Week 1 - Chapter 8 Discussion Problems

Jake Anderson

ECON 104 – UCLA

Outline

Problem 8.5 (a)

Consider the simple regression model

$$y_i = \beta_1 + \beta_2 x_{i2} + e_i.$$

Suppose $N = 5$ and the values of x_{i2} are $(1, 2, 3, 4, 5)$. Let the true values of the parameters be $\beta_1 = 1$, $\beta_2 = 1$. Let the true random error values, which are never known in reality, be

$$e_i = (1, -1, 0, 6, -6).$$

a. Calculate the values of y_i.

Problem 8.5 (a)

Consider the simple regression model

$$y_i = \beta_1 + \beta_2 x_{i2} + e_i.$$

Suppose $N = 5$ and the values of x_{i2} are $(1, 2, 3, 4, 5)$. Let the true values of the parameters be $\beta_1 = 1$, $\beta_2 = 1$. Let the true random error values, which are never known in reality, be

$$e_i = (1, -1, 0, 6, -6).$$

(a) Calculate the values of y_i.

Solution:
They give us the values for everything we need in the equation, so we can plug in for each:

$$y_1 = 1 + 1 \cdot 1 + 1 = 3$$
$$y_2 = 1 + 1 \cdot 2 - 1 = 2$$
$$y_3 = 1 + 1 \cdot 3 + 0 = 4$$
$$y_4 = 1 + 1 \cdot 4 + 6 = 11$$
$$y_5 = 1 + 1 \cdot 5 - 6 = 0$$

Problem 8.5 (b)

(b) The OLS estimates ($\hat{\beta}_1$ and $\hat{\beta}_2$) of the parameters are $b_1 = 3.1$ and $b_2 = 0.3$. Compute the least squares residual, $\hat{e}_1$, for the first observation, and $\hat{e}_4$, for the fourth observation. What is the sum of all the least squares residuals? In this example, what is the sum of the true random errors? Is the sum of the residuals always equal to the sum of the random errors? Explain.

Solution:

Model used for residuals:

$$\hat{e}_i = y_i - \hat{b}_1 - \hat{b}_2 x_{i2}.$$

From part (a):

$$y_i = \beta_1 + \beta_2 x_{i2} + e_i = 1 + x_{i2} + e_i, \qquad y_1 = 3, \quad y_4 = 11$$

Now plug into the fitted model ($\hat{y}_i = 3.1 + 0.3 x_{i2}$):

$$\hat{e}_1 = y_1 - \hat{y}_1 = 3 - (3.1 + 0.3 \cdot 1) = 3 - 3.4 = -0.4$$

$$\hat{e}_4 = y_4 - \hat{y}_4 = 11 - (3.1 + 0.3 \cdot 4) = 11 - 4.3 = 6.7$$

Problem 8.5 (c)

- It is hypothesized that the data are heteroskedastic with the variance of the first three random errors being σ_1^2, and the variance of the last two random errors being σ_2^2. We regress the squared residuals $\hat{e}_i^2$ on the indicator variable z_i, where $z_i = 0$, $i = 1, 2, 3$ and $z_i = 1$, $i = 4, 5$. The overall model F-statistic value is 12.86. Does this value provide evidence of heteroskedasticity at the 5% level of significance? What is the p-value for this F-value (requires computer)?

Problem 8.5 (c)

- It is hypothesized that the data are heteroskedastic with the variance of the first three random errors being σ_1^2, and the variance of the last two random errors being σ_2^2. We regress the squared residuals $\hat{e}_i^2$ on the indicator variable z_i, where $z_i = 0$, $i = 1, 2, 3$ and $z_i = 1$, $i = 4, 5$. The overall model F-statistic value is 12.86. Does this value provide evidence of heteroskedasticity at the 5% level of significance? What is the p-value for this F-value (requires computer)?

Solution:

The hypotheses of homoskedasticity: $H_0: \ \sigma_1^2 = \sigma_2^2$, $H_1: \ \sigma_1^2 \neq \sigma_2^2$.

The auxiliary regression is

$$\hat{e}_i^2 = \gamma_0 + \gamma_1 z_i + u_i,$$

so testing for heteroskedasticity is equivalent to testing $H_0 : \gamma_1 = 0$.

With $N = 5$ observations and two estimated parameters, the F-test has degrees of freedom $(1, 5 - 2 = 3)$. The test statistic is $F = 12.86$, which yields

$$p = \Pr(F_{1,3} \geq 12.86) \approx 0.037.$$

Since $p < 0.05$, (and $12.86 > F_{crit} = 10.13$) we reject the null hypothesis and conclude that there is evidence of heteroskedasticity at the 5% significance level.

Problem 8.5 (d)

- $R^2 = 0.8108$ from the regression in (c). Use this value to carry out the LM (Breusch–Pagan) test for heteroskedasticity at the 5% level of significance. What is the p-value for this test (requires computer)?

Solution

For the Lagrange Multiplier (Breusch–Pagan) test, the statistic is

$$LM = NR^2.$$

Here $N = 5$ and $R^2 = 0.8108$, so

$$LM = 5(0.8108) = 4.054.$$

Under H_0 (homoskedasticity), $LM \sim \chi^2_q$ where q is the number of regressors in the auxiliary regression excluding the intercept. Since the auxiliary regression uses only z_i, we have $q = 1$. Thus,

$$p = \Pr(\chi^2_1 \geq 4.054) \approx 0.0441.$$

Because $p < 0.05$, (and $4.054 > \chi^2_{crit} = 3.841$) we reject H_0 at the 5% level and conclude there is evidence of heteroskedasticity.

