The Simple Linear Regression Model

From Scatter Plot to Population Model

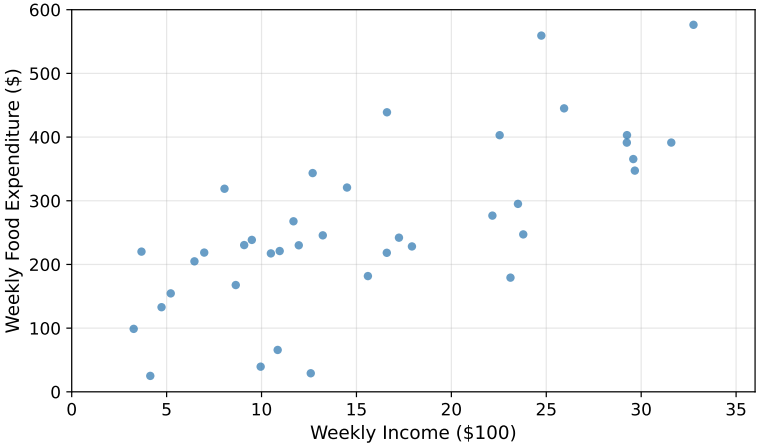
Jake Anderson

March 21, 2026

Outline

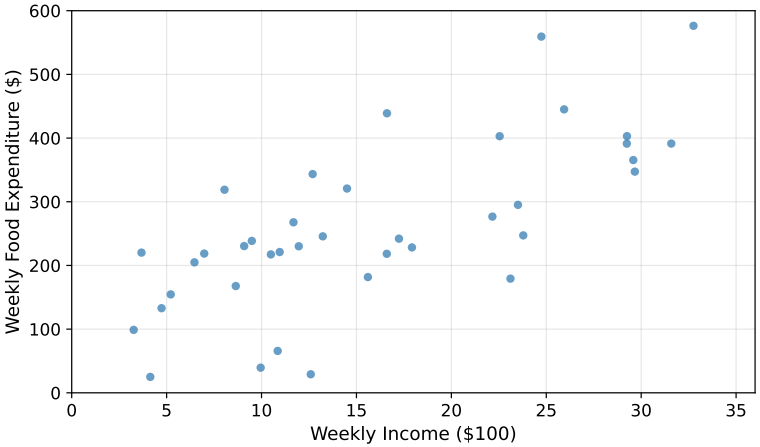
Back to the Food Expenditure Example

In Lecture 1 you saw a scatter plot of **weekly food expenditure** vs. **weekly household income** for 40 households from southern Australia.



Back to the Food Expenditure Example

In Lecture 1 you saw a scatter plot of **weekly food expenditure** vs. **weekly household income** for 40 households from southern Australia.



A positive relationship is visible. But can we be more precise? What is the *average* food expenditure

What We Want to Know

Two concrete questions:

1. If weekly income goes up by \$100, how much does *average* weekly food expenditure rise?
2. Can we predict food expenditure for a household with income \$2000/week?

What We Want to Know

Two concrete questions:

1. If weekly income goes up by \$100, how much does *average* weekly food expenditure rise?
2. Can we predict food expenditure for a household with income \$2000/week?

To answer these, we need a **model**: a mathematical description of how y (food expenditure) relates to x (income).

What We Want to Know

Two concrete questions:

1. If weekly income goes up by \$100, how much does *average* weekly food expenditure rise?
2. Can we predict food expenditure for a household with income \$2000/week?

To answer these, we need a **model**: a mathematical description of how y (food expenditure) relates to x (income).

We previewed the equation $y = \beta_1 + \beta_2 x + e$ in Lecture 1. Now let's build it from scratch and understand every piece.

Outline

A Deterministic Starting Point

Imagine every household follows a fixed spending rule: \$83 base plus \$10 for every \$100 of weekly income goes to food.

$$y = 83 + 10x, \quad \text{where } x \text{ is weekly income in hundreds of dollars}$$

A Deterministic Starting Point

Imagine every household follows a fixed spending rule: \$83 base plus \$10 for every \$100 of weekly income goes to food.

$$y = 83 + 10x, \quad \text{where } x \text{ is weekly income in hundreds of dollars}$$

This gives clean predictions and a clear marginal effect: $\Delta y / \Delta x = 10$.

A Deterministic Starting Point

Imagine every household follows a fixed spending rule: \$83 base plus \$10 for every \$100 of weekly income goes to food.

$$y = 83 + 10x, \quad \text{where } x \text{ is weekly income in hundreds of dollars}$$

This gives clean predictions and a clear marginal effect: $\Delta y / \Delta x = 10$.

Let's check this against the data.

The Deterministic Model Falls Apart

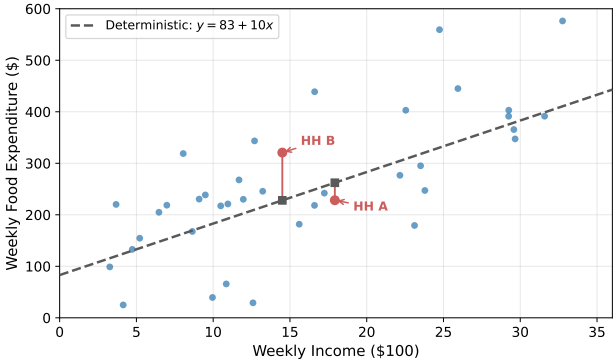
Using $y = 83 + 10x$, consider two households:

- **Household A:** income $x = 20$. Model predicts $y = 283$. Actual spending: \$350.
- **Household B:** income $x = 15$. Model predicts $y = 233$. Actual spending: \$120.

The Deterministic Model Falls Apart

Using $y = 83 + 10x$, consider two households:

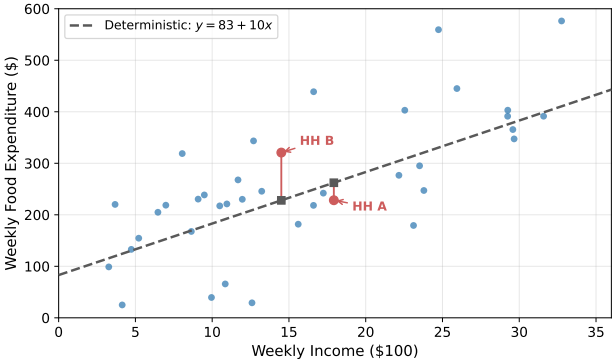
- **Household A:** income $x = 20$. Model predicts $y = 283$. Actual spending: \$350.
- **Household B:** income $x = 15$. Model predicts $y = 233$. Actual spending: \$120.



The Deterministic Model Falls Apart

Using $y = 83 + 10x$, consider two households:

- **Household A:** income $x = 20$. Model predicts $y = 283$. Actual spending: \$350.
- **Household B:** income $x = 15$. Model predicts $y = 233$. Actual spending: \$120.



