Introduction to Heteroskedasticity

Non-Constant Variance and What to Do About It

Jake Anderson

March 3, 2026

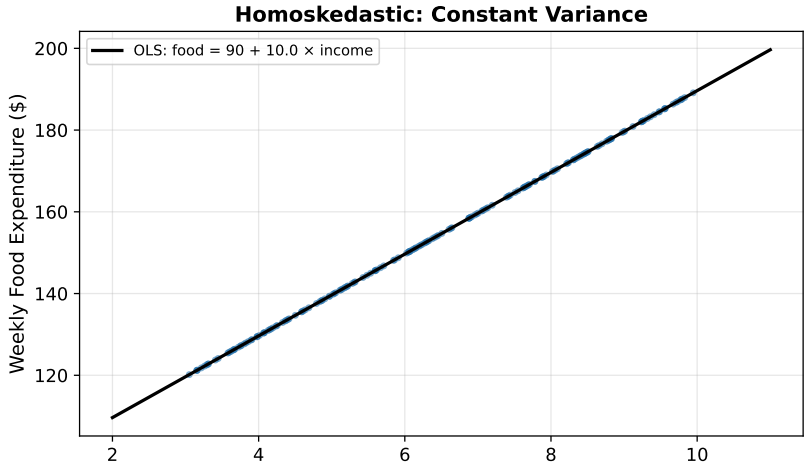
Outline

1. Motivation
2. What Goes Wrong
3. Detecting Heteroskedasticity
4. Fixing It: Robust Standard Errors
5. Fixing It: WLS
6. Fixing It: FGLS
7. Worked Example: Food Expenditure (HGL 8.1)

Food Expenditure vs. Income: The Simple Case

Suppose we model weekly food expenditure as a function of household income:

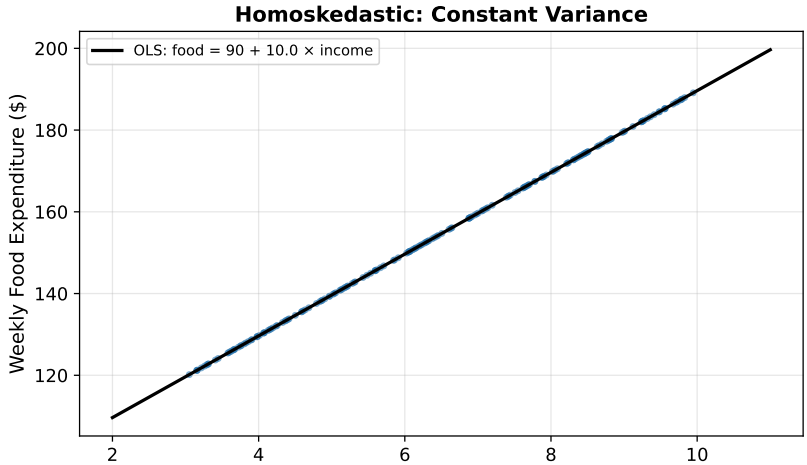
$$\text{food}_i = \beta_0 + \beta_1 \, \text{income}_i + e_i$$



Food Expenditure vs. Income: The Simple Case

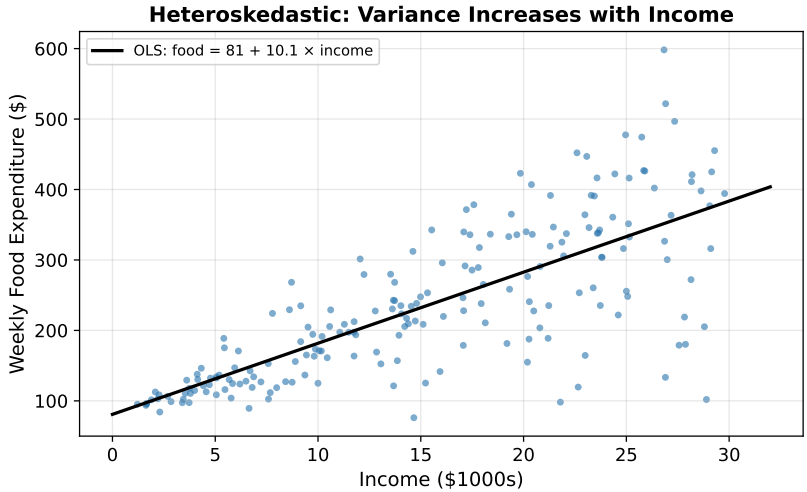
Suppose we model weekly food expenditure as a function of household income:

$$\text{food}_i = \beta_0 + \beta_1 \, \text{income}_i + e_i$$



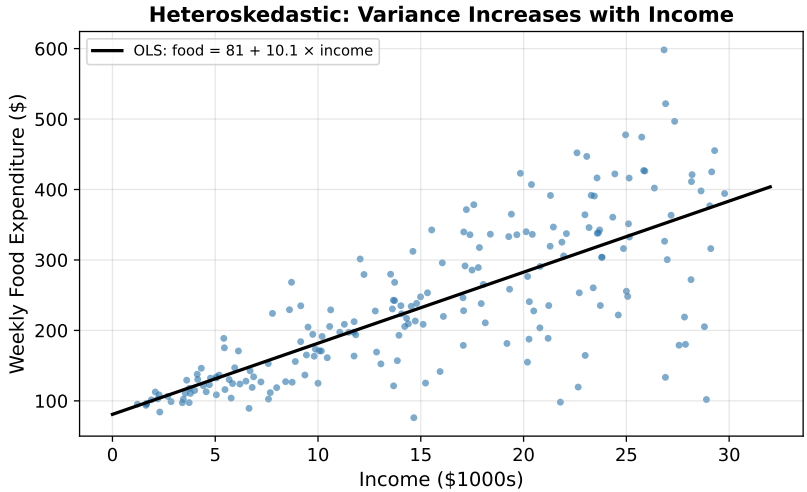
What If the Spread Changes?

Now consider the full income range:



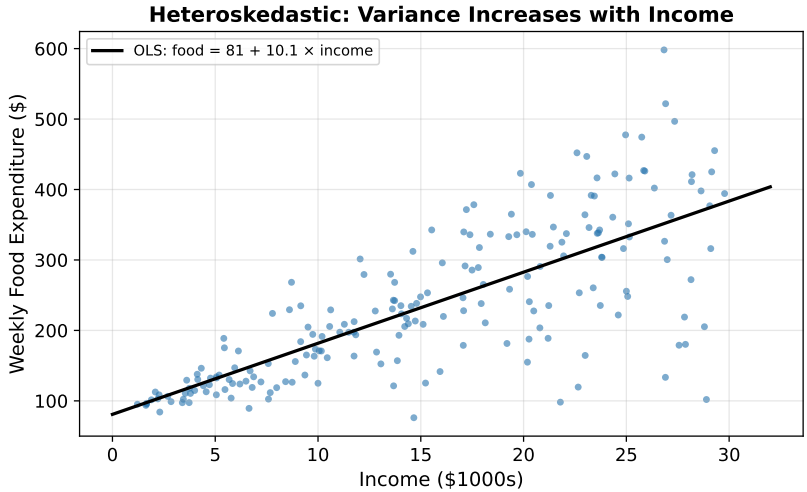
What If the Spread Changes?