Problem 8.5 (e)

(e) We now regress $\ln(\hat{e}_i^2)$ on z_i. The estimated coefficient of z_i is 3.81. We discover that the software reports using only $N = 4$ observations in this calculation. Why?

Solution

The regression uses $\ln(\hat{e}_i^2)$ as the dependent variable. If any least-squares residual is exactly zero, then $\hat{e}_i^2 = 0$ and

$$\ln(\hat{e}_i^2) = \ln(0)$$

is undefined $(-\infty)$. The software therefore drops that observation from the regression and the regression is carried out using only $N = 4$ observations.

Moral of the story: Software does things to "help" you, but be careful!
Think Pair Share: Using your 103 skills, how do I interpret the coefficient on z_i in the variance regression?

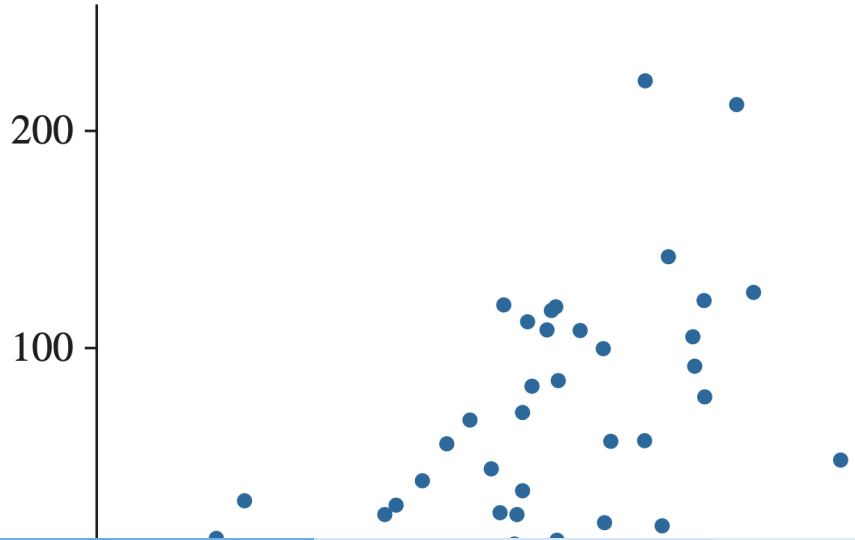
Problem 8.5 (f): Feasible GLS

(f) In order to carry out feasible generalized least squares using information from the regression in part (e), we first create the transformed variables $(y_i^*, x_{i1}^*, x_{i2}^*)$. List the values of the transformed observations for $i = 1$ and $i = 4$.

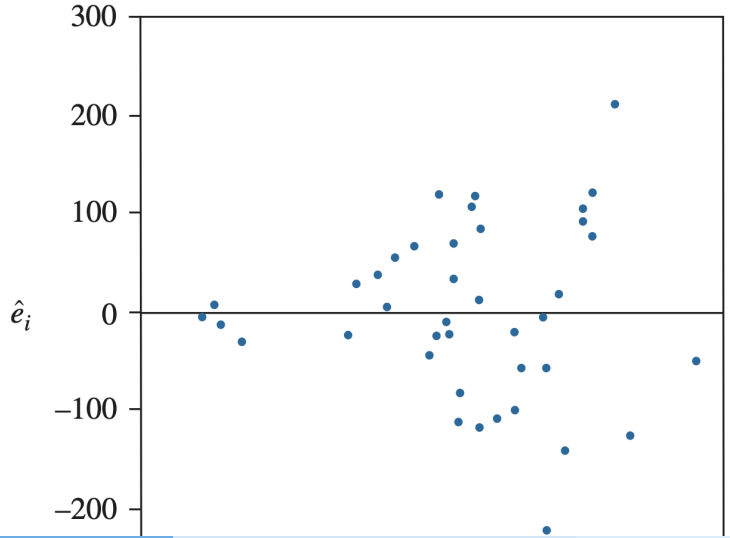
Solution Picture Prerequisites:

