from a *separate selection process* (e.g., wages unobserved because the person does not work)
     - Do you have a valid exclusion restriction? $\Longrightarrow$ **Heckman Selection Model**
     - No exclusion restriction? $\Longrightarrow$ Consider **bounds** or **sensitivity analysis**
   - **No:** the zeros are *corner solutions* (the person would choose a negative value but is constrained)
     - Same mechanism for participation and amount? $\Longrightarrow$ **Tobit**
     - Different mechanisms? $\Longrightarrow$ **Two-Part Model**

Decision Flowchart: Heckman vs. Tobit vs. OLS

1. Is your outcome missing for a non-random subset of observations?
   - **Yes:** the missing values come from a *separate selection process* (e.g., wages unobserved because the person does not work)
     - Do you have a valid exclusion restriction? $\Longrightarrow$ **Heckman Selection Model**
     - No exclusion restriction? $\Longrightarrow$ Consider **bounds** or **sensitivity analysis**
   - **No:** the zeros are *corner solutions* (the person would choose a negative value but is constrained)
     - Same mechanism for participation and amount? $\Longrightarrow$ **Tobit**
     - Different mechanisms? $\Longrightarrow$ **Two-Part Model**
2. Is there no selection or censoring at all?
   - $\Longrightarrow$ **OLS** is fine

Decision Flowchart: Heckman vs. Tobit vs. OLS

1. Is your outcome missing for a non-random subset of observations?
   - **Yes:** the missing values come from a *separate selection process* (e.g., wages unobserved because the person does not work)
     - Do you have a valid exclusion restriction? $\Longrightarrow$ **Heckman Selection Model**
     - No exclusion restriction? $\Longrightarrow$ Consider **bounds** or **sensitivity analysis**
   - **No:** the zeros are *corner solutions* (the person would choose a negative value but is constrained)
     - Same mechanism for participation and amount? $\Longrightarrow$ **Tobit**
     - Different mechanisms? $\Longrightarrow$ **Two-Part Model**
2. Is there no selection or censoring at all?
   - $\Longrightarrow$ **OLS** is fine

$\Longrightarrow$ Selection (Heckman) and censoring (Tobit) address different problems. The distinction is economic: does the outcome *exist but go unobserved*, or is it *constrained to a boundary*?

Outline

Summary: Back to Wages and Education

1. **The data problem:** 40% of individuals do not work, so their wages are **missing**. Workers have more education, fewer kids, and lower spouse income than non-workers

Summary: Back to Wages and Education

1. **The data problem:** 40% of individuals do not work, so their wages are **missing**. Workers have more education, fewer kids, and lower spouse income than non-workers
2. **OLS on workers is biased:** the education coefficient is 0.085 instead of the true 0.100. Unobserved factors that raise wages also raise the probability of working ($\rho = 0.6$)

Summary: Back to Wages and Education

1. **The data problem:** 40% of individuals do not work, so their wages are **missing**. Workers have more education, fewer kids, and lower spouse income than non-workers
2. **OLS on workers is biased:** the education coefficient is 0.085 instead of the true 0.100. Unobserved factors that raise wages also raise the probability of working ($\rho = 0.6$)
3. **The Heckman model** adds a selection equation (probit) and corrects the wage equation with the inverse Mills ratio. The corrected education coefficient is 0.111, close to the true 0.100

Summary: Back to Wages and Education

1. **The data problem:** 40% of individuals do not work, so their wages are **missing**. Workers have more education, fewer kids, and lower spouse income than non-workers
2. **OLS on workers is biased:** the education coefficient is 0.085 instead of the true 0.100. Unobserved factors that raise wages also raise the probability of working ($\rho = 0.6$)
3. **The Heckman model** adds a selection equation (probit) and corrects the wage equation with the inverse Mills ratio. The corrected education coefficient is 0.111, close to the true 0.100
4. **The exclusion restriction** (children, spouse income affect selection but not wages) is what makes the model credible

Summary: Back to Wages and Education

1. **The data problem:** 40% of individuals do not work, so their wages are **missing**. Workers have more education, fewer kids, and lower spouse income than non-workers
2. **OLS on workers is biased:** the education coefficient is 0.085 instead of the true 0.100. Unobserved factors that raise wages also raise the probability of working ($\rho = 0.6$)
3. **The Heckman model** adds a selection equation (probit) and corrects the wage equation with the inverse Mills ratio. The corrected education coefficient is 0.111, close to the true 0.100
4. **The exclusion restriction** (children, spouse income affect selection but not wages) is what makes the model credible
5. **Testing:** if the IMR coefficient is significant, selection bias is present and the correction is needed

Summary: Back to Wages and Education

1. **The data problem:** 40% of individuals do not work, so their wages are **missing**. Workers have more education, fewer kids, and lower spouse income than non-workers
2. **OLS on workers is biased:** the education coefficient is 0.085 instead of the true 0.100. Unobserved factors that raise wages also raise the probability of working ($\rho = 0.6$)
3. **The Heckman model** adds a selection equation (probit) and corrects the wage equation with the inverse Mills ratio. The corrected education coefficient is 0.111, close to the true 0.100
4. **The exclusion restriction** (children, spouse income affect selection but not wages) is what makes the model credible
5. **Testing:** if the IMR coefficient is significant, selection bias is present and the correction is needed
6. **Heckman vs. Tobit:** Tobit is for censoring (corner solutions). Heckman is for selection (missing data from a separate decision process)

James Heckman received the Nobel Prize in Economics in 2000, in part for developing this model.