Now consider the full income range:



What If the Spread Changes?

Now consider the full income range:



Why Does Variance Change?

Heteroskedasticity is common whenever:

- **Scale effects:** Higher income $\implies$ more discretion in spending
- **Learning:** Experienced firms have tighter cost control than new ones
- **Aggregation:** County-level data averages over different population sizes
- **Model misspecification:** Omitted variables whose effect grows with x

Why Does Variance Change?

Heteroskedasticity is common whenever:

- **Scale effects:** Higher income $\implies$ more discretion in spending
- **Learning:** Experienced firms have tighter cost control than new ones
- **Aggregation:** County-level data averages over different population sizes
- **Model misspecification:** Omitted variables whose effect grows with x

Formally, we write:

$$\text{Var}(e_i \mid x_i) = \sigma_i^2$$

where σ_i^2 depends on x_i (or some other observable). The assumption $\sigma_i^2 = \sigma^2$ for all i is what we call **homoskedasticity**.

OLS Under Heteroskedasticity: The Good News

Even with heteroskedasticity, OLS is still:

- **Unbiased:** $\mathrm{E}[\hat{\beta}] = \beta$ (requires only $\mathrm{E}[e_i \mid x_i] = 0$)
- **Consistent:** $\hat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$

OLS Under Heteroskedasticity: The Good News

Even with heteroskedasticity, OLS is still:

- **Unbiased:** $\mathrm{E}[\hat{\beta}] = \beta$ (requires only $\mathrm{E}[e_i \mid x_i] = 0$)
- **Consistent:** $\hat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$

Why? The OLS formula:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x}) e_i}{\sum_i (x_i - \bar{x})^2}$$

depends on $\mathrm{E}[e_i \mid x_i] = 0$, not on $\mathrm{Var}(e_i \mid x_i) = \sigma^2$.

OLS Under Heteroskedasticity: The Good News

Even with heteroskedasticity, OLS is still:

- **Unbiased:** $\mathrm{E}[\hat{\beta}] = \beta$ (requires only $\mathrm{E}[e_i \mid x_i] = 0$)
- **Consistent:** $\hat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$

Why? The OLS formula:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x}) e_i}{\sum_i (x_i - \bar{x})^2}$$

depends on $\mathrm{E}[e_i \mid x_i] = 0$, not on $\mathrm{Var}(e_i \mid x_i) = \sigma^2$.

$\Longrightarrow$ The coefficient estimates themselves are fine. The problem is elsewhere.

OLS Under Heteroskedasticity: The Bad News

The usual OLS standard error formula **assumes homoskedasticity**:

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2} \quad \text{where } \hat{\sigma}^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

OLS Under Heteroskedasticity: The Bad News

The usual OLS standard error formula **assumes homoskedasticity**:

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2} \quad \text{where } \hat{\sigma}^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

This averages all squared residuals into a **single** $\hat{\sigma}^2$. But if variance differs across observations, this average is wrong:

- Observations with **large** variance get too little weight in the SE
- Observations with **small** variance get too much weight

OLS Under Heteroskedasticity: The Bad News

The usual OLS standard error formula **assumes homoskedasticity**:

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2} \quad \text{where } \hat{\sigma}^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

This averages all squared residuals into a **single** $\hat{\sigma}^2$. But if variance differs across observations, this average is wrong:

- Observations with **large** variance get too little weight in the SE
- Observations with **small** variance get too much weight

Consequences:

1. Standard errors are **biased** (could be too small or too large)
2. t-statistics and p-values are **unreliable**
3. Confidence intervals have **wrong coverage**
4. OLS is no longer **BLUE** (Best Linear Unbiased Estimator)

OLS Under Heteroskedasticity: The Bad News

The usual OLS standard error formula **assumes homoskedasticity**:

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2} \quad \text{where } \hat{\sigma}^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

This averages all squared residuals into a **single** $\hat{\sigma}^2$. But if variance differs across observations, this average is wrong:

- Observations with **large** variance get too little weight in the SE
- Observations with **small** variance get too much weight

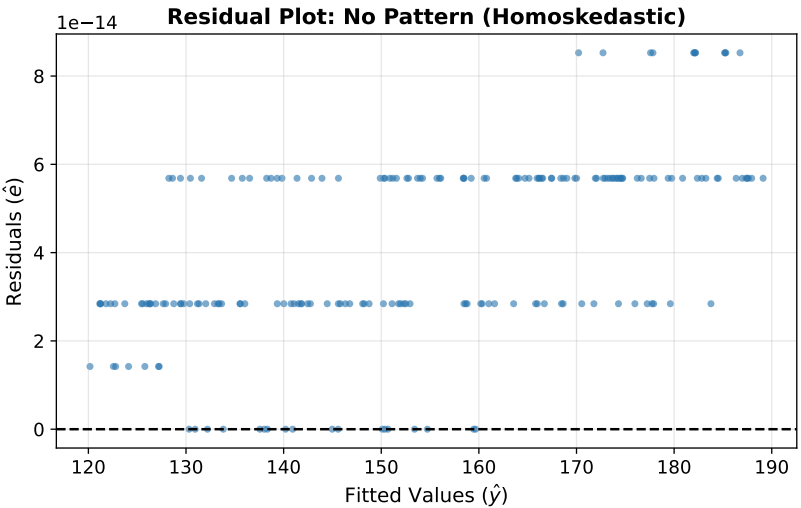
Consequences:

1. Standard errors are **biased** (could be too small or too large)
2. t-statistics and p-values are **unreliable**
3. Confidence intervals have **wrong coverage**
4. OLS is no longer **BLUE** (Best Linear Unbiased Estimator)

$\Longrightarrow$ You cannot trust hypothesis tests from OLS when heteroskedasticity is present.