1. Untransformed Data
2. Transformed Data
3. Remember the fundamental equation of FGLS

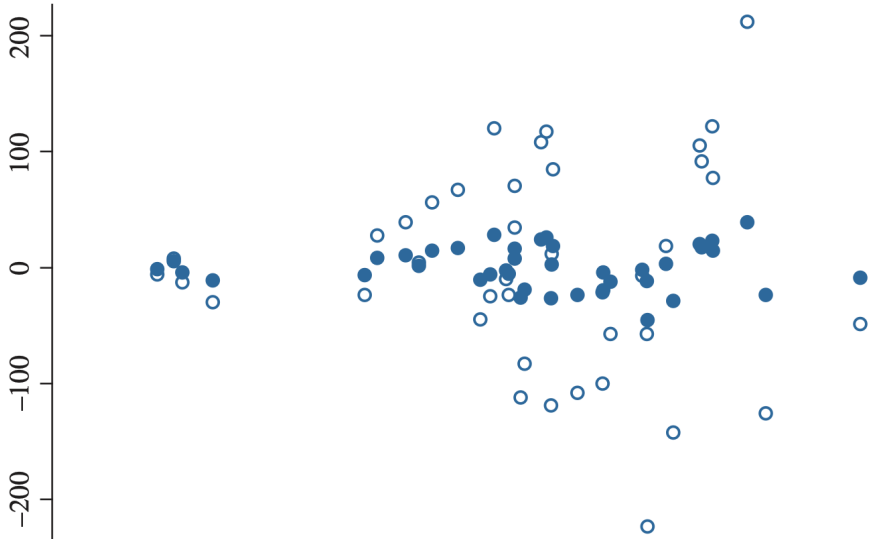
Untransformed Data



Same Data, but looking at the residuals instead of Y



Comparing residuals to scaled residuals



Remember the fundamental equation of FGLS

$$\text{Var}(e_i^* \mid x_i) = \text{Var}\left(\frac{e_i}{\sqrt{x_i}} \,\middle|\, x_i \right) = \frac{1}{x_i} \text{Var}(e_i \mid x_i) = \frac{1}{x_i} \sigma^2 x_i = \sigma^2$$

- Here, the skedastic function is $h(x_i) = x_i$, so we divide by the square root of x_i to transform the data.
- This comes from an assumption we make of the form of the variance function.

Problem 8.5 (f): Feasible GLS

(f) In order to carry out feasible generalized least squares using information from the regression in part (e), we first create the transformed variables $(y_i^*, x_{i1}^*, x_{i2}^*)$. List the values of the transformed observations for $i = 1$ and $i = 4$.

Solution (following section 8.5.1 in HGL):

1. Estimate the original model by OLS and save residuals $\hat{e}_i$.
2. Use $\hat{e}_i$ and z_i to estimate a variance model.
3. Compute the estimated skedastic function $\hat{h}(z_i)$.
4. Divide each observation by $\sqrt{\hat{h}(z_i)}$ to form $(y_i^*, x_{i1}^*, x_{i2}^*)$.

Step (1): OLS residuals

Original model:

$$y_i = \beta_1 + \beta_2 x_{i2} + e_i, \qquad x_{i1} \equiv 1.$$

OLS fit (given in the problem):

$$\hat{y}_i = b_1 + b_2 x_{i2} = 3.1 + 0.3\, x_{i2}.$$

OLS residuals: (we computed these in part (b))

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (3.1 + 0.3\, x_{i2}).$$

These $\hat{e}_i$ are the inputs for the variance model in Step (2).

Step (2): estimate a variance model (part (e))

From the variance regression in part (e),

$$\ln(\hat{e}_i^2) = \alpha + 3.81\, z_i$$

the fitted variance model is

$$\hat{\sigma}_i^2 = e^{\alpha + 3.81\, z_i}$$

Rewrite this as

$$\hat{\sigma}_i^2 = \underbrace{e^{\alpha}}_{\text{common (baseline) variance}} \times \underbrace{e^{3.81\, z_i}}_{\text{group-specific multiplier}}$$

So the estimated variance differs by group through the factor $e^{3.81\, z_i}$.

Step (3): compute the skedastic function $\hat{h}(z_i)$

Write the estimated variance as

$$\hat{\sigma}_i^2 = \hat{\sigma}^2 \, \hat{h}_i, \qquad \hat{\sigma}^2 \equiv e^{\alpha} \qquad \hat{h}_i \equiv e^{3.81 \, z_i}.$$

Since $z_i \in \{0, 1\}$:

$$\hat{h}_i = \begin{cases} 1, & z_i = 0 \ (i = 1, 2, 3), \\ 45.150, & z_i = 1 \ (i = 4, 5). \end{cases}$$

(For simplicity, we drop the hat and write h_i in subsequent steps.)

Step (4): transform the data and report $i = 1$ and $i = 4$

FGLS divides each observation by the estimated standard deviation: $\hat{\sigma}_i = \hat{\sigma}\sqrt{h_i}$. Dividing all observations by the common constant $\hat{\sigma}$ does not change OLS estimates, so we implement the transform using $\sqrt{h_i}$:

$$y_i^* = \frac{y_i}{\sqrt{h_i}}, \qquad x_{i1}^* = \frac{1}{\sqrt{h_i}}, \qquad x_{i2}^* = \frac{x_{i2}}{\sqrt{h_i}}.$$

Compute $\sqrt{h_i}$:

$$\sqrt{h_i} = \begin{cases} 1, & z_i = 0, \\ \sqrt{45.150} = 6.719, & z_i = 1. \end{cases} \qquad \frac{1}{6.719} = 0.149.$$

i	y_i	x_{i1}	x_{i2}	y_i^*	x_{i1}^*	x_{i2}^*
1	3	1	1	3	1	1
2	2	1	2	2	1	2
3	4	1	3	4	1	3
4	11	1	4	1.637	0.149	0.595
5	0	1	5	0	0.149	0.744

Gut check about the FGLS weights

From the log-variance regression,

$$\ln(\sigma_i^2 \mid z_i = 1) - \ln(\sigma_i^2 \mid z_i = 0) = 3.81.$$

This implies

$$\frac{\sigma^2_{z=1}}{\sigma^2_{z=0}} = e^{3.81} \approx 45, \qquad \frac{\sigma_{z=1}}{\sigma_{z=0}} \approx \sqrt{45} \approx 6.7.$$