Comparison: OLS vs. Heckman on Our Data

	OLS (workers)	**Heckman**	**True**
Education	0.085	0.111	0.100
Experience	0.040	0.039	0.040
IMR ($\hat{\delta}$)	–	0.260	0.240
Bias in educ	-15%	$+11\%$	–

Comparison: OLS vs. Heckman on Our Data

	OLS (workers)	**Heckman**	**True**
Education	0.085	0.111	0.100
Experience	0.040	0.039	0.040
IMR ($\hat{\delta}$)	–	0.260	0.240
Bias in educ	-15%	$+11\%$	–

The Heckman two-step does not recover the true parameter perfectly (0.111 vs. 0.100), but the bias is smaller and in the opposite direction. With larger samples, both converge to the true values.

Comparison: OLS vs. Heckman on Our Data

	OLS (workers)	**Heckman**	**True**
Education	0.085	0.111	0.100
Experience	0.040	0.039	0.040
IMR ($\hat{\delta}$)	–	0.260	0.240
Bias in educ	-15%	$+11\%$	–

The Heckman two-step does not recover the true parameter perfectly (0.111 vs. 0.100), but the bias is smaller and in the opposite direction. With larger samples, both converge to the true values.

$\Longrightarrow$ The selection correction works because it accounts for the fact that workers are not a random sample. Ignoring selection systematically underestimates the return to education in this setting.

Thank you!

jakeanderson@g.ucla.edu

Qualitative and Limited Dependent Variables

An Overview of Models for Non-Continuous Outcomes

Jake Anderson

May 16, 2026

Outline

The Problem: Non-Continuous Outcomes

Everything so far assumes y is continuous and unbounded. But many economic outcomes are not:

- **Binary:** work or not, default or not, buy or not
- **Unordered categories:** car / bus / train / bike
- **Ordered categories:** strongly disagree $\rightarrow$ strongly agree
- **Counts:** doctor visits, patents filed, arrests
- **Censored:** hours worked (piled up at zero)

The Problem: Non-Continuous Outcomes

Everything so far assumes y is continuous and unbounded. But many economic outcomes are not:

- **Binary:** work or not, default or not, buy or not
- **Unordered categories:** car / bus / train / bike
- **Ordered categories:** strongly disagree $\rightarrow$ strongly agree
- **Counts:** doctor visits, patents filed, arrests
- **Censored:** hours worked (piled up at zero)

$\Longrightarrow$ OLS is the wrong tool for all of these. This deck introduces the right ones.

OLS on a Binary Outcome: The Linear Probability Model

Suppose $y \in \{0, 1\}$ (e.g., drives to work or not). If we run OLS:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

OLS on a Binary Outcome: The Linear Probability Model

Suppose $y \in \{0, 1\}$ (e.g., drives to work or not). If we run OLS:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

This is the **Linear Probability Model (LPM)**. The fitted value $\hat{y}$ is interpreted as a probability. But there are problems:

OLS on a Binary Outcome: The Linear Probability Model

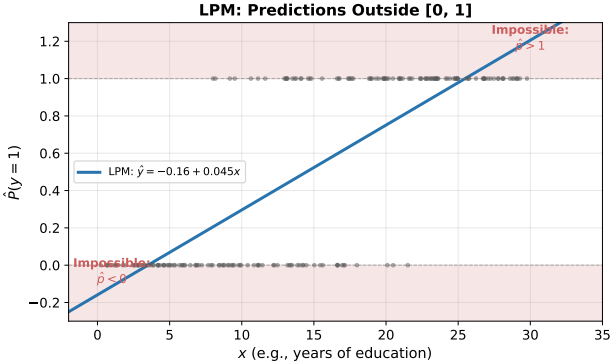
Suppose $y \in \{0, 1\}$ (e.g., drives to work or not). If we run OLS:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

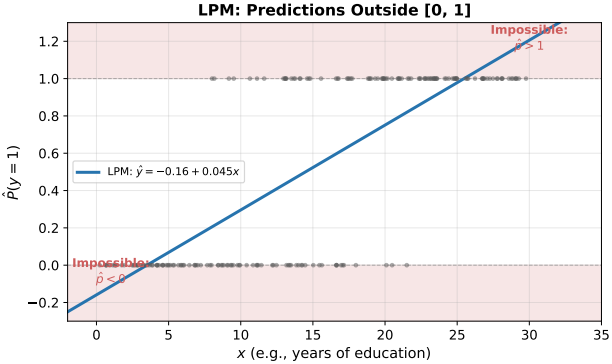
This is the **Linear Probability Model (LPM)**. The fitted value $\hat{y}$ is interpreted as a probability. But there are problems:

1. **Predictions outside** $[0, 1]$**:** OLS can predict $\hat{p} = -0.3$ or $\hat{p} = 1.4$
2. **Heteroskedasticity:** $\text{Var}(y \mid x) = p(1 - p)$ depends on x
3. **Constant marginal effects:** a one-unit change in x always changes probability by β_1, but probabilities are bounded

LPM: Predictions Outside $[0, 1]$



LPM: Predictions Outside $[0, 1]$



$\Longrightarrow$ The LPM's linear structure cannot respect the $[0, 1]$ bounds. We need a function that maps $x'\beta$ into $[0, 1]$.

The Latent Variable Framework

Many binary outcomes reflect an underlying continuous quantity we cannot observe. Call it y^*:

$$y_i^* = x_i'\beta + e_i$$

The Latent Variable Framework

Many binary outcomes reflect an underlying continuous quantity we cannot observe. Call it y^*:

$$y_i^* = x_i'\beta + e_i$$

We observe:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

The Latent Variable Framework

Many binary outcomes reflect an underlying continuous quantity we cannot observe. Call it y^*:

$$y_i^* = x_i'\beta + e_i$$

We observe:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

The probability of $y = 1$ depends on the distribution of e_i:

$$P(y_i = 1) = P(e_i > -x_i'\beta) = 1 - F(-x_i'\beta)$$

The Latent Variable Framework

Many binary outcomes reflect an underlying continuous quantity we cannot observe. Call it y^*:

$$y_i^* = x_i'\beta + e_i$$

We observe:

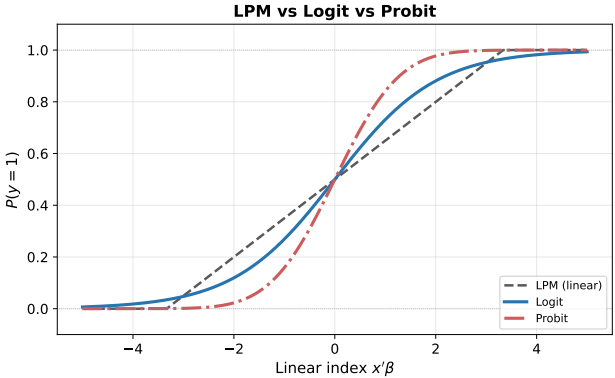
$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

The probability of $y = 1$ depends on the distribution of e_i:

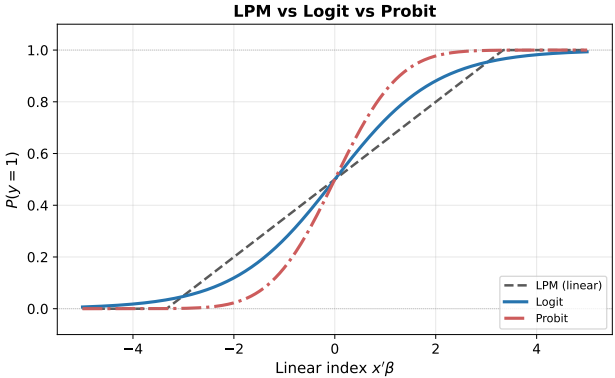
$$P(y_i = 1) = P(e_i > -x_i'\beta) = 1 - F(-x_i'\beta)$$

- If $e_i \sim N(0,1)$: $P(y = 1) = \Phi(x'\beta) \implies$ **Probit**
- If $e_i \sim$ Logistic: $P(y = 1) = \Lambda(x'\beta) = \dfrac{e^{x'\beta}}{1 + e^{x'\beta}} \implies$ **Logit**

The S-Curve: Logit and Probit vs LPM



The S-Curve: Logit and Probit vs LPM



Both logit and probit guarantee $\hat{p} \in [0, 1]$. The two S-curves are nearly identical in practice. Logit has slightly heavier tails.

Estimation: Maximum Likelihood

We cannot use OLS for probit/logit. Instead, we use **Maximum Likelihood Estimation (MLE)**: find the β that makes the observed data least surprising.

Estimation: Maximum Likelihood

We cannot use OLS for probit/logit. Instead, we use **Maximum Likelihood Estimation (MLE)**: find the β that makes the observed data least surprising.

For each observation:

$$f(y_i) = [\Phi(x_i'\beta)]^{y_i} \, [1 - \Phi(x_i'\beta)]^{1-y_i}$$

Estimation: Maximum Likelihood

We cannot use OLS for probit/logit. Instead, we use **Maximum Likelihood Estimation (MLE)**: find the β that makes the observed data least surprising.

For each observation:

$$f(y_i) = [\Phi(x_i'\beta)]^{y_i} \, [1 - \Phi(x_i'\beta)]^{1-y_i}$$

Log-likelihood for the whole sample:

$$\ln L = \sum_{i=1}^{N} \Big[y_i \ln \Phi(x_i'\beta) + (1 - y_i) \ln(1 - \Phi(x_i'\beta)) \Big]$$

Estimation: Maximum Likelihood

We cannot use OLS for probit/logit. Instead, we use **Maximum Likelihood Estimation (MLE)**: find the β that makes the observed data least surprising.

For each observation:

$$f(y_i) = [\Phi(x_i'\beta)]^{y_i} \, [1 - \Phi(x_i'\beta)]^{1-y_i}$$

Log-likelihood for the whole sample:

$$\ln L = \sum_{i=1}^{N} \Big[y_i \ln \Phi(x_i'\beta) + (1 - y_i) \ln(1 - \Phi(x_i'\beta)) \Big]$$

$\Longrightarrow$ MLE picks the β that maximizes this. In large samples, MLE is consistent, asymptotically normal, and efficient.

Marginal Effects: Why Coefficients Are Not Enough

In probit/logit, the coefficient β_k is **not** the marginal effect.

Marginal Effects: Why Coefficients Are Not Enough

In probit/logit, the coefficient β_k is **not** the marginal effect.

Probit:

$$\frac{\partial P}{\partial x_k} = \phi(x'\beta) \cdot \beta_k$$

Marginal Effects: Why Coefficients Are Not Enough

In probit/logit, the coefficient β_k is **not** the marginal effect.

Probit:

$$\frac{\partial P}{\partial x_k} = \phi(x'\beta) \cdot \beta_k$$

Logit:

$$\frac{\partial P}{\partial x_k} = \Lambda(x'\beta)(1 - \Lambda(x'\beta)) \cdot \beta_k$$

Marginal Effects: Why Coefficients Are Not Enough

In probit/logit, the coefficient β_k is **not** the marginal effect.

Probit:

$$\frac{\partial P}{\partial x_k} = \phi(x'\beta) \cdot \beta_k$$

Logit:

$$\frac{\partial P}{\partial x_k} = \Lambda(x'\beta)(1 - \Lambda(x'\beta)) \cdot \beta_k$$

$\Longrightarrow$ The marginal effect depends on where you are on the S-curve:

- Near $p = 0.5$ (middle): large effect
- Near $p = 0$ or $p = 1$ (tails): small effect

Marginal Effects: Why Coefficients Are Not Enough

In probit/logit, the coefficient β_k is **not** the marginal effect.