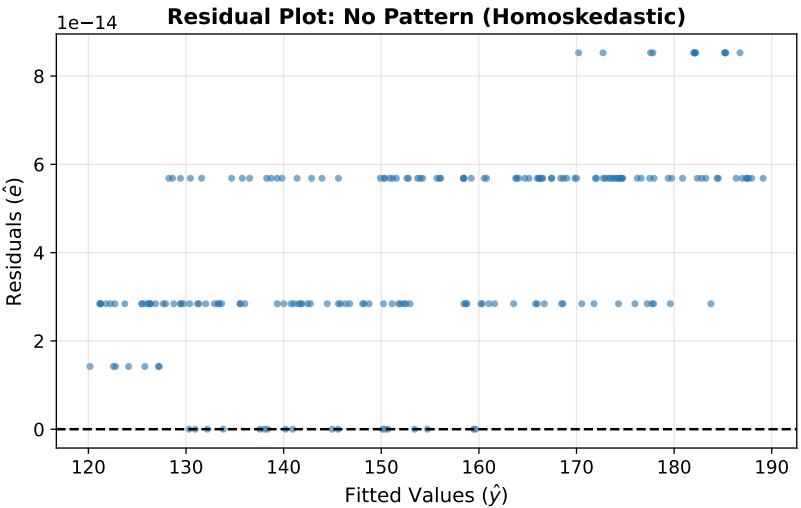
Visual Detection: Residual Plots

Plot residuals $\hat{e}_i$ against fitted values $\hat{y}_i$ (or against x_i):

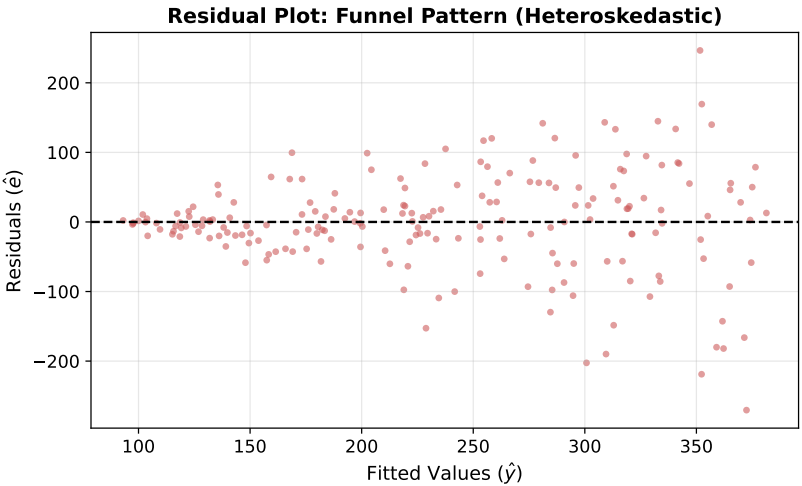


Visual Detection: Residual Plots

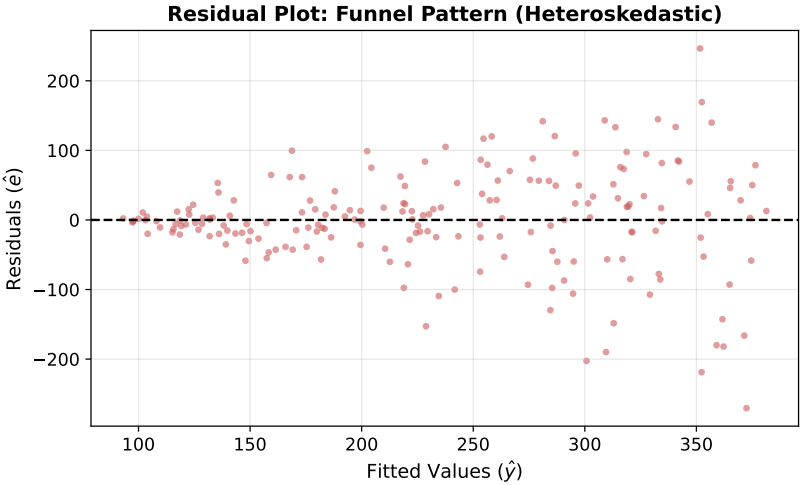
Plot residuals $\hat{e}_i$ against fitted values $\hat{y}_i$ (or against x_i):



Residual Plot: The Funnel



Residual Plot: The Funnel



Residuals fan out as $\hat{y}$ increases $\implies$ strong evidence of heteroskedasticity. This "funnel" or

Breusch-Pagan (BP) Test

Idea: If heteroskedasticity is present, squared residuals should be related to x.

Breusch-Pagan (BP) Test

Idea: If heteroskedasticity is present, squared residuals should be related to x.

Steps:

1. Run OLS: $y_i = \beta_0 + \beta_1 x_i + e_i$. Obtain residuals $\hat{e}_i$.

Breusch-Pagan (BP) Test

Idea: If heteroskedasticity is present, squared residuals should be related to x.

Steps:

1. Run OLS: $y_i = \beta_0 + \beta_1 x_i + e_i$. Obtain residuals $\hat{e}_i$.
2. Run **auxiliary regression**: $\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + v_i$

Breusch-Pagan (BP) Test

Idea: If heteroskedasticity is present, squared residuals should be related to x.

Steps:

1. Run OLS: $y_i = \beta_0 + \beta_1 x_i + e_i$. Obtain residuals $\hat{e}_i$.
2. Run **auxiliary regression**: $\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + v_i$
3. Test H_0: $\gamma_1 = 0 \iff$ homoskedasticity
4. Test statistic: $\text{BP} = n \cdot R^2_{\text{aux}} \sim \chi^2(1)$ under H_0

Breusch-Pagan (BP) Test

Idea: If heteroskedasticity is present, squared residuals should be related to x.

Steps:

1. Run OLS: $y_i = \beta_0 + \beta_1 x_i + e_i$. Obtain residuals $\hat{e}_i$.
2. Run **auxiliary regression**: $\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + v_i$
3. Test H_0: $\gamma_1 = 0 \iff$ homoskedasticity
4. Test statistic: $\text{BP} = n \cdot R^2_{\text{aux}} \sim \chi^2(1)$ under H_0

If BP is large (small p-value) $\implies$ reject homoskedasticity.

Breusch-Pagan (BP) Test

Idea: If heteroskedasticity is present, squared residuals should be related to x.