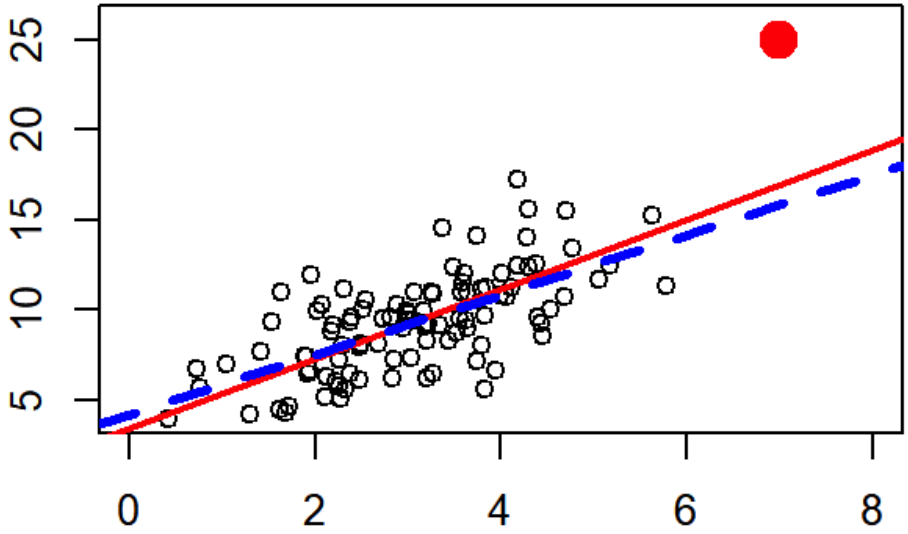
FGLS divides each observation by its estimated standard deviation. Therefore,

$$w_{z=1} = \frac{1}{\sigma_{z=1}} \approx \frac{1}{6.7} \approx 0.15.$$

Interpretation:

- Observations in the high-variance group are much noisier.
- FGLS downweights them so they influence the fit less.
- One $z = 1$ observation counts like about 15% of a $z = 0$ observation.
- This give us more stable estimates and smaller standard errors compared to OLS.

Note on Influential Observations



Outline

Question 8.21: Store-level averaging in the LPM

In Example 8.9 we estimated the linear probability model

$$COKE = \beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI + e,$$

where $COKE = 1$ if a shopper purchased Coke and $COKE = 0$ if a shopper purchased Pepsi. The variable $PRATIO$ is the relative price ratio of Coke to Pepsi, and $DISP\_COKE$ and $DISP\_PEPSI$ are indicator variables equal to one if the relevant display is present.

Suppose we have 1140 observations on randomly selected shoppers from 50 different grocery stores. Each grocery store has its own settings for $PRATIO$, $DISP\_COKE$, and $DISP\_PEPSI$. Let (i, j) denote the jth shopper at the ith store. Then the model can be written as

$$COKE_{ij} = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i + e_{ij}.$$

Question 8.21: Store-level averaging in the LPM

In Example 8.9 we estimated the linear probability model

$$COKE = \beta_1 + \beta_2 PRATIO + \beta_3 DISP\_COKE + \beta_4 DISP\_PEPSI + e,$$

where $COKE = 1$ if a shopper purchased Coke and $COKE = 0$ if a shopper purchased Pepsi. The variable $PRATIO$ is the relative price ratio of Coke to Pepsi, and $DISP\_COKE$ and $DISP\_PEPSI$ are indicator variables equal to one if the relevant display is present.

Suppose we have 1140 observations on randomly selected shoppers from 50 different grocery stores. Each grocery store has its own settings for $PRATIO$, $DISP\_COKE$, and $DISP\_PEPSI$. Let (i, j) denote the jth shopper at the ith store. Then the model can be written as

$$COKE_{ij} = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i + e_{ij}.$$

Averaging this equation over all shoppers in store i gives

$$\overline{COKE}_i = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i + \overline{e}_i,$$

where

$$\overline{e}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} e_{ij}, \qquad \overline{COKE}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} COKE_{ij},$$

and N_i is the number of sampled shoppers in the ith store.

Question 8.21: Do You Believe It?

First things first: **Does this model make sense to you?**

We start from the individual-level model (shopper j in store i):

$$COKE_{ij} = \beta_1 + \beta_2\, PRATIO_i + \beta_3\, DISP\_COKE_i + \beta_4\, DISP\_PEPSI_i + e_{ij}.$$

1. $COKE_{ij} = 1$ if shopper j buys Coke, 0 if Pepsi.
2. Store-level regressors $\left(PRATIO_i, DISP\_COKE_i, DISP\_PEPSI_i\right)$ are constant within store i.

Part (a): interpretation of $\overline{COKE}_i$

(a) What is the interpretation of $\overline{COKE}_i$ for store i?

Answer. Because $COKE_{ij} \in \{0, 1\}$,

$$\overline{COKE}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} COKE_{ij}$$

is the *sample share (fraction)* of sampled shoppers in store i who purchased Coke. Equivalently, it is the *sample mean* of a 0 / 1 indicator in store i.