Probit:

$$\frac{\partial P}{\partial x_k} = \phi(x'\beta) \cdot \beta_k$$

Logit:

$$\frac{\partial P}{\partial x_k} = \Lambda(x'\beta)(1 - \Lambda(x'\beta)) \cdot \beta_k$$

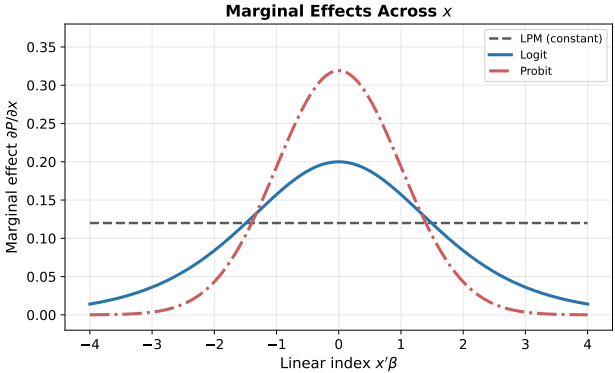
$\Longrightarrow$ The marginal effect depends on where you are on the S-curve:

- Near $p = 0.5$ (middle): large effect
- Near $p = 0$ or $p = 1$ (tails): small effect

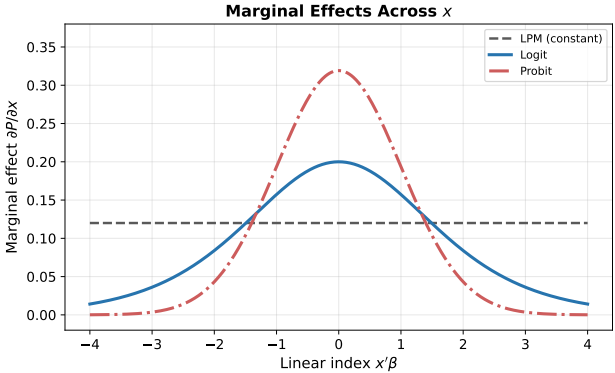
Common practice: report the **Average Marginal Effect (AME)**:

$$\widehat{AME} = \frac{1}{N} \sum_{i=1}^{N} \phi(\hat{\beta}_0 + \hat{\beta}_1 x_i) \cdot \hat{\beta}_1$$

Marginal Effects: LPM vs Logit/Probit



Marginal Effects: LPM vs Logit/Probit



The LPM assumes the same marginal effect everywhere. Logit/probit capture the fact that a change in x has the biggest impact on probability near $p = 0.5$.

Comparing LPM, Probit, and Logit

Coefficient scaling (approximate):

$$\hat{\beta}_{\text{Logit}} \approx 4\,\hat{\beta}_{\text{LPM}}, \qquad \hat{\beta}_{\text{Probit}} \approx 2.5\,\hat{\beta}_{\text{LPM}}, \qquad \hat{\beta}_{\text{Logit}} \approx 1.6\,\hat{\beta}_{\text{Probit}}$$

Comparing LPM, Probit, and Logit

Coefficient scaling (approximate):

$$\hat{\beta}_{\text{Logit}} \approx 4\,\hat{\beta}_{\text{LPM}}, \qquad \hat{\beta}_{\text{Probit}} \approx 2.5\,\hat{\beta}_{\text{LPM}}, \qquad \hat{\beta}_{\text{Logit}} \approx 1.6\,\hat{\beta}_{\text{Probit}}$$

	LPM	**Probit**	**Logit**
Estimation	OLS	MLE	MLE
$\hat{p} \in [0,1]$?	No	Yes	Yes
Marginal effects	Constant	Vary with x	Vary with x
Interpretation	Direct	Via ϕ	Via odds ratio

Comparing LPM, Probit, and Logit

Coefficient scaling (approximate):

$$\hat{\beta}_{\text{Logit}} \approx 4\,\hat{\beta}_{\text{LPM}}, \qquad \hat{\beta}_{\text{Probit}} \approx 2.5\,\hat{\beta}_{\text{LPM}}, \qquad \hat{\beta}_{\text{Logit}} \approx 1.6\,\hat{\beta}_{\text{Probit}}$$

	LPM	**Probit**	**Logit**
Estimation	OLS	MLE	MLE
$\hat{p} \in [0,1]$?	No	Yes	Yes
Marginal effects	Constant	Vary with x	Vary with x
Interpretation	Direct	Via ϕ	Via odds ratio

$\Longrightarrow$ In practice, all three give similar predicted probabilities and AMEs. Use probit/logit when you need predictions in $[0,1]$; use LPM as a quick baseline.

More Than Two Choices

What if the dependent variable has three or more **unordered** categories?

- Transportation mode: car, bus, train, bike
- College choice: no college, 2-year, 4-year
- Insurance: none, public, public + add-on

More Than Two Choices

What if the dependent variable has three or more **unordered** categories?

- Transportation mode: car, bus, train, bike
- College choice: no college, 2-year, 4-year
- Insurance: none, public, public + add-on

The **multinomial logit** extends binary logit to J categories. With one category as the base (say $j = 1$):

$$P(y_i = j \mid x_i) = \frac{e^{x_i' \beta_j}}{\sum_{k=1}^{J} e^{x_i' \beta_k}}$$

More Than Two Choices

What if the dependent variable has three or more **unordered** categories?