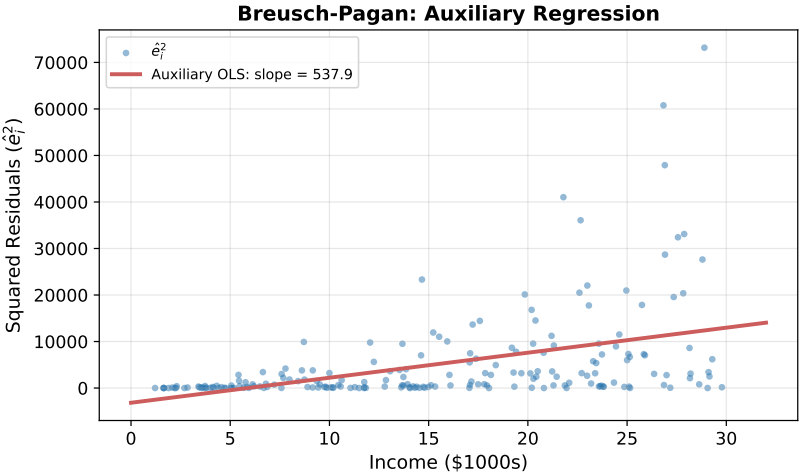
Steps:

1. Run OLS: $y_i = \beta_0 + \beta_1 x_i + e_i$. Obtain residuals $\hat{e}_i$.
2. Run **auxiliary regression**: $\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + v_i$
3. Test H_0: $\gamma_1 = 0 \iff$ homoskedasticity
4. Test statistic: $\text{BP} = n \cdot R^2_{\text{aux}} \sim \chi^2(1)$ under H_0

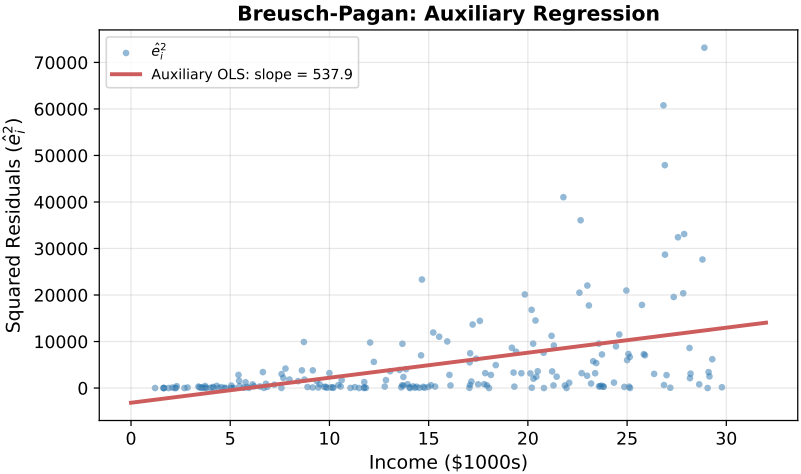
If BP is large (small p-value) $\implies$ reject homoskedasticity.

With multiple regressors: include all x's in the auxiliary regression. Then $\text{BP} \sim \chi^2(k)$ where $k =$ number of regressors.

BP Test: The Auxiliary Regression



BP Test: The Auxiliary Regression



The auxiliary regression line slopes upward $\Longrightarrow$ variance increases with income. A flat line would indicate homoskedasticity.

White Test

The BP test assumes σ_i^2 is a **linear** function of x. The White test is more flexible:

White Test

The BP test assumes σ_i^2 is a **linear** function of x. The White test is more flexible:

Auxiliary regression:

$$\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + v_i$$

White Test

The BP test assumes σ_i^2 is a **linear** function of x. The White test is more flexible:

Auxiliary regression:

$$\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + v_i$$

With multiple regressors, include all x's, their squares, and cross-products.

White Test

The BP test assumes σ_i^2 is a **linear** function of x. The White test is more flexible:

Auxiliary regression:

$$\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + v_i$$

With multiple regressors, include all x's, their squares, and cross-products.

Test statistic: White $= n \cdot R^2_{\text{aux}} \sim \chi^2(q)$
where q = number of regressors in the auxiliary regression (excluding intercept).

White Test

The BP test assumes σ_i^2 is a **linear** function of x. The White test is more flexible:

Auxiliary regression:

$$\hat{e}_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + v_i$$

With multiple regressors, include all x's, their squares, and cross-products.

Test statistic: White $= n \cdot R^2_{\text{aux}} \sim \chi^2(q)$
where q = number of regressors in the auxiliary regression (excluding intercept).

Advantage: detects **nonlinear** forms of heteroskedasticity.
Disadvantage: with many regressors, the auxiliary regression has many terms $\implies$ low power.

Goldfeld-Quandt (GQ) Test

Idea: If variance increases with x, the residual variance in the "high x" group should exceed the "low x" group.

Goldfeld-Quandt (GQ) Test

Idea: If variance increases with x, the residual variance in the "high x" group should exceed the "low x" group.

Steps:

1. Sort observations by x_i
2. Drop the middle c observations (typically $c \approx n/5$)
3. Run **separate OLS** on the bottom group and the top group
4. Compute $\text{GQ} = \dfrac{\text{SSE}_{\text{top}}/(n_{\text{top}} - k)}{\text{SSE}_{\text{bottom}}/(n_{\text{bottom}} - k)}$

Goldfeld-Quandt (GQ) Test

Idea: If variance increases with x, the residual variance in the "high x" group should exceed the "low x" group.

Steps:

1. Sort observations by x_i
2. Drop the middle c observations (typically $c \approx n/5$)
3. Run **separate OLS** on the bottom group and the top group
4. Compute $\text{GQ} = \dfrac{\text{SSE}_{\text{top}}/(n_{\text{top}} - k)}{\text{SSE}_{\text{bottom}}/(n_{\text{bottom}} - k)}$

Under H_0 (homoskedasticity): $\text{GQ} \sim F(n_{\text{top}} - k,\ n_{\text{bottom}} - k)$

Goldfeld-Quandt (GQ) Test

Idea: If variance increases with x, the residual variance in the "high x" group should exceed the "low x" group.

Steps:

1. Sort observations by x_i
2. Drop the middle c observations (typically $c \approx n/5$)
3. Run **separate OLS** on the bottom group and the top group
4. Compute $\text{GQ} = \dfrac{\text{SSE}_{\text{top}}/(n_{\text{top}} - k)}{\text{SSE}_{\text{bottom}}/(n_{\text{bottom}} - k)}$

Under H_0 (homoskedasticity): $\text{GQ} \sim F(n_{\text{top}} - k,\ n_{\text{bottom}} - k)$

$\Longrightarrow$ Simple and intuitive, but requires choosing which variable to sort by.

The Easiest Fix: Robust SEs

Instead of assuming $\text{Var}(e_i \mid x_i) = \sigma^2$, estimate a **heteroskedasticity-consistent** (HC) variance:

The Easiest Fix: Robust SEs

Instead of assuming $\text{Var}(e_i \mid x_i) = \sigma^2$, estimate a **heteroskedasticity-consistent** (HC) variance:

HC0 (White):

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})^2 \, \hat{e}_i^2}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$$

The Easiest Fix: Robust SEs

Instead of assuming $\text{Var}(e_i \mid x_i) = \sigma^2$, estimate a **heteroskedasticity-consistent** (HC) variance:

HC0 (White):

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})^2 \, \hat{e}_i^2}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$$

Each observation contributes its **own** squared residual $\hat{e}_i^2$ rather than the pooled $\hat{\sigma}^2$.