$\Longrightarrow$ At the same income, different households spend very different amounts. What is the model

Adding the Error Term

What explains the variation in food expenditure *beyond* income?

- Household composition, dietary preferences, location
- Whether the household eats out or cooks at home
- Impulse shopping, sales, seasonal effects

Adding the Error Term

What explains the variation in food expenditure *beyond* income?

- Household composition, dietary preferences, location
- Whether the household eats out or cooks at home
- Impulse shopping, sales, seasonal effects

Collectively, these factors are represented by a single random variable e:

$$y = \beta_1 + \beta_2 x + e$$

Adding the Error Term

What explains the variation in food expenditure *beyond* income?

- Household composition, dietary preferences, location
- Whether the household eats out or cooks at home
- Impulse shopping, sales, seasonal effects

Collectively, these factors are represented by a single random variable e:

$$y = \beta_1 + \beta_2 x + e$$

- β_1: intercept (expected food expenditure when income is zero)
- β_2: slope (change in *expected* food expenditure per \$100 increase in income)
- e: random error (everything else affecting y)

Adding the Error Term

What explains the variation in food expenditure *beyond* income?

- Household composition, dietary preferences, location
- Whether the household eats out or cooks at home
- Impulse shopping, sales, seasonal effects

Collectively, these factors are represented by a single random variable e:

$$y = \beta_1 + \beta_2 x + e$$

- β_1: intercept (expected food expenditure when income is zero)
- β_2: slope (change in *expected* food expenditure per \$100 increase in income)
- e: random error (everything else affecting y)

β_1 and β_2 are unknown **population parameters**. We never observe them directly.

The Model for Each Observation

We have $N = 40$ households, randomly sampled. Each one satisfies the same behavioral rule:

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

The Model for Each Observation

We have $N = 40$ households, randomly sampled. Each one satisfies the same behavioral rule:

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

- y_i: food expenditure for household i (the **dependent variable**)
- x_i: income for household i (the **independent variable**)
- e_i: the error for household i (unobservable)

The Model for Each Observation

We have $N = 40$ households, randomly sampled. Each one satisfies the same behavioral rule:

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

- y_i: food expenditure for household i (the **dependent variable**)
- x_i: income for household i (the **independent variable**)
- e_i: the error for household i (unobservable)

Because households are randomly sampled, the pairs (y_i, x_i) are **independent and identically distributed (iid)**. For our derivations, we condition on the observed x values and study the behavior of the errors and estimators.

The Model for Each Observation

We have $N = 40$ households, randomly sampled. Each one satisfies the same behavioral rule:

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

- y_i: food expenditure for household i (the **dependent variable**)
- x_i: income for household i (the **independent variable**)
- e_i: the error for household i (unobservable)

Because households are randomly sampled, the pairs (y_i, x_i) are **independent and identically distributed (iid)**. For our derivations, we condition on the observed x values and study the behavior of the errors and estimators.

This equation is the **data generating process (DGP)**: it describes how the observable data arise from the unknown parameters plus randomness.

Decomposing y_i

The model splits each observation into two parts:

$$y_i = \underbrace{\beta_1 + \beta_2 x_i}_{\text{systematic component}} + \underbrace{e_i}_{\text{random component}}$$

Decomposing y_i

The model splits each observation into two parts:

$$y_i = \underbrace{\beta_1 + \beta_2 x_i}_{\text{systematic component}} + \underbrace{e_i}_{\text{random component}}$$

- The **systematic component** is the part of y_i that depends on x_i through the model.
- The **random component** e_i is everything else: the variation in y_i not captured by x_i.

Decomposing y_i

The model splits each observation into two parts:

$$y_i = \underbrace{\beta_1 + \beta_2 x_i}_{\text{systematic component}} + \underbrace{e_i}_{\text{random component}}$$

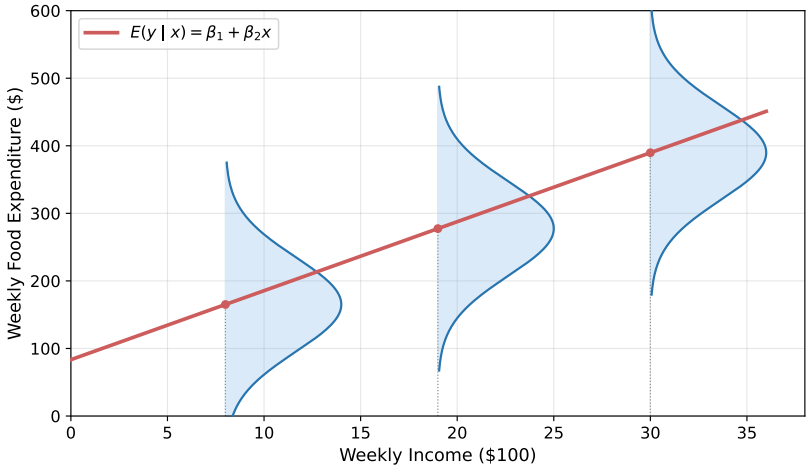
- The **systematic component** is the part of y_i that depends on x_i through the model.
- The **random component** e_i is everything else: the variation in y_i not captured by x_i.

The systematic component will turn out to be the *conditional mean* $\mathrm{E}(y_i \mid x_i)$, once we impose an assumption on e_i.

Outline

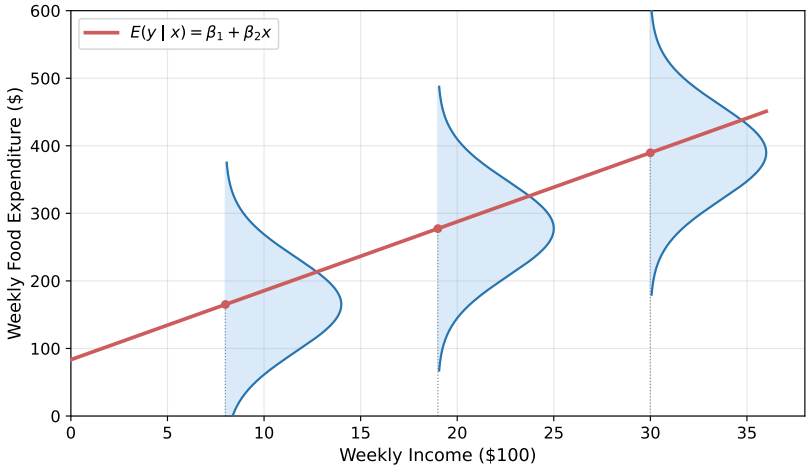
What Does “Average Food Expenditure at a Given Income” Mean?

At any income level x, there is a **distribution** of food expenditure values across households. Some spend more, some less.



What Does “Average Food Expenditure at a Given Income” Mean?