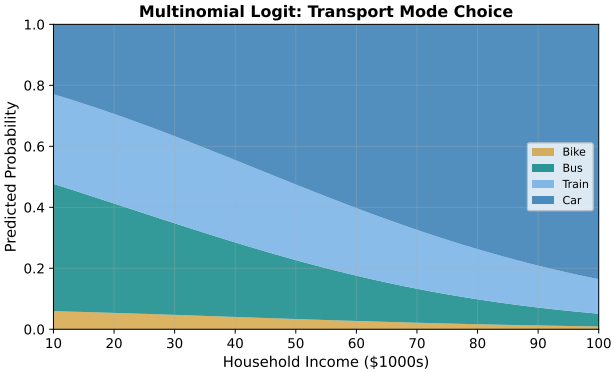
- Transportation mode: car, bus, train, bike
- College choice: no college, 2-year, 4-year
- Insurance: none, public, public + add-on

The **multinomial logit** extends binary logit to J categories. With one category as the base (say $j = 1$):

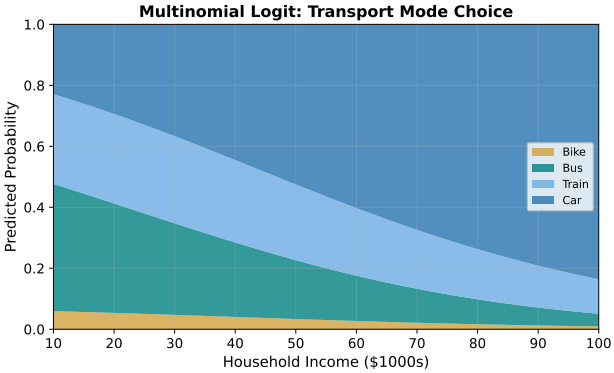
$$P(y_i = j \mid x_i) = \frac{e^{x_i' \beta_j}}{\sum_{k=1}^{J} e^{x_i' \beta_k}}$$

- Estimate $J - 1$ sets of coefficients (one per non-base category)
- Coefficients show the effect on the log-odds relative to the base
- Marginal effects are not the raw coefficients

Multinomial Logit: Predicted Probabilities



Multinomial Logit: Predicted Probabilities



As income rises, predicted choice shares shift from bus/bike toward car. The probabilities always sum to 1 across alternatives.

Independence of Irrelevant Alternatives (IIA)

Multinomial logit assumes the ratio of probabilities for any two choices does not depend on what other alternatives are available.

Independence of Irrelevant Alternatives (IIA)

Multinomial logit assumes the ratio of probabilities for any two choices does not depend on what other alternatives are available.

The red bus / blue bus problem:

- Initially: $P(\text{car}) = 0.5$, $P(\text{red bus}) = 0.5$
- Add an identical blue bus

Independence of Irrelevant Alternatives (IIA)

Multinomial logit assumes the ratio of probabilities for any two choices does not depend on what other alternatives are available.

The red bus / blue bus problem:

- Initially: $P(\text{car}) = 0.5$, $P(\text{red bus}) = 0.5$
- Add an identical blue bus

IIA predicts: $P(\text{car}) = P(\text{red bus}) = P(\text{blue bus}) = 0.33$

Independence of Irrelevant Alternatives (IIA)

Multinomial logit assumes the ratio of probabilities for any two choices does not depend on what other alternatives are available.

The red bus / blue bus problem:

- Initially: $P(\text{car}) = 0.5$, $P(\text{red bus}) = 0.5$
- Add an identical blue bus

IIA predicts: $P(\text{car}) = P(\text{red bus}) = P(\text{blue bus}) = 0.33$
But realistically: $P(\text{car}) = 0.5$, $P(\text{red bus}) = P(\text{blue bus}) = 0.25$

Independence of Irrelevant Alternatives (IIA)

Multinomial logit assumes the ratio of probabilities for any two choices does not depend on what other alternatives are available.

The red bus / blue bus problem:

- Initially: $P(\text{car}) = 0.5$, $P(\text{red bus}) = 0.5$
- Add an identical blue bus

IIA predicts: $P(\text{car}) = P(\text{red bus}) = P(\text{blue bus}) = 0.33$
But realistically: $P(\text{car}) = 0.5$, $P(\text{red bus}) = P(\text{blue bus}) = 0.25$

$\implies$ Adding a clone of an existing option should not steal share from a completely different option.
Test IIA with the Hausman-McFadden test; if it fails, consider nested logit or mixed logit.

When Categories Have a Natural Ranking

Sometimes the categories are ordered but the distances between them are unknown:

- Survey responses: strongly disagree $\rightarrow$ strongly agree
- Health satisfaction: low, medium, high
- Bond ratings: AAA, AA, A, BBB, ...

When Categories Have a Natural Ranking

Sometimes the categories are ordered but the distances between them are unknown:

- Survey responses: strongly disagree $\rightarrow$ strongly agree
- Health satisfaction: low, medium, high
- Bond ratings: AAA, AA, A, BBB, ...

The **ordered probit/logit** model assumes an underlying latent variable y^*:

$$y_i^* = x_i'\beta + e_i$$

When Categories Have a Natural Ranking

Sometimes the categories are ordered but the distances between them are unknown:

- Survey responses: strongly disagree $\rightarrow$ strongly agree
- Health satisfaction: low, medium, high
- Bond ratings: AAA, AA, A, BBB, ...

The **ordered probit/logit** model assumes an underlying latent variable y^*:

$$y_i^* = x_i'\beta + e_i$$

The observed outcome depends on where y^* falls relative to threshold parameters (**cutpoints**) $\mu_1, \mu_2, \ldots$:

$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < y_i^* \leq \mu_2 \\ 3 & \text{if } y_i^* > \mu_2 \end{cases}$$

When Categories Have a Natural Ranking

Sometimes the categories are ordered but the distances between them are unknown:

- Survey responses: strongly disagree $\rightarrow$ strongly agree
- Health satisfaction: low, medium, high
- Bond ratings: AAA, AA, A, BBB, ...