The Easiest Fix: Robust SEs

Instead of assuming $\text{Var}(e_i \mid x_i) = \sigma^2$, estimate a **heteroskedasticity-consistent** (HC) variance:

HC0 (White):

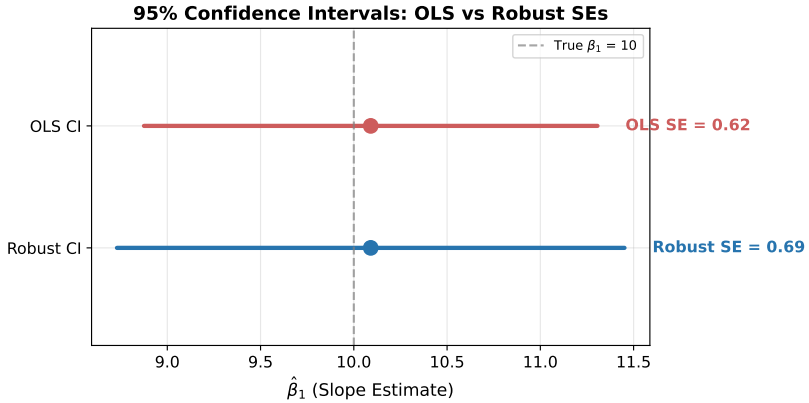
$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})^2 \, \hat{e}_i^2}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$$

Each observation contributes its **own** squared residual $\hat{e}_i^2$ rather than the pooled $\hat{\sigma}^2$.

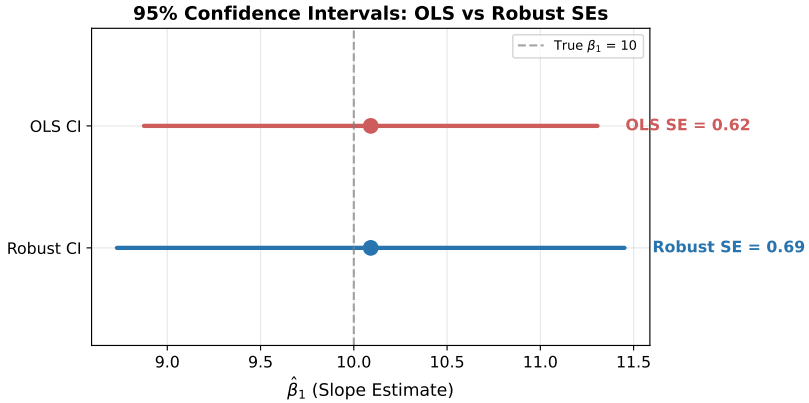
HC1 (small-sample correction): multiply by $\dfrac{n}{n-k}$

HC1 is the default "robust" SE in most software. With large n, HC0 and HC1 are nearly identical.

Robust SEs: What Changes?



Robust SEs: What Changes?



- The **point estimate** $\hat{\beta}_1$ does not change (same OLS regression)
- Only the **standard error** (and therefore the CI width) changes
- With heteroskedasticity, OLS SEs are typically **too small** $\implies$ robust CIs are wider

Robust SEs in R

Using the car and lmtest packages:

```
library(car) library(lmtest)
model <- lm(food   income, data = food_data)
# Usual OLS standard errors summary(model)
# Robust (HC1) standard errors coeftest(model, vcov. = hccm(model, type = "hc1"))
```

Robust SEs in R

Using the `car` and `lmtest` packages:

```
library(car) library(lmtest)
model <- lm(food   income, data = food_data)
# Usual OLS standard errors summary(model)
# Robust (HC1) standard errors coeftest(model, vcov. = hccm(model, type = "hc1"))
```

$\Longrightarrow$ Always report robust SEs unless you have strong reason to believe homoskedasticity holds. Many applied researchers use robust SEs by default.

Weighted Least Squares: The Idea

Robust SEs fix inference but don't improve **efficiency**. Can we do better?

Weighted Least Squares: The Idea

Robust SEs fix inference but don't improve **efficiency**. Can we do better?

Intuition: Observations with small variance carry more information about the regression line. We should trust them more.

Weighted Least Squares: The Idea

Robust SEs fix inference but don't improve **efficiency**. Can we do better?

Intuition: Observations with small variance carry more information about the regression line. We should trust them more.

WLS weights: $w_i = 1/\sigma_i^2$

- Low-variance observations $\implies$ large weight
- High-variance observations $\implies$ small weight

Weighted Least Squares: The Idea

Robust SEs fix inference but don't improve **efficiency**. Can we do better?

Intuition: Observations with small variance carry more information about the regression line. We should trust them more.

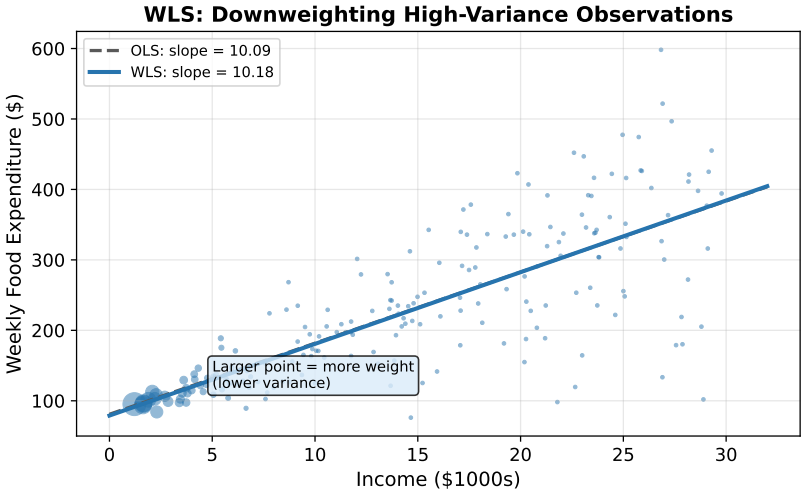
WLS weights: $w_i = 1/\sigma_i^2$

- Low-variance observations $\implies$ large weight
- High-variance observations $\implies$ small weight