At any income level x, there is a **distribution** of food expenditure values across households. Some spend more, some less.



From $E(e \mid x) = 0$ to the Regression Function

Suppose we assume that the error term has zero conditional mean:

$$E(e_i \mid x_i) = 0$$

From $E(e \mid x) = 0$ to the Regression Function

Suppose we assume that the error term has zero conditional mean:

$$E(e_i \mid x_i) = 0$$

Then take the conditional expectation of both sides of $y_i = \beta_1 + \beta_2 x_i + e_i$. Since β_1 and β_2 are constants and x_i is fixed when we condition on it, they pass through the expectation:

$$\begin{aligned} E(y_i \mid x_i) &= E(\beta_1 + \beta_2 x_i + e_i \mid x_i) \\ &= \beta_1 + \beta_2 x_i + \underbrace{E(e_i \mid x_i)}_{=0} \end{aligned}$$

The Population Regression Function

$$E(y_i \mid x_i) = \beta_1 + \beta_2 x_i$$

The average value of y given x is a linear function of x.

From $E(e \mid x) = 0$ to the Regression Function

Suppose we assume that the error term has zero conditional mean:

$$E(e_i \mid x_i) = 0$$

Then take the conditional expectation of both sides of $y_i = \beta_1 + \beta_2 x_i + e_i$. Since β_1 and β_2 are constants and x_i is fixed when we condition on it, they pass through the expectation:

$$\begin{aligned} E(y_i \mid x_i) &= E(\beta_1 + \beta_2 x_i + e_i \mid x_i) \\ &= \beta_1 + \beta_2 x_i + \underbrace{E(e_i \mid x_i)}_{=0} \end{aligned}$$

From $E(e \mid x) = 0$ to the Regression Function

Suppose we assume that the error term has zero conditional mean:

$$E(e_i \mid x_i) = 0$$

Then take the conditional expectation of both sides of $y_i = \beta_1 + \beta_2 x_i + e_i$. Since β_1 and β_2 are constants and x_i is fixed when we condition on it, they pass through the expectation:

$$\begin{aligned} E(y_i \mid x_i) &= E(\beta_1 + \beta_2 x_i + e_i \mid x_i) \\ &= \beta_1 + \beta_2 x_i + \underbrace{E(e_i \mid x_i)}_{=0} \end{aligned}$$

The Population Regression Function

$$E(y_i \mid x_i) = \beta_1 + \beta_2 x_i$$

The average value of y given x is a linear function of x.

Interpreting β_1 and β_2

$$\mathsf{E}(y \mid x) = \beta_1 + \beta_2 x$$

Interpreting β_1 and β_2

$$\mathrm{E}(y \mid x) = \beta_1 + \beta_2 x$$

Slope (β_2):

$$\beta_2 = \frac{\Delta\, \mathrm{E}(y \mid x)}{\Delta x}$$

A one-unit increase in x (here, \$100 in weekly income) changes expected food expenditure by β_2 dollars, holding all else constant.

Interpreting β_1 and β_2

$$\mathrm{E}(y \mid x) = \beta_1 + \beta_2 x$$

Slope (β_2):

$$\beta_2 = \frac{\Delta\, \mathrm{E}(y \mid x)}{\Delta x}$$

A one-unit increase in x (here, \$100 in weekly income) changes expected food expenditure by β_2 dollars, holding all else constant.

Intercept (β_1):
The expected value of y when $x = 0$. For the food expenditure model, this would be the expected food expenditure at zero income.

Interpreting β_1 and β_2

$$\mathrm{E}(y \mid x) = \beta_1 + \beta_2 x$$

Slope (β_2):

$$\beta_2 = \frac{\Delta\, \mathrm{E}(y \mid x)}{\Delta x}$$

A one-unit increase in x (here, \$100 in weekly income) changes expected food expenditure by β_2 dollars, holding all else constant.

Intercept (β_1):
The expected value of y when $x = 0$. For the food expenditure model, this would be the expected food expenditure at zero income.

Is that meaningful? In this dataset, the lowest income is \$369/week. Nobody has zero income.

Interpreting β_1 and β_2

$$\mathrm{E}(y \mid x) = \beta_1 + \beta_2 x$$

Slope (β_2):

$$\beta_2 = \frac{\Delta\, \mathrm{E}(y \mid x)}{\Delta x}$$

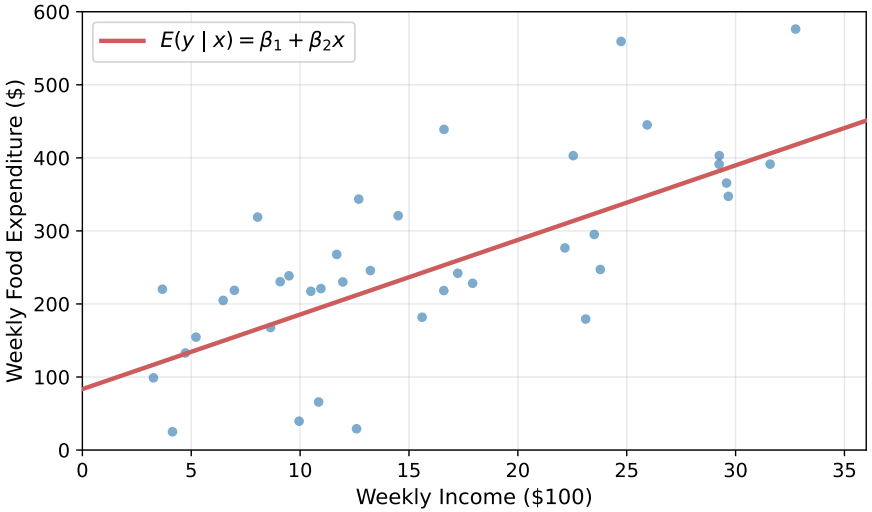
A one-unit increase in x (here, \$100 in weekly income) changes expected food expenditure by β_2 dollars, holding all else constant.

Intercept (β_1):
The expected value of y when $x = 0$. For the food expenditure model, this would be the expected food expenditure at zero income.