The **ordered probit/logit** model assumes an underlying latent variable y^*:

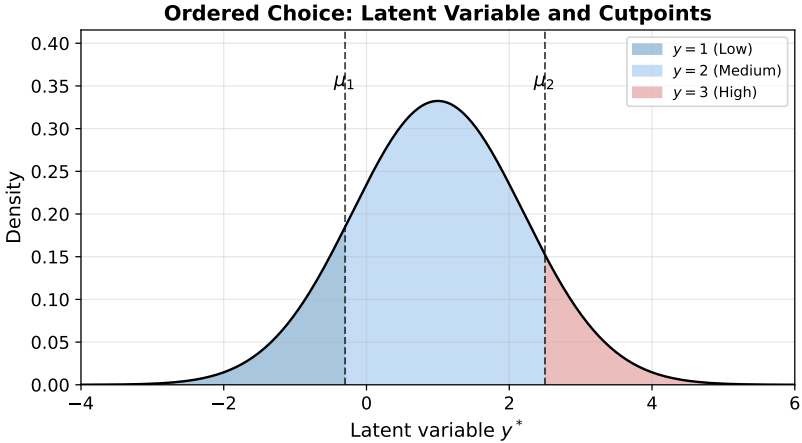
$$y_i^* = x_i'\beta + e_i$$

The observed outcome depends on where y^* falls relative to threshold parameters (**cutpoints**) $\mu_1, \mu_2, \ldots$:

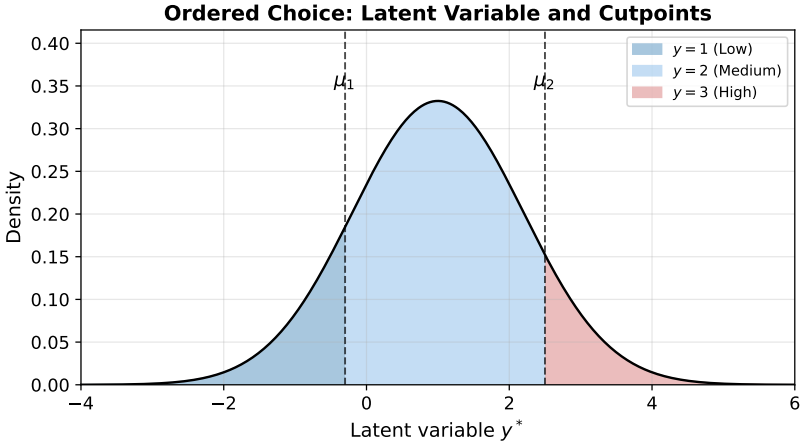
$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < y_i^* \leq \mu_2 \\ 3 & \text{if } y_i^* > \mu_2 \end{cases}$$

The cutpoints μ are estimated along with β.

Ordered Choice: The Latent Variable



Ordered Choice: The Latent Variable



A change in x shifts the entire distribution of y^*, simultaneously changing the probability of every category. The marginal effects must sum to zero across categories.

Ordered Choice: Marginal Effects

For a continuous variable in a three-category model:

Ordered Choice: Marginal Effects

For a continuous variable in a three-category model:

$$\frac{\partial P(y=1)}{\partial x_k} = -\phi(\mu_1 - x'\beta) \cdot \beta_k$$

$$\frac{\partial P(y=2)}{\partial x_k} = [\phi(\mu_1 - x'\beta) - \phi(\mu_2 - x'\beta)] \cdot \beta_k$$

$$\frac{\partial P(y=3)}{\partial x_k} = \phi(\mu_2 - x'\beta) \cdot \beta_k$$

Ordered Choice: Marginal Effects

For a continuous variable in a three-category model:

$$\frac{\partial P(y=1)}{\partial x_k} = -\phi(\mu_1 - x'\beta) \cdot \beta_k$$

$$\frac{\partial P(y=2)}{\partial x_k} = [\phi(\mu_1 - x'\beta) - \phi(\mu_2 - x'\beta)] \cdot \beta_k$$

$$\frac{\partial P(y=3)}{\partial x_k} = \phi(\mu_2 - x'\beta) \cdot \beta_k$$

$\implies$ The sign of β_k tells you the direction for the highest and lowest categories, but the middle categories could go either way.

Ordered Choice: Marginal Effects

For a continuous variable in a three-category model:

$$\frac{\partial P(y=1)}{\partial x_k} = -\phi(\mu_1 - x'\beta) \cdot \beta_k$$

$$\frac{\partial P(y=2)}{\partial x_k} = [\phi(\mu_1 - x'\beta) - \phi(\mu_2 - x'\beta)] \cdot \beta_k$$

$$\frac{\partial P(y=3)}{\partial x_k} = \phi(\mu_2 - x'\beta) \cdot \beta_k$$

$\implies$ The sign of β_k tells you the direction for the highest and lowest categories, but the middle categories could go either way.

For binary variables: compute the **discrete difference** (change in each category's probability when the dummy goes from 0 to 1).

Counts: Non-Negative Integers

Some outcomes are counts: doctor visits, patents, arrests. Counts are non-negative integers, often right-skewed with many zeros.

Counts: Non-Negative Integers

Some outcomes are counts: doctor visits, patents, arrests. Counts are non-negative integers, often right-skewed with many zeros.

The **Poisson model** assumes:

$$P(Y = y) = \frac{e^{-\mu}\,\mu^{y}}{y!}, \qquad y = 0, 1, 2, \ldots$$

where $\mu = E(Y) = \text{Var}(Y)$.

Counts: Non-Negative Integers

Some outcomes are counts: doctor visits, patents, arrests. Counts are non-negative integers, often right-skewed with many zeros.

The **Poisson model** assumes:

$$P(Y = y) = \frac{e^{-\mu}\,\mu^{y}}{y!}, \qquad y = 0, 1, 2, \ldots$$

where $\mu = E(Y) = \text{Var}(Y)$.

We model the conditional mean as:

$$\mu_i = \exp(x_i'\beta)$$

Counts: Non-Negative Integers

Some outcomes are counts: doctor visits, patents, arrests. Counts are non-negative integers, often right-skewed with many zeros.

The **Poisson model** assumes:

$$P(Y = y) = \frac{e^{-\mu}\,\mu^{y}}{y!}, \qquad y = 0, 1, 2, \ldots$$

where $\mu = E(Y) = \text{Var}(Y)$.