Part (b): assumptions

(b) Assume:

i. $E(COKE_{ij} \mid x_{ij}) = P_i$
ii. $\text{Var}(COKE_{ij} \mid x_{ij}) = P_i(1 - P_i)$

Show:

i. $E(\overline{COKE}_i \mid X) = P_i$
ii. $\text{Var}(\overline{COKE}_i \mid X) = P_i(1 - P_i)/N_i$

Part (b): assumptions

(b) Assume:

i. $E(COKE_{ij} \mid x_{ij}) = P_i$
ii. $\text{Var}(COKE_{ij} \mid x_{ij}) = P_i(1 - P_i)$

Show:

i. $E(\overline{COKE}_i \mid X) = P_i$
ii. $\text{Var}(\overline{COKE}_i \mid X) = P_i(1 - P_i)/N_i$

First:

- What kind of random variable is this?

Part (b): assumptions

(b) Assume:

i. $E(COKE_{ij} \mid x_{ij}) = P_i$
ii. $\text{Var}(COKE_{ij} \mid x_{ij}) = P_i(1 - P_i)$

Show:

i. $E(\overline{COKE}_i \mid X) = P_i$
ii. $\text{Var}(\overline{COKE}_i \mid X) = P_i(1 - P_i)/N_i$

First:

- What kind of random variable is this?

Answer: Bernoulli!

Part (b): compute $E(\overline{COKE}_i \mid X)$

Start from the definition:

$$\overline{COKE}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} COKE_{ij}.$$

Use linearity of expectation:

$$\begin{aligned} E(\overline{COKE}_i \mid X) &= \frac{1}{N_i} \sum_{j=1}^{N_i} E(COKE_{ij} \mid X) \\ &= \frac{1}{N_i} \sum_{j=1}^{N_i} P_i \\ &= \frac{1}{N_i} P_i \sum_{j=1}^{N_i} 1 \\ &= \frac{1}{N_i} N_i P_i \\ &= P_i. \end{aligned}$$

Part (b): compute $\text{Var}(\overline{COKE}_i \mid X)$

Use the variance of an average:

$$\text{Var}(\overline{COKE}_i \mid X) = \text{Var}\left(\frac{1}{N_i} \sum_{j=1}^{N_i} COKE_{ij} \;\middle|\; X \right) = \frac{1}{N_i^2} \text{Var}\left(\sum_{j=1}^{N_i} COKE_{ij} \;\middle|\; X \right).$$

Expand the variance:

$$\text{Var}\left(\sum_{j=1}^{N_i} COKE_{ij} \mid X \right) = \sum_{j=1}^{N_i} \text{Var}(COKE_{ij} \mid X) + 2 \sum_{j<k} \text{Cov}(COKE_{ij}, COKE_{ik} \mid X).$$

Under zero covariances, the covariance terms drop:

$$\begin{aligned} \text{Var}(\overline{COKE}_i \mid X) &= \frac{1}{N_i^2} \sum_{j=1}^{N_i} \text{Var}(COKE_{ij} \mid X) \\ &= \frac{1}{N_i^2} \sum_{j=1}^{N_i} P_i(1 - P_i) \\ &= \frac{1}{N_i^2} N_i P_i(1 - P_i) \end{aligned}$$

Takeaway: what averaging does to variance

- $\overline{COKE}_i$ is a mean of N_i Bernoulli draws with success probability P_i.
- Mean stays the same: $E(\overline{COKE}_i \mid X) = P_i$.
- Variance shrinks with sample size:

$$\text{Var}(\overline{COKE}_i \mid X) = \frac{P_i(1 - P_i)}{N_i}.$$

Takeaway: what averaging does to variance

- $\overline{COKE}_i$ is a mean of N_i Bernoulli draws with success probability P_i.
- Mean stays the same: $E(\overline{COKE}_i \mid X) = P_i$.
- Variance shrinks with sample size:

$$\text{Var}(\overline{COKE}_i \mid X) = \frac{P_i(1 - P_i)}{N_i}.$$

- Larger $N_i \Rightarrow$ more precise store-level share.

Outline

Question 8.27: Olympics Medal Analysis

There were 64 countries who competed in the 1992 Olympics and won at least one medal.

For each of these countries:

- **MEDALTOT**: total number of medals won
- **POP**: population in millions
- **GDP**: GDP in billions of 1995 dollars

We will exclude the United Kingdom and use the remaining $N = 63$ observations.

Part (a): OLS Estimation

- Estimate the model:

$$MEDALTOT = \beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + e$$

by OLS.

Part (a): OLS Estimation

- Estimate the model:

$$MEDALTOT = \beta_1 + \beta_2 \ln(POP) + \beta_3 \ln(GDP) + e$$

by OLS.

R Output:

```
##
## Call:
## lm(formula = medaltot ~ log(pop) + log(gdp), data = olympics)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.127 -12.790  -4.657   6.383  85.193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  113.535     21.395   5.307  1.7e-06 ***
## log(pop)       2.764      2.070   1.335   0.1868
## log(gdp)       4.270      1.718   2.486   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part (b): Testing for Heteroskedasticity

- Calculate $\hat{e}_i^2$ from the regression in (a). Regress $\hat{e}_i^2$ on $\ln(POP)$ and $\ln(GDP)$.

Part (b): Testing for Heteroskedasticity

(b) Calculate $\hat{e}_i^2$ from the regression in (a). Regress $\hat{e}_i^2$ on $\ln(POP)$ and $\ln(GDP)$.

Auxiliary Regression Output:

```
##
## Call:
## lm(formula = resid ~ log(pop) + log(gdp), data = olympics)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -906.9 -333.2 -239.5    9.4 6384.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3691.84    1139.75   3.239  0.00196 **
## log(pop)      133.60     110.27   1.212  0.23044
## log(gdp)      112.38      91.51   1.228  0.22422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1042 on 60 degrees of freedom
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.1031
```

Part (c): Robust Standard Errors

- Reestimate the model using heteroskedasticity robust standard errors.

Part (c): Robust Standard Errors

- Reestimate the model using heteroskedasticity robust standard errors.

Robust SE Output:

```
t test of coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.5354    33.8438  3.3547 0.001382 **
log(pop)      2.7643     1.7891  1.5451 0.127580
log(gdp)      4.2705     1.7797  2.3995 0.019540 *
```

Part (c): Robust Standard Errors

(e) Reestimate the model using heteroskedasticity robust standard errors.

Robust SE Output:

```
t test of coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.5354    33.8438  3.3547 0.001382 **
log(pop)      2.7643     1.7891  1.5451 0.127580
log(gdp)      4.2705     1.7797  2.3995 0.019540 *
```

Key Change: Standard errors adjusted due to heteroskedasticity correction

- ln(GDP): SE increased from 1.72 to 1.78
- ln(POP): SE decreased from 2.07 to 1.79

Part (c): Testing GDP Relationship

Question: Is there a relationship between medals and GDP at 10% and 5% significance?

Part (c): Testing GDP Relationship

Question: Is there a relationship between medals and GDP at 10% and 5% significance?

Step 1: Set up hypotheses

- H_0: $\beta_3 = 0$ (no relationship)
- H_1: $\beta_3 > 0$ (positive relationship)
- One-sided test (we expect positive effect)

Part (c): Testing GDP Relationship

Question: Is there a relationship between medals and GDP at 10% and 5% significance?

Step 1: Set up hypotheses

- H_0: $\beta_3 = 0$ (no relationship)
- H_1: $\beta_3 > 0$ (positive relationship)
- One-sided test (we expect positive effect)

Step 2: Calculate test statistic

$$t = \frac{\hat{\beta}_3}{\text{Robust SE}(\hat{\beta}_3)} = \frac{4.2705}{1.7797} = 2.399$$

Part (c): Testing GDP Relationship

Question: Is there a relationship between medals and GDP at 10% and 5% significance?

Step 1: Set up hypotheses

- H_0: $\beta_3 = 0$ (no relationship)
- H_1: $\beta_3 > 0$ (positive relationship)
- One-sided test (we expect positive effect)

Step 2: Calculate test statistic

$$t = \frac{\hat{\beta}_3}{\text{Robust SE}(\hat{\beta}_3)} = \frac{4.2705}{1.7797} = 2.399$$

Step 3: Find critical values

- $t_{0.10,60} = 1.296$ (10% significance, one-sided)
- $t_{0.05,60} = 1.671$ (5% significance, one-sided)

Part (d): Testing Population Relationship

Question: Is there a relationship between medals and population at 10% and 5% significance?

Part (d): Testing Population Relationship

Question: Is there a relationship between medals and population at 10% and 5% significance?

Step 1: Set up hypotheses

- H_0: $\beta_2 = 0$ (no relationship)
- H_1: $\beta_2 > 0$ (positive relationship)

Part (d): Testing Population Relationship

Question: Is there a relationship between medals and population at 10% and 5% significance?

Step 1: Set up hypotheses

- H_0: $\beta_2 = 0$ (no relationship)
- H_1: $\beta_2 > 0$ (positive relationship)

Step 2: Calculate test statistic

$$t = \frac{\hat{\beta}_2}{\text{Robust SE}(\hat{\beta}_2)} = \frac{2.7643}{1.7891} = 1.545$$

Part (d): Testing Population Relationship

Question: Is there a relationship between medals and population at 10% and 5% significance?

Step 1: Set up hypotheses

- H_0: $\beta_2 = 0$ (no relationship)
- H_1: $\beta_2 > 0$ (positive relationship)

Step 2: Calculate test statistic

$$t = \frac{\hat{\beta}_2}{\text{Robust SE}(\hat{\beta}_2)} = \frac{2.7643}{1.7891} = 1.545$$

Step 3: Compare to critical values

- $t_{0.10,60} = 1.296 \rightarrow 1.545 > 1.296$ ✓
- $t_{0.05,60} = 1.671 \rightarrow 1.545 < 1.671$ X
- **Weak evidence** of positive relationship (significant at 10% but not 5%)

Part (e): Predicting UK Medals

- Find point and 95% interval estimates for expected medals won by the UK (population = 58 million, GDP = \$1010 billion).

Part (e): Predicting UK Medals

- Find point and 95% interval estimates for expected medals won by the UK (population = 58 million, GDP = \$1010 billion).

Step 1: Calculate point estimate

$$\hat{E}[MEDALTOT \mid POP = 58, GDP = 1010] = \hat{\beta}_1 + \hat{\beta}_2 \ln(58) + \hat{\beta}_3 \ln(1010)$$

Part (e): Predicting UK Medals