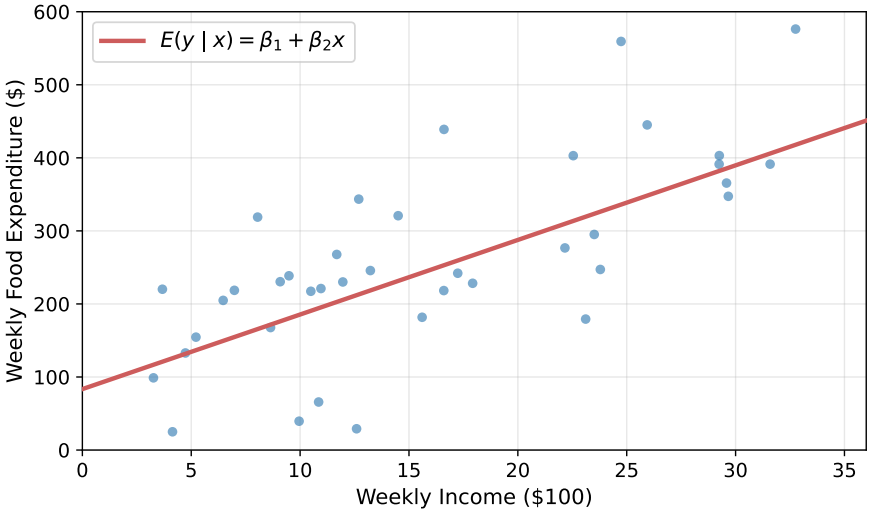
Is that meaningful? In this dataset, the lowest income is \$369/week. Nobody has zero income.

$\Longrightarrow$ The intercept is often just a mathematical anchor for the line, not an economically meaningful quantity.

The Regression Line Through the Data



The Regression Line Through the Data



Each point y_i is the sum of the predicted value on the line ($E(y_i \mid x_i) = \beta_1 + \beta_2 x_i$) plus the error ($e_i$)

Outline

Why Do We Need Assumptions?

The model $y_i = \beta_1 + \beta_2 x_i + e_i$ is just notation. Without assumptions about e_i and x_i, we cannot:

- guarantee that our estimates are correct on average (unbiased)
- calculate how precise our estimates are
- construct confidence intervals or test hypotheses

Why Do We Need Assumptions?

The model $y_i = \beta_1 + \beta_2 x_i + e_i$ is just notation. Without assumptions about e_i and x_i, we cannot:

- guarantee that our estimates are correct on average (unbiased)
- calculate how precise our estimates are
- construct confidence intervals or test hypotheses

We will state six assumptions, labeled SR1–SR6. The first five are needed for OLS to work well. The sixth (normality) is optional.

Why Do We Need Assumptions?

The model $y_i = \beta_1 + \beta_2 x_i + e_i$ is just notation. Without assumptions about e_i and x_i, we cannot:

- guarantee that our estimates are correct on average (unbiased)
- calculate how precise our estimates are
- construct confidence intervals or test hypotheses

We will state six assumptions, labeled SR1–SR6. The first five are needed for OLS to work well. The sixth (normality) is optional.

For each assumption, we will ask: **what goes wrong if it fails?**

SR1: The Linear Model

SR1: Econometric Model

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

SR1: The Linear Model

SR1: Econometric Model

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

This says the relationship between y and x is **linear in the parameters**. The conditional mean $E(y \mid x)$ is a straight line.

SR1: Econometric Model

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

This says the relationship between y and x is **linear in the parameters**. The conditional mean $E(y \mid x)$ is a straight line.

What if it fails?

If the true relationship is, say, $E(y \mid x) = \beta_1 + \beta_2 x^2$, then fitting a straight line produces systematically wrong predictions. The errors e_i would have a pattern (positive at low and high x, negative in the middle), which violates later assumptions too.

SR1: Econometric Model

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

This says the relationship between y and x is **linear in the parameters**. The conditional mean $E(y \mid x)$ is a straight line.

What if it fails?

If the true relationship is, say, $E(y \mid x) = \beta_1 + \beta_2 x^2$, then fitting a straight line produces systematically wrong predictions. The errors e_i would have a pattern (positive at low and high x, negative in the middle), which violates later assumptions too.

$\Longrightarrow$ We will revisit nonlinear models (quadratic, log-linear) in a later lecture.

SR2: Strict Exogeneity

SR2: Zero Conditional Mean

$$\mathsf{E}(e_i \mid x_i) = 0$$

SR2: Strict Exogeneity

SR2: Zero Conditional Mean

$$\mathrm{E}(e_i \mid x_i) = 0$$

In words: knowing the value of income tells you *nothing* about the average error. The "everything else" affecting food expenditure is unrelated to income, on average.

SR2: Strict Exogeneity

SR2: Zero Conditional Mean

$$\mathrm{E}(e_i \mid x_i) = 0$$

In words: knowing the value of income tells you *nothing* about the average error. The "everything else" affecting food expenditure is unrelated to income, on average.

Two implications (necessary but not sufficient):

1. $\mathrm{E}(e_i) = 0$: the unconditional mean of the error is zero
2. $\mathrm{Cov}(e_i, x_i) = 0$: the error is uncorrelated with x

SR2: Strict Exogeneity

SR2: Zero Conditional Mean

$$\mathrm{E}(e_i \mid x_i) = 0$$

In words: knowing the value of income tells you *nothing* about the average error. The "everything else" affecting food expenditure is unrelated to income, on average.

Two implications (necessary but not sufficient):

1. $\mathrm{E}(e_i) = 0$: the unconditional mean of the error is zero
2. $\mathrm{Cov}(e_i, x_i) = 0$: the error is uncorrelated with x

Without SR2, OLS estimates are **biased**: on average, the estimated slope does not equal the true slope. If x is correlated with the error, this bias does not disappear as the sample grows; OLS is also **inconsistent**.

SR2 Violation: An Example

Consider a wage model: $\text{WAGE}_i = \beta_1 + \beta_2 \, \text{EDUC}_i + e_i$

SR2 Violation: An Example

Consider a wage model: $\text{WAGE}_i = \beta_1 + \beta_2 \, \text{EDUC}_i + e_i$

The error e_i contains ability, motivation, family connections. These factors are plausibly **correlated with education**: people with higher ability tend to get more education.

SR2 Violation: An Example

Consider a wage model: $\text{WAGE}_i = \beta_1 + \beta_2 \, \text{EDUC}_i + e_i$

The error e_i contains ability, motivation, family connections. These factors are plausibly **correlated with education**: people with higher ability tend to get more education.

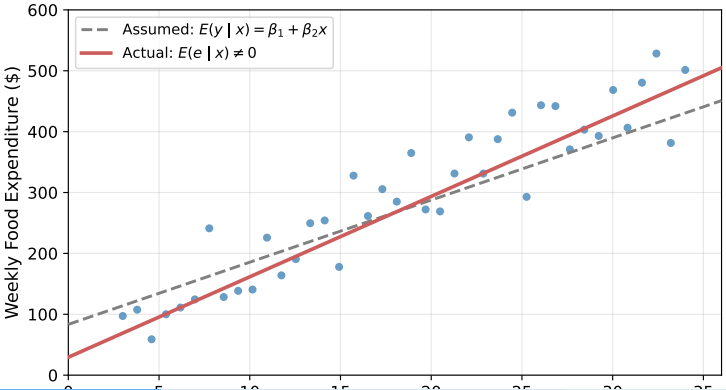
$\implies$ $\text{E}(e_i \mid \text{EDUC}_i) \neq 0$, so EDUC is **endogenous**.