We model the conditional mean as:

$$\mu_i = \exp(x_i'\beta)$$

$\Longrightarrow$ The exponential ensures $\mu > 0$. A one-unit increase in x_k multiplies the expected count by e^{β_k}.

The Overdispersion Problem

Poisson assumes $E(Y) = \text{Var}(Y)$ (**equidispersion**). In real data, the variance almost always exceeds the mean (**overdispersion**).

The Overdispersion Problem

Poisson assumes $E(Y) = \text{Var}(Y)$ (**equidispersion**). In real data, the variance almost always exceeds the mean (**overdispersion**).

If overdispersion is present:

- Poisson coefficient estimates are still **consistent**
- But standard errors are **too small**
- Hypothesis tests become unreliable

The Overdispersion Problem

Poisson assumes $E(Y) = \text{Var}(Y)$ (**equidispersion**). In real data, the variance almost always exceeds the mean (**overdispersion**).

If overdispersion is present:

- Poisson coefficient estimates are still **consistent**
- But standard errors are **too small**
- Hypothesis tests become unreliable

The **negative binomial model** relaxes equidispersion:

$$\text{Var}(Y) = \mu + \alpha\mu^2$$

The Overdispersion Problem

Poisson assumes $E(Y) = \text{Var}(Y)$ (**equidispersion**). In real data, the variance almost always exceeds the mean (**overdispersion**).

If overdispersion is present:

- Poisson coefficient estimates are still **consistent**
- But standard errors are **too small**
- Hypothesis tests become unreliable

The **negative binomial model** relaxes equidispersion:

$$\text{Var}(Y) = \mu + \alpha\mu^2$$

When $\alpha = 0$: reduces to Poisson. When $\alpha > 0$: allows overdispersion.

The Overdispersion Problem

Poisson assumes $E(Y) = \text{Var}(Y)$ (**equidispersion**). In real data, the variance almost always exceeds the mean (**overdispersion**).

If overdispersion is present:

- Poisson coefficient estimates are still **consistent**
- But standard errors are **too small**
- Hypothesis tests become unreliable

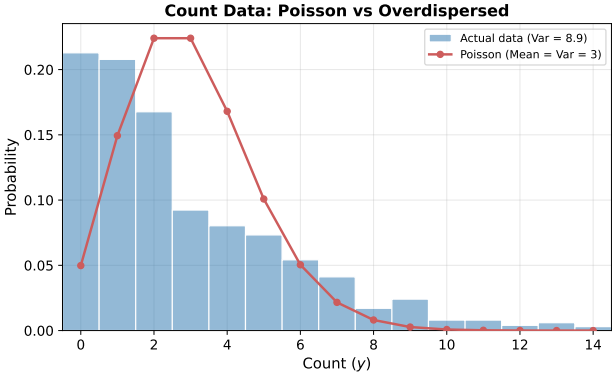
The **negative binomial model** relaxes equidispersion:

$$\text{Var}(Y) = \mu + \alpha\mu^2$$

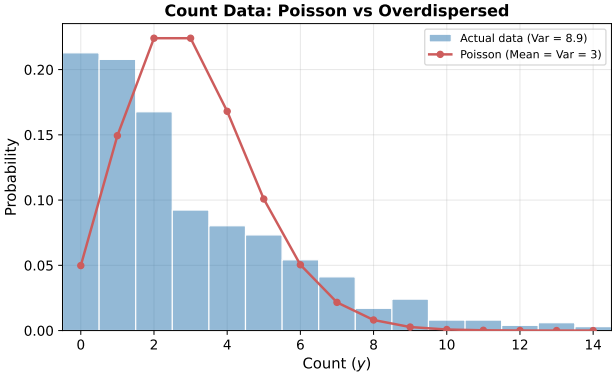
When $\alpha = 0$: reduces to Poisson. When $\alpha > 0$: allows overdispersion.

$\implies$ Test for overdispersion by testing $H_0 \colon \alpha = 0$.

Count Data: Poisson vs Overdispersed



Count Data: Poisson vs Overdispersed



The actual data has a longer right tail and more zeros than Poisson predicts. The negative binomial accommodates this extra variability.

Censored vs Truncated Data

Censored: everyone is in the sample, but some values are "clipped."

- Hours worked: we observe 0 for non-workers, but their "desired hours" might be negative

Censored vs Truncated Data

Censored: everyone is in the sample, but some values are "clipped."

- Hours worked: we observe 0 for non-workers, but their "desired hours" might be negative

Truncated: some observations are excluded entirely.

- If we only survey earners above the poverty line, we never see anyone below it

Censored vs Truncated Data

Censored: everyone is in the sample, but some values are "clipped."

- Hours worked: we observe 0 for non-workers, but their "desired hours" might be negative

Truncated: some observations are excluded entirely.

- If we only survey earners above the poverty line, we never see anyone below it

The observed censored variable:

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Censored vs Truncated Data

Censored: everyone is in the sample, but some values are "clipped."

- Hours worked: we observe 0 for non-workers, but their "desired hours" might be negative

Truncated: some observations are excluded entirely.

- If we only survey earners above the poverty line, we never see anyone below it

The observed censored variable:

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

$\Longrightarrow$ Censoring creates a pile-up at zero. OLS on the censored data attenuates the slope toward zero (similar to measurement error bias).