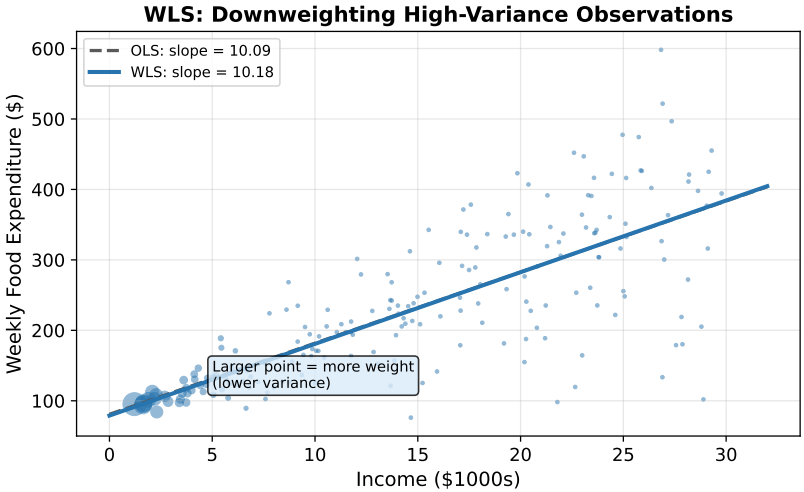
WLS minimizes:

$$\sum_i w_i(y_i - \beta_0 - \beta_1 x_i)^2 = \sum_i \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma_i^2}$$

WLS: Visual Intuition



WLS: Visual Intuition



Larger points have more weight (lower variance). The WLS line is pulled toward the low-income,

WLS as a Transformation

Divide the entire equation by σ_i:

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{e_i}{\sigma_i}$$

WLS as a Transformation

Divide the entire equation by σ_i:

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{e_i}{\sigma_i}$$

Now the transformed error $e_i^* = e_i/\sigma_i$ has:

$$\text{Var}(e_i^*) = \frac{\text{Var}(e_i)}{\sigma_i^2} = \frac{\sigma_i^2}{\sigma_i^2} = 1$$

WLS as a Transformation

Divide the entire equation by σ_i:

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{e_i}{\sigma_i}$$

Now the transformed error $e_i^* = e_i/\sigma_i$ has:

$$\text{Var}(e_i^*) = \frac{\text{Var}(e_i)}{\sigma_i^2} = \frac{\sigma_i^2}{\sigma_i^2} = 1$$

$\implies$ OLS on the transformed model is **homoskedastic**. This is exactly WLS.

WLS as a Transformation

Divide the entire equation by σ_i:

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{e_i}{\sigma_i}$$

Now the transformed error $e_i^* = e_i/\sigma_i$ has:

$$\text{Var}(e_i^*) = \frac{\text{Var}(e_i)}{\sigma_i^2} = \frac{\sigma_i^2}{\sigma_i^2} = 1$$

$\implies$ OLS on the transformed model is **homoskedastic**. This is exactly WLS.

WLS is **BLUE** when the weights are correct. It is more efficient than OLS.

WLS: The Catch

WLS requires knowing σ_i^2 (or at least σ_i^2 up to a proportionality constant).

WLS: The Catch

WLS requires knowing σ_i^2 (or at least σ_i^2 up to a proportionality constant).

Common assumptions:

- $\text{Var}(e_i) = \sigma^2 x_i \implies w_i = 1/x_i$
- $\text{Var}(e_i) = \sigma^2 x_i^2 \implies w_i = 1/x_i^2$
- $\text{Var}(e_i) = \sigma^2/n_i$ (grouped data) $\implies w_i = n_i$

WLS: The Catch

WLS requires knowing σ_i^2 (or at least σ_i^2 up to a proportionality constant).

Common assumptions:

- $\text{Var}(e_i) = \sigma^2 x_i \implies w_i = 1/x_i$
- $\text{Var}(e_i) = \sigma^2 x_i^2 \implies w_i = 1/x_i^2$
- $\text{Var}(e_i) = \sigma^2/n_i$ (grouped data) $\implies w_i = n_i$

R note: The `weights` argument in `lm()` expects $w_i = 1/\sigma_i^2$, not σ_i^2:

```
# If Var(e_i) proportional to income: wls_model <- lm(food   income, data = food_data, weights = 1 / incor
```

WLS: The Catch

WLS requires knowing σ_i^2 (or at least σ_i^2 up to a proportionality constant).

Common assumptions:

- $\text{Var}(e_i) = \sigma^2 x_i \implies w_i = 1/x_i$
- $\text{Var}(e_i) = \sigma^2 x_i^2 \implies w_i = 1/x_i^2$
- $\text{Var}(e_i) = \sigma^2/n_i$ (grouped data) $\implies w_i = n_i$

R note: The `weights` argument in `lm()` expects $w_i = 1/\sigma_i^2$, not σ_i^2:

```
# If Var(e_i) proportional to income: wls_model <- lm(food   income, data = food_data, weights = 1 / incor
```

$\implies$ If we know the form of heteroskedasticity, WLS gives us efficient estimates. But what if we don't know σ_i^2?

Feasible GLS: Estimate the Variance Function

When σ_i^2 is unknown, we **estimate** it from the data in two steps.

Feasible GLS: Estimate the Variance Function

When σ_i^2 is unknown, we **estimate** it from the data in two steps.

Step 1: Run OLS and get residuals $\hat{e}_i$. Then regress:

$$\ln(\hat{e}_i^2) = \gamma_0 + \gamma_1 \ln(x_i) + v_i$$

This estimates how variance scales with x.

Feasible GLS: Estimate the Variance Function

When σ_i^2 is unknown, we **estimate** it from the data in two steps.

Step 1: Run OLS and get residuals $\hat{e}_i$. Then regress:

$$\ln(\hat{e}_i^2) = \gamma_0 + \gamma_1 \ln(x_i) + v_i$$

This estimates how variance scales with x.

Step 2: Compute estimated weights:

$$\hat{\sigma}_i^2 = \exp(\hat{\gamma}_0 + \hat{\gamma}_1 \ln x_i), \qquad \hat{w}_i = 1/\hat{\sigma}_i^2$$

Run WLS using these estimated weights.

Feasible GLS: Estimate the Variance Function

When σ_i^2 is unknown, we **estimate** it from the data in two steps.

Step 1: Run OLS and get residuals $\hat{e}_i$. Then regress:

$$\ln(\hat{e}_i^2) = \gamma_0 + \gamma_1 \ln(x_i) + v_i$$

This estimates how variance scales with x.