SR2 Violation: An Example

Consider a wage model: $\text{WAGE}_i = \beta_1 + \beta_2\,\text{EDUC}_i + e_i$

The error e_i contains ability, motivation, family connections. These factors are plausibly **correlated with education**: people with higher ability tend to get more education.

$\implies \text{E}(e_i \mid \text{EDUC}_i) \neq 0$, so EDUC is **endogenous**.

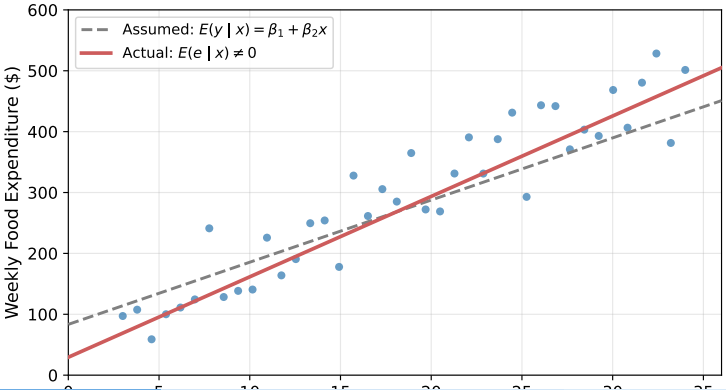


SR2 Violation: An Example

Consider a wage model: $\text{WAGE}_i = \beta_1 + \beta_2\,\text{EDUC}_i + e_i$

The error e_i contains ability, motivation, family connections. These factors are plausibly **correlated with education**: people with higher ability tend to get more education.

$\implies \text{E}(e_i \mid \text{EDUC}_i) \neq 0$, so EDUC is **endogenous**.



SR2: Food Expenditure Model

Is strict exogeneity plausible for our food expenditure model?

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad \text{where } e_i \text{ includes tastes, location, impulses}$$

SR2: Food Expenditure Model

Is strict exogeneity plausible for our food expenditure model?

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad \text{where } e_i \text{ includes tastes, location, impulses}$$

The question: given a household's income, can we predict whether their "everything else" is positive or negative?

SR2: Food Expenditure Model

Is strict exogeneity plausible for our food expenditure model?

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad \text{where } e_i \text{ includes tastes, location, impulses}$$

The question: given a household's income, can we predict whether their "everything else" is positive or negative?

- A household with income \$2000/week that also happens to love fine dining $\implies e_i > 0$
- But is this *correlated with income*? Do higher-income households systematically have more expensive tastes, beyond what income itself predicts?

SR2: Food Expenditure Model

Is strict exogeneity plausible for our food expenditure model?

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad \text{where } e_i \text{ includes tastes, location, impulses}$$

The question: given a household's income, can we predict whether their "everything else" is positive or negative?

- A household with income \$2000/week that also happens to love fine dining $\implies e_i > 0$
- But is this *correlated with income*? Do higher-income households systematically have more expensive tastes, beyond what income itself predicts?

$\implies$ For this model, $\mathrm{E}(e_i \mid x_i) = 0$ is at least arguable. Whether it holds is ultimately an empirical and theoretical judgment, not something we can prove from the data.

SR3: Homoskedasticity

SR3: Conditional Homoskedasticity

$$\text{Var}(e_i \mid x_i) = \sigma^2 \quad \text{(constant)}$$

SR3: Homoskedasticity

SR3: Conditional Homoskedasticity

$$\text{Var}(e_i \mid x_i) = \sigma^2 \quad \text{(constant)}$$

The spread of the error is the **same at every income level**. Whether a household earns \$500/week or \$3000/week, the variability of food expenditure around its conditional mean is the same.

SR3: Homoskedasticity

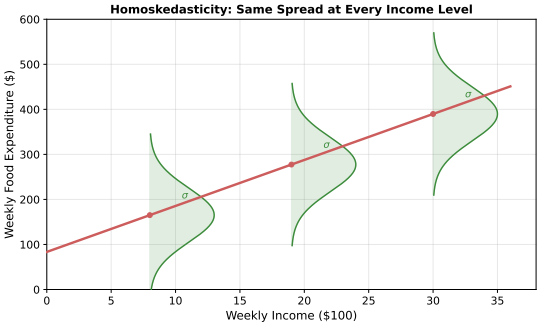
SR3: Conditional Homoskedasticity

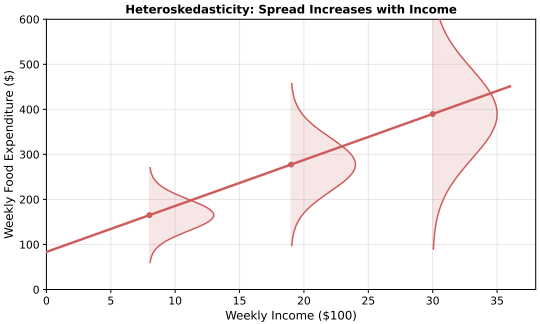
$$\text{Var}(e_i \mid x_i) = \sigma^2 \quad \text{(constant)}$$

The spread of the error is the **same at every income level**. Whether a household earns \$500/week or \$3000/week, the variability of food expenditure around its conditional mean is the same.

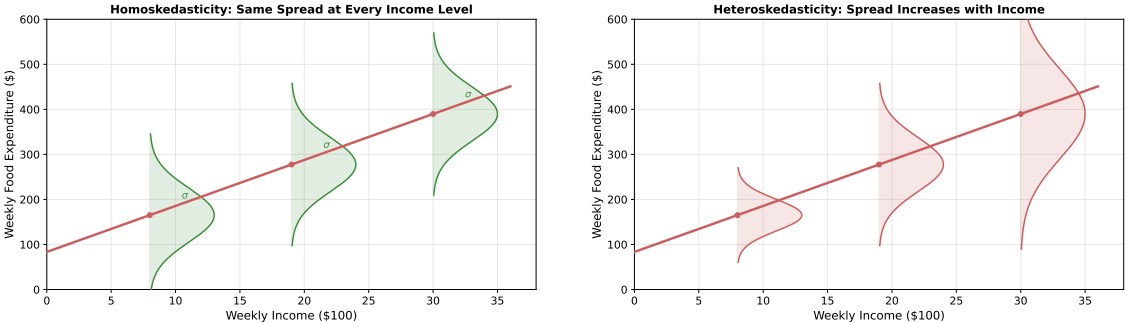
Since $\text{Var}(y_i \mid x_i) = \text{Var}(e_i \mid x_i)$, this also means $\text{Var}(y_i \mid x_i) = \sigma^2$.