The Tobit Model

The **Tobit model** handles censored data. It combines a probit (for whether $y > 0$) with a linear regression (for the magnitude when positive):

$$y_i^* = x_i'\beta + e_i, \qquad e_i \sim N(0, \sigma^2)$$

The Tobit Model

The **Tobit model** handles censored data. It combines a probit (for whether $y > 0$) with a linear regression (for the magnitude when positive):

$$y_i^* = x_i'\beta + e_i, \qquad e_i \sim N(0, \sigma^2)$$

A change in x affects the outcome through two channels:

1. **Extensive margin:** changing $P(y > 0)$
2. **Intensive margin:** changing $E(y \mid y > 0)$

The Tobit Model

The **Tobit model** handles censored data. It combines a probit (for whether $y > 0$) with a linear regression (for the magnitude when positive):

$$y_i^* = x_i'\beta + e_i, \qquad e_i \sim N(0, \sigma^2)$$

A change in x affects the outcome through two channels:

1. **Extensive margin:** changing $P(y > 0)$
2. **Intensive margin:** changing $E(y \mid y > 0)$

Limitation: Tobit assumes the same β governs both margins. If the decision to participate depends on different factors than the amount, Tobit is misspecified.

The Tobit Model

The **Tobit model** handles censored data. It combines a probit (for whether $y > 0$) with a linear regression (for the magnitude when positive):

$$y_i^* = x_i'\beta + e_i, \qquad e_i \sim N(0, \sigma^2)$$

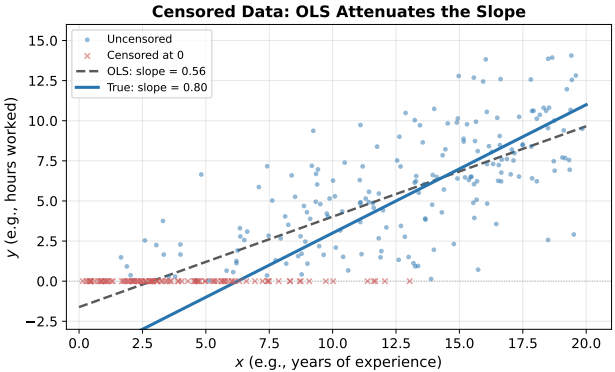
A change in x affects the outcome through two channels:

1. **Extensive margin:** changing $P(y > 0)$
2. **Intensive margin:** changing $E(y \mid y > 0)$

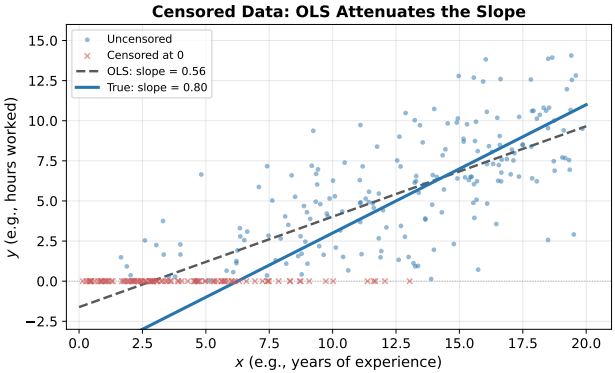
Limitation: Tobit assumes the same β governs both margins. If the decision to participate depends on different factors than the amount, Tobit is misspecified.

$\Longrightarrow$ The Heckman selection model relaxes this by allowing separate equations for the two stages.

Censored Data: OLS vs the True Relationship

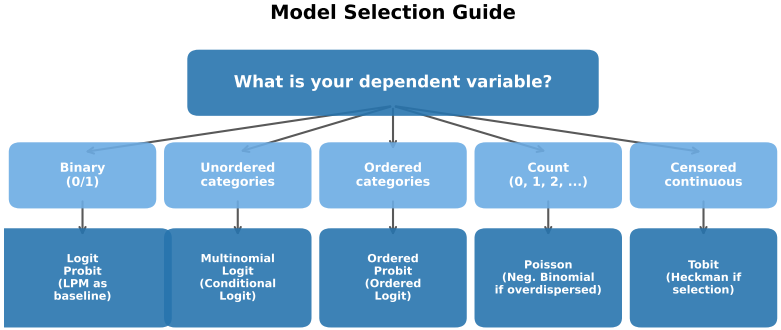


Censored Data: OLS vs the True Relationship



OLS pulls the slope toward zero because it treats the censored zeros as genuine low values. Tobit recovers the steeper true slope.

Choosing the Right Model



All estimated by Maximum Likelihood (except LPM, which uses OLS)

Interpret coefficients through marginal effects, not raw values

Model Selection: Summary Table

Dependent Variable	**Model**	**Estimation**
Binary (0/1)	LPM, Probit, Logit	OLS / MLE
Unordered categories	Multinomial Logit	MLE
Ordered categories	Ordered Probit/Logit	MLE
Count (0, 1, 2, ...)	Poisson, Neg. Binomial	MLE
Censored continuous	Tobit	MLE
Selected sample	Heckman Selection	Two-step / MLE

Model Selection: Summary Table

Dependent Variable	**Model**	**Estimation**
Binary (0/1)	LPM, Probit, Logit	OLS / MLE
Unordered categories	Multinomial Logit	MLE
Ordered categories	Ordered Probit/Logit	MLE
Count (0, 1, 2, ...)	Poisson, Neg. Binomial	MLE
Censored continuous	Tobit	MLE
Selected sample	Heckman Selection	Two-step / MLE

$\Longrightarrow$ The common thread: match the model to the structure of y. In all cases, interpret results through **marginal effects**, not raw coefficients.

Model Selection: Summary Table

Dependent Variable	**Model**	**Estimation**
Binary (0/1)	LPM, Probit, Logit	OLS / MLE
Unordered categories	Multinomial Logit	MLE
Ordered categories	Ordered Probit/Logit	MLE
Count (0, 1, 2, ...)	Poisson, Neg. Binomial	MLE
Censored continuous	Tobit	MLE
Selected sample	Heckman Selection	Two-step / MLE

$\Longrightarrow$ The common thread: match the model to the structure of y. In all cases, interpret results through **marginal effects**, not raw coefficients.

$\Longrightarrow$ For all MLE models: goodness of fit is measured by pseudo-R^2 and percent correctly predicted, not R^2.

Thank you!
jakeanderson@g.ucla.edu