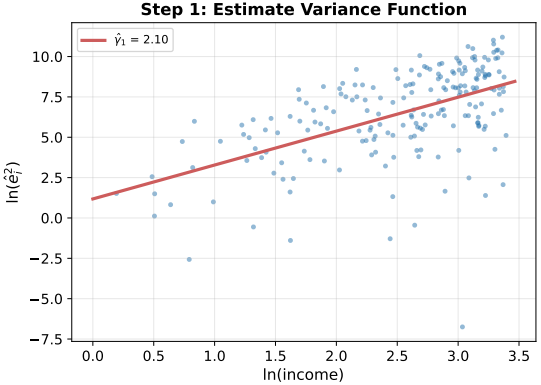
Step 2: Compute estimated weights:

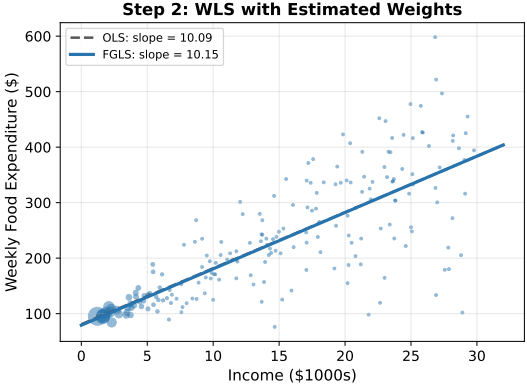
$$\hat{\sigma}_i^2 = \exp(\hat{\gamma}_0 + \hat{\gamma}_1 \ln x_i), \qquad \hat{w}_i = 1/\hat{\sigma}_i^2$$

Run WLS using these estimated weights.

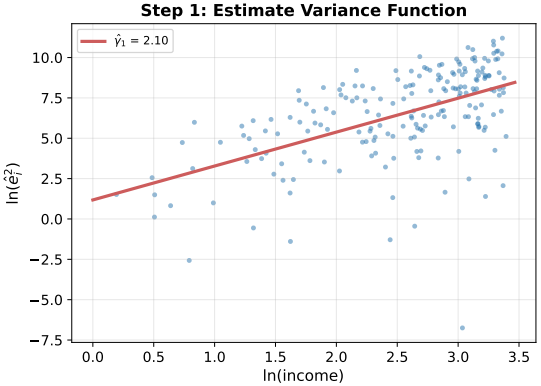
$\Longrightarrow$ FGLS is "feasible" because it replaces the unknown σ_i^2 with an estimate $\hat{\sigma}_i^2$. With large n, FGLS approaches the efficiency of true GLS.

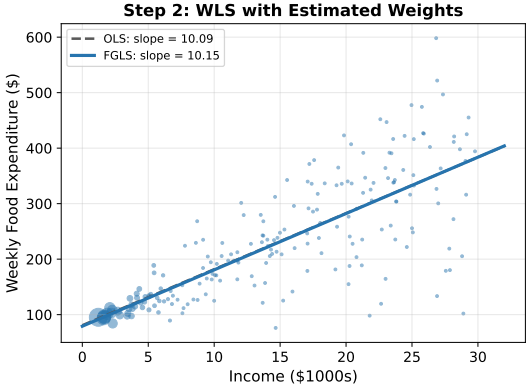
FGLS: The Two Steps





FGLS: The Two Steps





Left: the auxiliary regression reveals how variance scales with income. Right: FGLS uses these estimated weights to refit the model.

Exercise 8.1: Setup

Model: $\text{food}_i = \beta_1 + \beta_2 \, \text{income}_i + e_i$

Data: 40 households. Weekly food expenditure ($) and weekly income ($100s).

Exercise 8.1: Setup

Model: $\text{food}_i = \beta_1 + \beta_2 \, \text{income}_i + e_i$

Data: 40 households. Weekly food expenditure ($) and weekly income ($100s).

OLS results:

$$\widehat{\text{food}} = 83.42 + 10.21 \, \text{income}$$

	$\hat{\beta}_1$ (intercept)	$\hat{\beta}_2$ (income)
Estimate	83.42	10.21
OLS SE	43.41	2.09

Exercise 8.1: Setup

Model: $\text{food}_i = \beta_1 + \beta_2 \, \text{income}_i + e_i$

Data: 40 households. Weekly food expenditure ($) and weekly income ($100s).

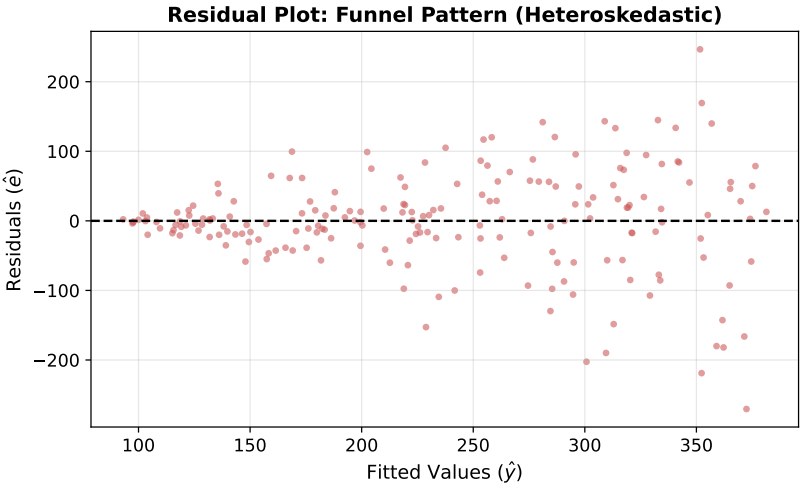
OLS results:

$$\widehat{\text{food}} = 83.42 + 10.21 \, \text{income}$$

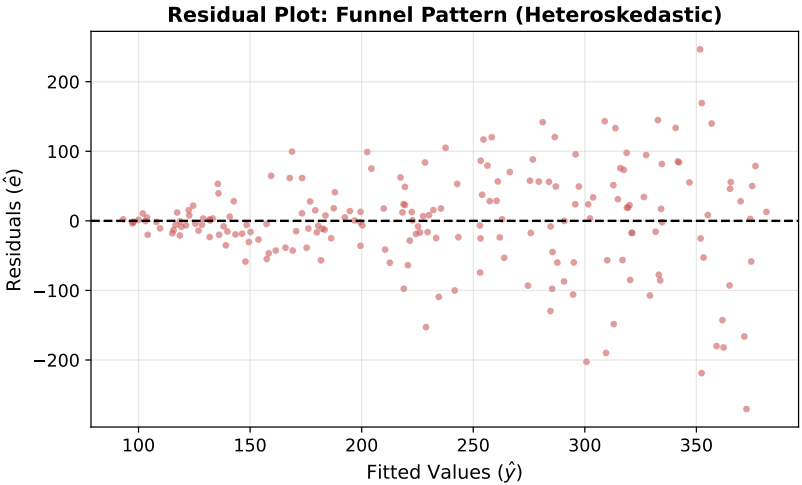
	$\hat{\beta}_1$ (intercept)	$\hat{\beta}_2$ (income)
Estimate	83.42	10.21
OLS SE	43.41	2.09

$t = 10.21/2.09 = 4.88 \implies$ significant at any conventional level. But can we trust this SE?

Residual Plot



Residual Plot



Clear funnel pattern $\Longrightarrow$ heteroskedasticity. The OLS SE of 2.09 is suspect.