SR3: Homoskedasticity vs. Heteroskedasticity



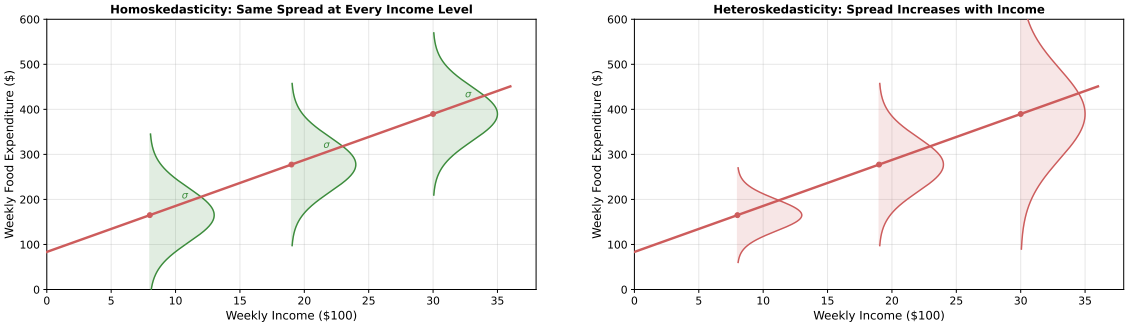


SR3: Homoskedasticity vs. Heteroskedasticity



Left: Same spread at every income level (SR3 holds).
Right: Higher-income households have more variable food spending (SR3 fails). This is **heteroskedasticity**.

SR3: Homoskedasticity vs. Heteroskedasticity



Left: Same spread at every income level (SR3 holds).
Right: Higher-income households have more variable food spending (SR3 fails). This is **heteroskedasticity**.

$\Longrightarrow$ If SR3 fails, OLS estimates are still unbiased, but their **standard errors are wrong**. Confidence intervals and hypothesis tests become unreliable.

SR4: Uncorrelated Errors

SR4: Uncorrelated Errors

$$\text{Cov}(e_i, e_j \mid x_i, x_j) = 0, \quad i \neq j$$

SR4: Uncorrelated Errors

SR4: Uncorrelated Errors

$$\text{Cov}(e_i, e_j \mid x_i, x_j) = 0, \quad i \neq j$$

Knowing that household i spent unusually much on food tells you *nothing* about whether household j did the same.

SR4: Uncorrelated Errors

$$\text{Cov}(e_i, e_j \mid x_i, x_j) = 0, \quad i \neq j$$

Knowing that household i spent unusually much on food tells you *nothing* about whether household j did the same.

When does this fail?

- **Time-series data:** this quarter's GDP shock affects next quarter's ($\implies$ serial correlation)
- **Spatial data:** households in the same neighborhood share local grocery prices ($\implies$ spatial correlation / clustering)

SR4: Uncorrelated Errors

$$\text{Cov}(e_i, e_j \mid x_i, x_j) = 0, \quad i \neq j$$

Knowing that household i spent unusually much on food tells you *nothing* about whether household j did the same.

When does this fail?

- **Time-series data:** this quarter's GDP shock affects next quarter's ($\implies$ serial correlation)
- **Spatial data:** households in the same neighborhood share local grocery prices ($\implies$ spatial correlation / clustering)

For our cross-sectional food data with randomly sampled households, SR4 is plausible.

SR4: Uncorrelated Errors

SR4: Uncorrelated Errors

$$\text{Cov}(e_i, e_j \mid x_i, x_j) = 0, \quad i \neq j$$

Knowing that household i spent unusually much on food tells you *nothing* about whether household j did the same.

When does this fail?

- **Time-series data:** this quarter's GDP shock affects next quarter's ($\implies$ serial correlation)
- **Spatial data:** households in the same neighborhood share local grocery prices ($\implies$ spatial correlation / clustering)

For our cross-sectional food data with randomly sampled households, SR4 is plausible.

$\implies$ If SR4 fails, the consequences are similar to SR3: OLS is still unbiased, but standard errors are wrong.

SR5: Variation in x

SR5: Explanatory Variable Must Vary

x_i takes at least two different values in the sample.

SR5: Variation in x

SR5: Explanatory Variable Must Vary

x_i takes at least two different values in the sample.

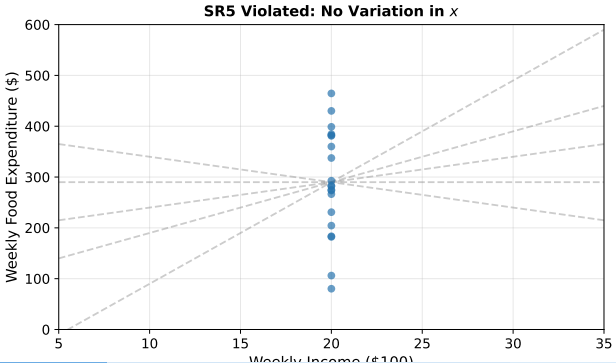
To estimate a slope, you need at least two distinct x values. Otherwise, you are trying to draw a line through a single vertical strip of points.

SR5: Variation in x

SR5: Explanatory Variable Must Vary

x_i takes at least two different values in the sample.

To estimate a slope, you need at least two distinct x values. Otherwise, you are trying to draw a line through a single vertical strip of points.

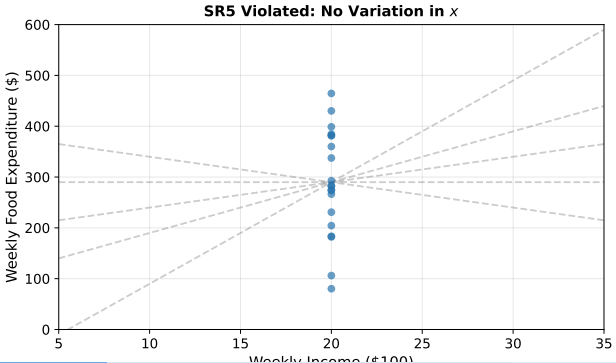


SR5: Variation in x

SR5: Explanatory Variable Must Vary

x_i takes at least two different values in the sample.

To estimate a slope, you need at least two distinct x values. Otherwise, you are trying to draw a line through a single vertical strip of points.

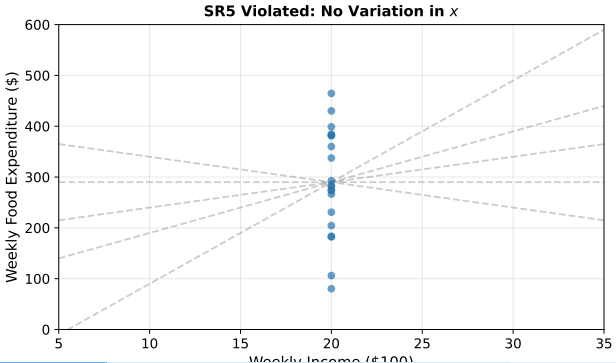


SR5: Variation in x

SR5: Explanatory Variable Must Vary

x_i takes at least two different values in the sample.