- Find point and 95% interval estimates for expected medals won by the UK (population = 58 million, GDP = \$1010 billion).

Step 1: Calculate point estimate

$$\hat{E}[MEDALTOT \mid POP = 58, GDP = 1010] = \hat{\beta}_1 + \hat{\beta}_2 \ln(58) + \hat{\beta}_3 \ln(1010)$$

$$= 113.535 + 2.764 \times \ln(58) + 4.270 \times \ln(1010)$$

Part (e): Predicting UK Medals

- Find point and 95% interval estimates for expected medals won by the UK (population = 58 million, GDP = \$1010 billion).

Step 1: Calculate point estimate

$$\hat{E}[MEDALTOT \mid POP = 58, GDP = 1010] = \hat{\beta}_1 + \hat{\beta}_2 \ln(58) + \hat{\beta}_3 \ln(1010)$$

$$= 113.535 + 2.764 \times \ln(58) + 4.270 \times \ln(1010)$$

$$= 113.535 + 2.764 \times 4.060 + 4.270 \times 6.918$$

Part (e): Predicting UK Medals

- Find point and 95% interval estimates for expected medals won by the UK (population = 58 million, GDP = \$1010 billion).

Step 1: Calculate point estimate

$$\hat{E}[MEDALTOT \mid POP = 58, GDP = 1010] = \hat{\beta}_1 + \hat{\beta}_2 \ln(58) + \hat{\beta}_3 \ln(1010)$$

$$= 113.535 + 2.764 \times \ln(58) + 4.270 \times \ln(1010)$$

$$= 113.535 + 2.764 \times 4.060 + 4.270 \times 6.918$$

$$= 113.535 + 11.222 + 29.544 = 154.301 \text{ medals}$$

Part (e): Confidence Interval Calculation

Step 2: Calculate standard error for the prediction

Part (e): Confidence Interval Calculation

Step 2: Calculate standard error for the prediction

Our prediction is:

$$\hat{y}_{UK} = \hat{\beta}_1 + \ln(58) \times \hat{\beta}_2 + \ln(1010) \times \hat{\beta}_3$$

Part (e): Confidence Interval Calculation

Step 2: Calculate standard error for the prediction
Our prediction is:

$$\hat{y}_{UK} = \hat{\beta}_1 + \ln(58) \times \hat{\beta}_2 + \ln(1010) \times \hat{\beta}_3$$

Problem: We need the SE of this combination of coefficients

Part (e): Confidence Interval Calculation

Step 2: Calculate standard error for the prediction
Our prediction is:

$$\hat{y}_{UK} = \hat{\beta}_1 + \ln(58) \times \hat{\beta}_2 + \ln(1010) \times \hat{\beta}_3$$

Problem: We need the SE of this combination of coefficients

Key insight: The variance of $\hat{y}_{UK}$ depends on:

- The variance of each $\hat{\beta}_j$ (from robust SE)
- The covariance between the $\hat{\beta}_j$'s
- The weights $[1, \ln(58), \ln(1010)]$

Part (e): Confidence Interval Calculation

Step 2: Calculate standard error for the prediction
Our prediction is:

$$\hat{y}_{UK} = \hat{\beta}_1 + \ln(58) \times \hat{\beta}_2 + \ln(1010) \times \hat{\beta}_3$$

Problem: We need the SE of this combination of coefficients

Key insight: The variance of $\hat{y}_{UK}$ depends on:

- The variance of each $\hat{\beta}_j$ (from robust SE)
- The covariance between the $\hat{\beta}_j$'s
- The weights $[1, \ln(58), \ln(1010)]$

Solution: Use R to compute the SE accounting for all these factors:

$$\text{SE}(\hat{y}_{UK}) = 46.836$$

Part (e): Confidence Interval Calculation

Step 2: Calculate standard error for the prediction
Our prediction is:

$$\hat{y}_{UK} = \hat{\beta}_1 + \ln(58) \times \hat{\beta}_2 + \ln(1010) \times \hat{\beta}_3$$

Problem: We need the SE of this combination of coefficients

Key insight: The variance of $\hat{y}_{UK}$ depends on:

- The variance of each $\hat{\beta}_j$ (from robust SE)
- The covariance between the $\hat{\beta}_j$'s
- The weights $[1, \ln(58), \ln(1010)]$

Solution: Use R to compute the SE accounting for all these factors:

$$\text{SE}(\hat{y}_{UK}) = 46.836$$

(Note: This uses the "delta method" - a formula that combines variances and covariances of the coefficients, which Rojas covers in his 103 course but was not covered in Convery's Fall 2025 offering of 103)

Part (e): Constructing the Interval

Step 3: Build 95% confidence interval

Part (e): Constructing the Interval

Step 3: Build 95% confidence interval

Formula:

$$\text{CI} = \hat{y}_{UK} \pm t_{0.025,60} \times \text{SE}(\hat{y}_{UK})$$

Part (e): Constructing the Interval

Step 3: Build 95% confidence interval

Formula:

$$\text{CI} = \hat{y}_{UK} \pm t_{0.025,60} \times \text{SE}(\hat{y}_{UK})$$

With $t_{0.025,60} = 2.000$:

$$\text{CI} = 154.301 \pm 2.000 \times 46.836$$

Part (e): Constructing the Interval

Step 3: Build 95% confidence interval

Formula:

$$\text{CI} = \hat{y}_{UK} \pm t_{0.025,60} \times \text{SE}(\hat{y}_{UK})$$

With $t_{0.025,60} = 2.000$:

$$\text{CI} = 154.301 \pm 2.000 \times 46.836$$

$$\text{CI} = 154.301 \pm 93.672$$

Part (e): Constructing the Interval

Step 3: Build 95% confidence interval

Formula:

$$\text{CI} = \hat{y}_{UK} \pm t_{0.025,60} \times \text{SE}(\hat{y}_{UK})$$

With $t_{0.025,60} = 2.000$:

$$\text{CI} = 154.301 \pm 2.000 \times 46.836$$

$$\text{CI} = 154.301 \pm 93.672$$

$$\text{CI} = [60.629, 247.973]$$

Part (e): Constructing the Interval

Step 3: Build 95% confidence interval

Formula:

$$\text{CI} = \hat{y}_{UK} \pm t_{0.025,60} \times \text{SE}(\hat{y}_{UK})$$

With $t_{0.025,60} = 2.000$:

$$\text{CI} = 154.301 \pm 2.000 \times 46.836$$

$$\text{CI} = 154.301 \pm 93.672$$

$$\text{CI} = [60.629, 247.973]$$

Interpretation: We are 95% confident that the expected number of medals for a country with UK's characteristics is between 61 and 248 medals.

Part (f): Evaluating the Prediction

- The UK won 20 medals in 1992. Was the model successful in predicting the mean number of medals for the UK?

Part (f): Evaluating the Prediction

- (f) The UK won 20 medals in 1992. Was the model successful in predicting the mean number of medals for the UK?

Step 1: Compare actual to predicted

- Actual UK medals: 20
- Predicted: 154.301
- Prediction error: $154.301 - 20 = 134.301$ medals (off by 672%!)

Part (f): Evaluating the Prediction

(f) The UK won 20 medals in 1992. Was the model successful in predicting the mean number of medals for the UK?

Step 1: Compare actual to predicted

- Actual UK medals: 20
- Predicted: 154.301
- Prediction error: $154.301 - 20 = 134.301$ medals (off by 672%!)

Step 2: Check if actual falls in confidence interval

- 95% CI: $[60.615, 247.987]$
- Is $20 \in [60.615, 247.987]$? **NO**
- The actual value is **below** the lower bound

Part (f): Why Did the Model Fail?

Step 3: Formal hypothesis test

Test $H_0 : E[MEDALTOT \mid UK] = 20$ vs. $H_1 : E[MEDALTOT \mid UK] \neq 20$

Part (f): Why Did the Model Fail?

Step 3: Formal hypothesis test

Test $H_0 : E[MEDALTOT \mid UK] = 20$ vs. $H_1 : E[MEDALTOT \mid UK] \neq 20$

$$t = \frac{154.301 - 20}{46.836} = \frac{134.301}{46.836} = 2.867$$

Part (f): Why Did the Model Fail?

Step 3: Formal hypothesis test

Test $H_0 : E[MEDALTOT \mid UK] = 20$ vs. $H_1 : E[MEDALTOT \mid UK] \neq 20$

$$t = \frac{154.301 - 20}{46.836} = \frac{134.301}{46.836} = 2.867$$

With $|t| = 2.867 > t_{0.025,60} = 2.000$, we **reject** H_0.

Part (f): Why Did the Model Fail?

Step 3: Formal hypothesis test

Test $H_0 : E[MEDALTOT \mid UK] = 20$ vs. $H_1 : E[MEDALTOT \mid UK] \neq 20$

$$t = \frac{154.301 - 20}{46.836} = \frac{134.301}{46.836} = 2.867$$

With $|t| = 2.867 > t_{0.025,60} = 2.000$, we **reject** H_0.

Step 4: Explain the failure

- **Out-of-sample prediction**: UK excluded from estimation
- **Model misspecification**: Missing important variables
  - Sports culture and investment
  - Historical Olympic performance
  - Home advantage effects
  - Specialization in certain sports
- **Heterogeneity**: UK may be systematically different from other countries

Part (f): Lessons Learned

1. **In-sample vs. out-of-sample**: Models may not generalize well

Part (f): Lessons Learned

1. **In-sample vs. out-of-sample**: Models may not generalize well
2. **Confidence intervals**: Useful for assessing prediction quality
   - If actual value outside CI $\rightarrow$ model likely misspecified

Part (f): Lessons Learned

1. **In-sample vs. out-of-sample**: Models may not generalize well
2. **Confidence intervals**: Useful for assessing prediction quality
   - If actual value outside CI $\rightarrow$ model likely misspecified
3. **Economic vs. statistical significance**:
   - GDP is statistically significant
   - But model still fails to predict UK accurately
   - Statistical significance $\neq$ good predictions

Part (f): Lessons Learned

1. **In-sample vs. out-of-sample**: Models may not generalize well
2. **Confidence intervals**: Useful for assessing prediction quality
   - If actual value outside CI $\rightarrow$ model likely misspecified
3. **Economic vs. statistical significance**:
   - GDP is statistically significant
   - But model still fails to predict UK accurately
   - Statistical significance $\neq$ good predictions
4. **Omitted variable bias**: Need to consider country-specific factors

Part (f): Lessons Learned

1. **In-sample vs. out-of-sample**: Models may not generalize well
2. **Confidence intervals**: Useful for assessing prediction quality
   - If actual value outside CI $\rightarrow$ model likely misspecified
3. **Economic vs. statistical significance**:
   - GDP is statistically significant
   - But model still fails to predict UK accurately
   - Statistical significance $\neq$ good predictions
4. **Omitted variable bias**: Need to consider country-specific factors

Bottom line: Always validate predictions, especially out-of-sample!

Thank you!
jakeanderson@g.ucla.edu