Goldfeld-Quandt Test

Steps for the food expenditure data:

1. Sort 40 observations by income

Goldfeld-Quandt Test

Steps for the food expenditure data:

1. Sort 40 observations by income
2. Drop middle 8 observations ($c = 40/5 = 8$)

Goldfeld-Quandt Test

Steps for the food expenditure data:

1. Sort 40 observations by income
2. Drop middle 8 observations ($c = 40/5 = 8$)
3. Bottom 16: low-income households $\implies$ SSE_1
4. Top 16: high-income households $\implies$ SSE_2

Goldfeld-Quandt Test

Steps for the food expenditure data:

1. Sort 40 observations by income
2. Drop middle 8 observations ($c = 40/5 = 8$)
3. Bottom 16: low-income households $\implies$ SSE_1
4. Top 16: high-income households $\implies$ SSE_2
5. $\text{GQ} = \dfrac{\text{SSE}_2/(16-2)}{\text{SSE}_1/(16-2)} = \dfrac{\text{SSE}_2}{\text{SSE}_1}$

Goldfeld-Quandt Test

Steps for the food expenditure data:

1. Sort 40 observations by income
2. Drop middle 8 observations ($c = 40/5 = 8$)
3. Bottom 16: low-income households $\implies$ SSE_1
4. Top 16: high-income households $\implies$ SSE_2
5. $\text{GQ} = \dfrac{\text{SSE}_2/(16-2)}{\text{SSE}_1/(16-2)} = \dfrac{\text{SSE}_2}{\text{SSE}_1}$

Under H_0: $\text{GQ} \sim F(14, 14)$

If $\text{GQ} > F_{0.05}(14, 14) = 2.48$ $\implies$ reject homoskedasticity.

Goldfeld-Quandt Test

Steps for the food expenditure data:

1. Sort 40 observations by income
2. Drop middle 8 observations ($c = 40/5 = 8$)
3. Bottom 16: low-income households $\implies$ SSE_1
4. Top 16: high-income households $\implies$ SSE_2
5. $\text{GQ} = \dfrac{\text{SSE}_2/(16-2)}{\text{SSE}_1/(16-2)} = \dfrac{\text{SSE}_2}{\text{SSE}_1}$

Under H_0: $\text{GQ} \sim F(14, 14)$

If $\text{GQ} > F_{0.05}(14, 14) = 2.48$ $\implies$ reject homoskedasticity.

$\implies$ The GQ test is intuitive: does the high-income group have a bigger residual sum of squares than the low-income group?

Robust SEs for Food Expenditure

Compare OLS vs. robust standard errors for $\hat{\beta}_2$:

Robust SEs for Food Expenditure

Compare OLS vs. robust standard errors for $\hat{\beta}_2$:

	SE($\hat{\beta}_2$)	95% CI for β_2	CI Width
OLS	2.09	$(5.97, 14.45)$	8.48
Robust (HC1)	3.19	$(3.71, 16.71)$	13.00

Robust SEs for Food Expenditure

Compare OLS vs. robust standard errors for $\hat{\beta}_2$:

	SE($\hat{\beta}_2$)	95% CI for β_2	CI Width
OLS	2.09	(5.97, 14.45)	8.48
Robust (HC1)	3.19	(3.71, 16.71)	13.00

- The robust SE is **52% larger** than the OLS SE
- The robust CI is wider $\implies$ the OLS CI was falsely precise
- Both reject H_0: $\beta_2 = 0$, but the evidence is weaker with correct SEs

Robust SEs for Food Expenditure

Compare OLS vs. robust standard errors for $\hat{\beta}_2$:

	SE($\hat{\beta}_2$)	95% CI for β_2	CI Width
OLS	2.09	(5.97, 14.45)	8.48
Robust (HC1)	3.19	(3.71, 16.71)	13.00

- The robust SE is **52% larger** than the OLS SE
- The robust CI is wider $\implies$ the OLS CI was falsely precise
- Both reject H_0: $\beta_2 = 0$, but the evidence is weaker with correct SEs

$\implies$ In this case, OLS SEs understate uncertainty. Robust SEs give honest inference.

Computing Robust CIs

95% CI formula (same as always, just swap in the robust SE):

$$\hat{\beta}_2 \pm t_{0.025,\, n-2} \times \mathrm{SE}_{\mathrm{robust}}(\hat{\beta}_2)$$

Computing Robust CIs

95% CI formula (same as always, just swap in the robust SE):

$$\hat{\beta}_2 \pm t_{0.025,\, n-2} \times \text{SE}_{\text{robust}}(\hat{\beta}_2)$$

Example:

$$10.21 \pm 2.024 \times 3.19 = 10.21 \pm 6.46 = (3.75, 16.67)$$

Computing Robust CIs

95% CI formula (same as always, just swap in the robust SE):

$$\hat{\beta}_2 \pm t_{0.025,\, n-2} \times \text{SE}_{\text{robust}}(\hat{\beta}_2)$$

Example:

$$10.21 \pm 2.024 \times 3.19 = 10.21 \pm 6.46 = (3.75, 16.67)$$

In R:

```
library(car); library(lmtest) model <- lm(food   income, data = food_data) # Robust t-test and CI coeftes
vcov. = hccm(model, type = "hc1")) confint(coeftest(model, vcov. = hccm(model, type = "hc1")))
```

Summary: Which Fix to Use?

Method	**When to Use**	**Limitation**
Robust SEs	Always safe; default in applied work	Does not improve efficiency
WLS	You know $\text{Var}(e_i) \propto h(x_i)$	Requires correct variance form
FGLS	You can estimate the variance function	Two-step; relies on auxiliary regression

Summary: Which Fix to Use?

Method	When to Use	Limitation
Robust SEs	Always safe; default in applied work	Does not improve efficiency
WLS	You know $\text{Var}(e_i) \propto h(x_i)$	Requires correct variance form
FGLS	You can estimate the variance function	Two-step; relies on auxiliary regression

Practical advice:

1. Always check residual plots after running OLS
2. Report robust SEs by default
3. Use WLS/FGLS only when you have a good model for the variance structure

Summary: Which Fix to Use?

Method	When to Use	Limitation
Robust SEs	Always safe; default in applied work	Does not improve efficiency
WLS	You know $\text{Var}(e_i) \propto h(x_i)$	Requires correct variance form
FGLS	You can estimate the variance function	Two-step; relies on auxiliary regression

Practical advice:

1. Always check residual plots after running OLS
2. Report robust SEs by default
3. Use WLS/FGLS only when you have a good model for the variance structure

$\Longrightarrow$ Robust SEs are the minimum standard. WLS/FGLS can do better if you model the variance correctly.

Thank you!
jakeanderson@g.ucla.edu