To estimate a slope, you need at least two distinct x values. Otherwise, you are trying to draw a line through a single vertical strip of points.



SR6: Normality (Optional)

SR6: Error Normality

$$e_i \mid x_i \sim N(0, \sigma^2)$$

SR6: Normality (Optional)

SR6: Error Normality

$$e_i \mid x_i \sim N(0, \sigma^2)$$

Combined with the other assumptions, this gives:

$$y_i \mid x_i \sim N(\beta_1 + \beta_2 x_i,\ \sigma^2)$$

SR6: Normality (Optional)

SR6: Error Normality

$$e_i \mid x_i \sim N(0, \sigma^2)$$

Combined with the other assumptions, this gives:

$$y_i \mid x_i \sim N(\beta_1 + \beta_2 x_i,\ \sigma^2)$$

Why normal? The error e_i is the sum of many small, unrelated factors (tastes, location, impulse buys, …). The intuition behind the CLT suggests that such sums tend toward a normal distribution.

SR6: Normality (Optional)

SR6: Error Normality

$$e_i \mid x_i \sim N(0, \sigma^2)$$

Combined with the other assumptions, this gives:

$$y_i \mid x_i \sim N(\beta_1 + \beta_2 x_i,\ \sigma^2)$$

Why normal? The error e_i is the sum of many small, unrelated factors (tastes, location, impulse buys, …). The intuition behind the CLT suggests that such sums tend toward a normal distribution.

Why optional? If N is large enough, the CLT also makes the *estimators* b_1 and b_2 approximately normal, even when the e_i are not. SR6 is most useful in small samples.

All Six Assumptions at a Glance

Label	Name	Statement
SR1	Linear model	$y_i = \beta_1 + \beta_2 x_i + e_i$
SR2	Zero conditional mean	$\mathrm{E}(e_i \mid x_i) = 0$
SR3	Homoskedasticity	$\mathrm{Var}(e_i \mid x_i) = \sigma^2$
SR4	Uncorrelated errors	$\mathrm{Cov}(e_i, e_j \mid x_i, x_j) = 0$ for $i \neq j$
SR5	x varies	x_i takes ≥ 2 distinct values
SR6	Normality (*opt.*)	$e_i \mid x_i \sim N(0, \sigma^2)$

All Six Assumptions at a Glance

Label	Name	Statement
SR1	Linear model	$y_i = \beta_1 + \beta_2 x_i + e_i$
SR2	Zero conditional mean	$\mathrm{E}(e_i \mid x_i) = 0$
SR3	Homoskedasticity	$\mathrm{Var}(e_i \mid x_i) = \sigma^2$
SR4	Uncorrelated errors	$\mathrm{Cov}(e_i, e_j \mid x_i, x_j) = 0$ for $i \neq j$
SR5	x varies	x_i takes ≥ 2 distinct values
SR6	Normality (*opt.*)	$e_i \mid x_i \sim N(0, \sigma^2)$

SR1–SR5 are needed for OLS to be the Best Linear Unbiased Estimator (Gauss–Markov theorem, coming in a future lecture). SR6 adds exact distributional results for small samples.

What Goes Wrong: A Summary

Violated	Example	Consequence
SR1	True relationship is quadratic	Systematic prediction errors
SR2	Ability correlated with education	OLS is **biased**
SR3	Rich households have more variable spending	Standard errors are wrong
SR4	Neighboring households share shocks	Standard errors are wrong
SR5	All households have same income	Slope is unidentified
SR6	Errors are skewed	t-tests unreliable in small samples

What Goes Wrong: A Summary

Violated	Example	Consequence
SR1	True relationship is quadratic	Systematic prediction errors
SR2	Ability correlated with education	OLS is **biased**
SR3	Rich households have more variable spending	Standard errors are wrong
SR4	Neighboring households share shocks	Standard errors are wrong
SR5	All households have same income	Slope is unidentified
SR6	Errors are skewed	t-tests unreliable in small samples

$\Longrightarrow$ SR2 is the most consequential: it determines whether OLS gives unbiased answers. The rest of the course will develop tools for detecting and correcting these violations.

Outline

Two Vocabularies

	Population (unknown)	Sample (computed)
Intercept	β_1	b_1
Slope	β_2	b_2
Error	e_i	$\hat{e}_i$ (residual)
Regression function	$\mathrm{E}(y \mid x) = \beta_1 + \beta_2 x$	$\hat{y} = b_1 + b_2 x$
Error variance	σ^2	$\hat{\sigma}^2$

Two Vocabularies

	Population (unknown)	Sample (computed)
Intercept	β_1	b_1
Slope	β_2	b_2
Error	e_i	$\hat{e}_i$ (residual)
Regression function	$\mathrm{E}(y \mid x) = \beta_1 + \beta_2 x$	$\hat{y} = b_1 + b_2 x$
Error variance	σ^2	$\hat{\sigma}^2$

Parameters ($\beta_1, \beta_2, \sigma^2$) are fixed but unknown constants that describe the population.

Estimates ($b_1, b_2, \hat{\sigma}^2$) are numbers we compute from a particular sample. Different samples give different estimates.

Two Vocabularies

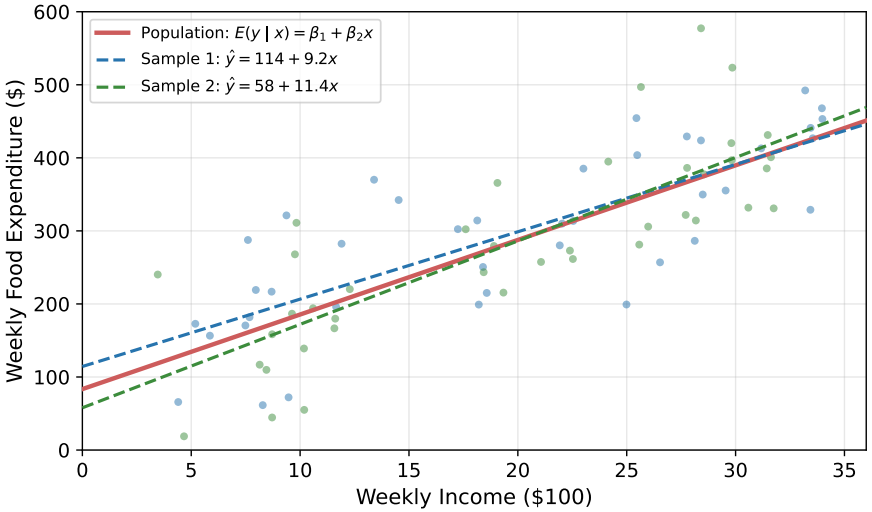
	Population (unknown)	Sample (computed)
Intercept	β_1	b_1
Slope	β_2	b_2
Error	e_i	$\hat{e}_i$ (residual)
Regression function	$\mathrm{E}(y \mid x) = \beta_1 + \beta_2 x$	$\hat{y} = b_1 + b_2 x$
Error variance	σ^2	$\hat{\sigma}^2$

Parameters ($\beta_1, \beta_2, \sigma^2$) are fixed but unknown constants that describe the population.

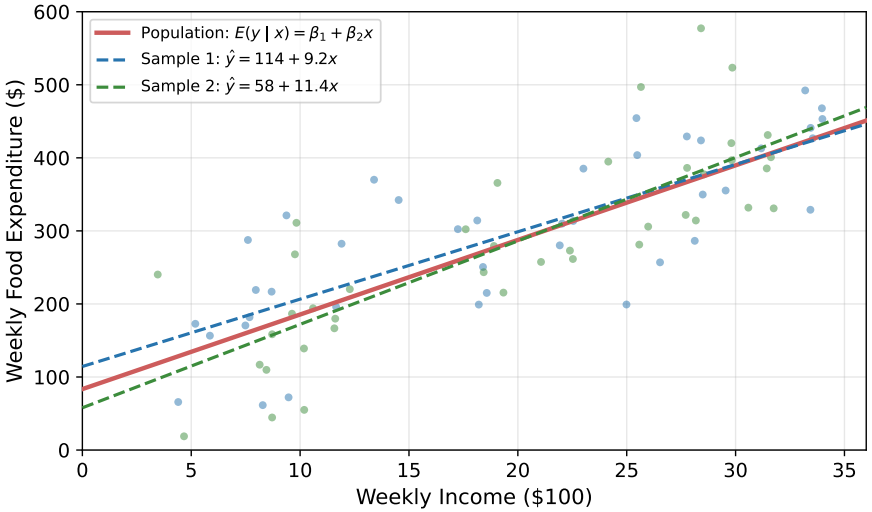
Estimates ($b_1, b_2, \hat{\sigma}^2$) are numbers we compute from a particular sample. Different samples give different estimates.

Estimators (b_1, b_2 as formulas, not numbers) are **random variables** whose properties (bias, variance) we can study.

Different Samples, Different Estimates

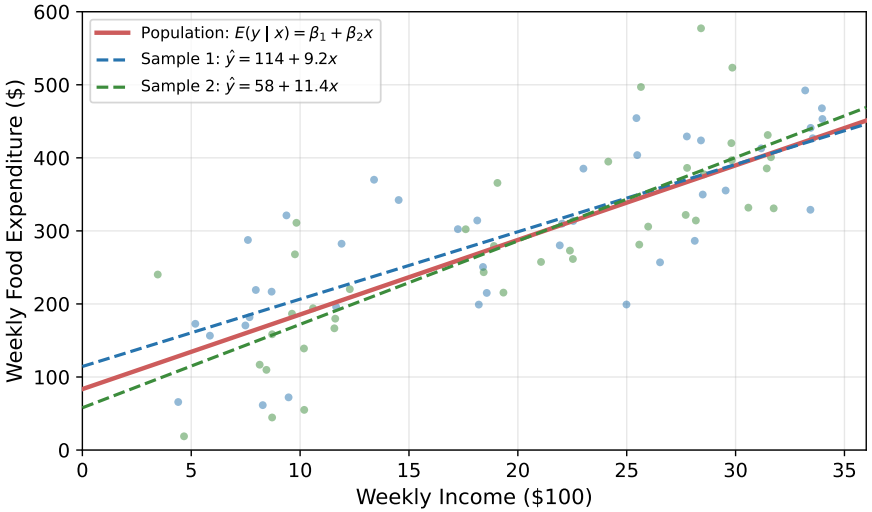


Different Samples, Different Estimates



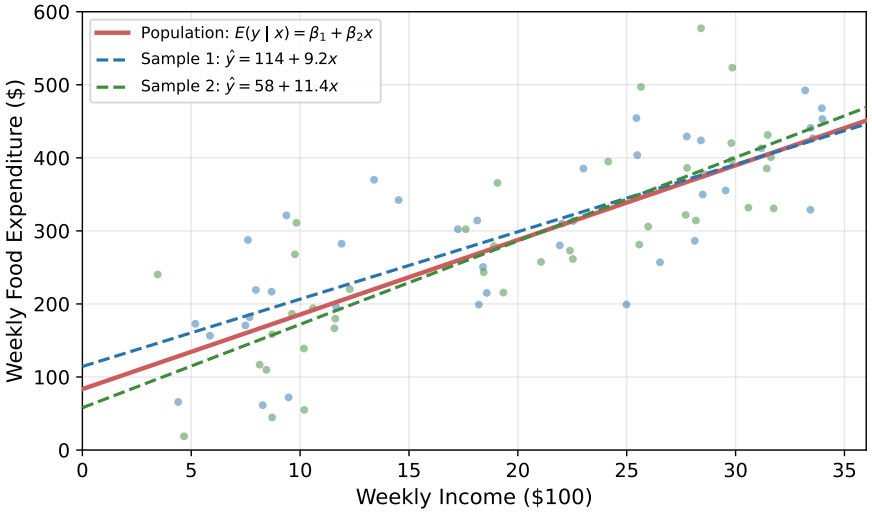
The red line is the true population regression function $E(\quad) = \beta_1 + \beta_2$

Different Samples, Different Estimates



The red line is the true population regression function $E(\quad) = \beta_1 + \beta_2$

Different Samples, Different Estimates



The red line is the true population regression function $E(\quad) = \beta_1 + \beta_2$

Preview: From Model to Estimation

Today we built the model. The next lecture answers: **how do we estimate β_1 and β_2?**

Preview: From Model to Estimation

Today we built the model. The next lecture answers: **how do we estimate β_1 and β_2?**

The idea behind **Ordinary Least Squares (OLS)**: choose b_1 and b_2 to minimize the sum of squared residuals:

$$\min_{b_1, b_2} \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2$$

Preview: From Model to Estimation

Today we built the model. The next lecture answers: **how do we estimate β_1 and β_2?**

The idea behind **Ordinary Least Squares (OLS)**: choose b_1 and b_2 to minimize the sum of squared residuals:

$$\min_{b_1, b_2} \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2$$

How do we solve this? That is next time.

Preview: From Model to Estimation

Today we built the model. The next lecture answers: **how do we estimate β_1 and β_2?**

The idea behind **Ordinary Least Squares (OLS)**: choose b_1 and b_2 to minimize the sum of squared residuals:

$$\min_{b_1, b_2} \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2$$

How do we solve this? That is next time.

$\Longrightarrow$ Under SR1–SR5, the OLS estimators have desirable properties that we will prove.

Thank you!
jakeanderson@g.ucla.edu