

# Introductory Econometrics Notes and Exercises

# Math Review A

## Notes

### Summation Proofs

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\&= \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i\bar{x} + \sum_{i=1}^n \bar{x}^2 \\&= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \bar{x} \sum_{i=1}^n \bar{x} \\&= \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\&= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\&= \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \\&= \sum_{i=1}^n (x_i^2 - \bar{x}x_i) \\&= \sum_{i=1}^n x_i(x_i - \bar{x})\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x} \bar{y}) \\&= \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n y_i \bar{x} + \sum_{i=1}^n \bar{x} \bar{y} \\&= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + \bar{y} \sum_{i=1}^n \bar{x} \\&= \sum_{i=1}^n x_i y_i - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{y}\bar{x} \\&= \sum_{i=1}^n x_i y_i - n\bar{y}\bar{x} \\&= \sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i \\&= \sum_{i=1}^n (x_i y_i - \bar{x} y_i) \\&= \sum_{i=1}^n y_i (x_i - \bar{x}) = \sum_{i=1}^n x_i (y_i - \bar{y})\end{aligned}$$

## Natural Logarithm

1.  $\ln(xy) = \ln(x) + \ln(y)$
2.  $\ln(\frac{x}{y}) = \ln(x) - \ln(y)$
3.  $\ln(x^c) = c \ln(x)$

The difference in lns can be used to approximate proportionate changes. Let  $x_0$  and  $x_1$  be positive values. Then, for small changes in  $x$

$$\ln(x_1) - \ln(x_0) \approx \frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$$

Thus,

$$100 \cdot \Delta \ln(x) \approx \% \Delta x$$

## Elasticity

The **elasticity** of  $y$  with respect to  $x$  equals

$$\frac{\% \Delta y}{\% \Delta x} = \frac{(y_1 - y_0)/y_0}{(x_1 - x_0)/x_0} = \frac{\Delta y/y_0}{\Delta x/x_0} = \frac{\Delta y}{\Delta x} \cdot \frac{x_0}{y_0}$$

Defining a linear model  $y = \beta_0 + \beta_1 x$ , the elasticity of  $y$  with respect to  $x$  equals

$$\frac{\% \Delta y}{\% \Delta x} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\beta_1 \Delta x}{\Delta x} \cdot \frac{x}{\beta_0 + \beta_1 x} = \beta_1 \cdot \frac{x}{\beta_0 + \beta_1 x} \approx \frac{\Delta \ln(y)}{\Delta \ln(x)}$$

If we use the above approximation for both  $x$  and  $y$ , then the elasticity is approximately equal to  $\frac{\Delta \ln(y)}{\Delta \ln(x)}$ . Thus, a **constant elasticity model** is approximated by

$$\ln(y) = \beta_0 + \beta_1 \ln(x)$$

where  $\beta_1$  is the approximate elasticity of  $y$  with respect to  $x$ .

A **semi-elasticity model** approximates the percentage change in  $y$  with respect to a unit change in  $x$  and takes the form

$$\ln(y) = \beta_0 + \beta_1 x$$

where  $\beta_1$  is the semi-elasticity of  $y$  with respect to  $x$ . In other words,  $\% \Delta y = 100 \beta_1 \Delta x \rightarrow \beta_1 \approx \frac{\% \Delta y}{100 \Delta x}$ .

Another relationship of some interest is

$$y = \beta_0 + \beta_1 \ln(x)$$

Using calculus, we can derive

$$\Delta y = \beta_1 \Delta \ln(x)$$

and thus

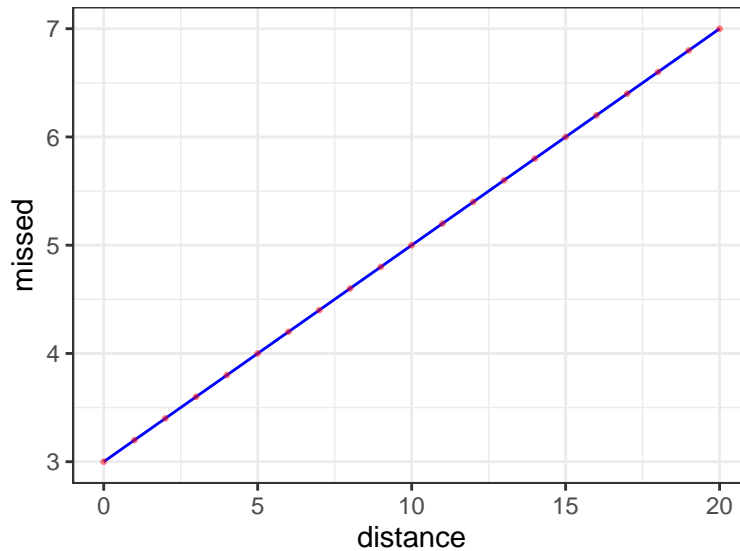
$$\beta_1 = \frac{\Delta y}{\Delta \ln(x)} \approx \frac{\Delta y}{\frac{\% \Delta x}{100}}$$

In other words,  $\beta_1/100$  is the unit change in  $y$  when  $x$  increases by 1%.

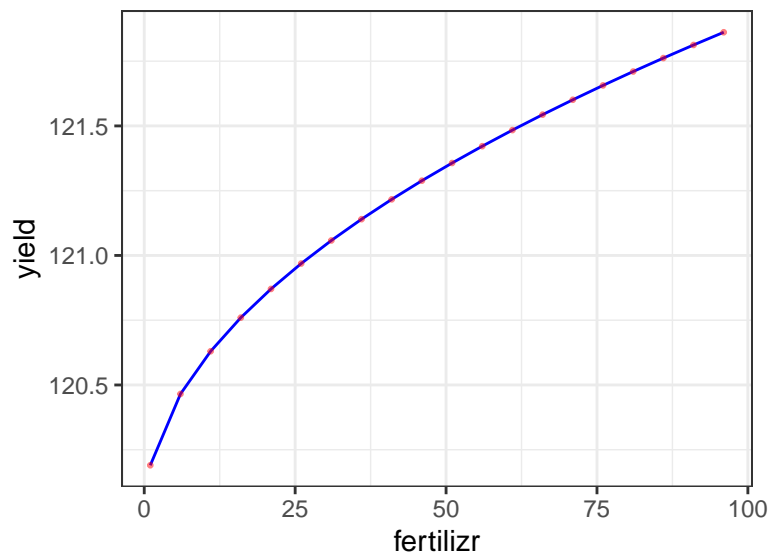
## Exercises

1. (a) Mean = 566  
(b) Median = 505  
(c) Mean = 5.66 Median = 5.05  
(d) Mean = 586 Median = 505

2. (a) See below
- (b) 4 classes
- (c) 2 classes



3. -21 CDs. This suggests using linear functions to describe demand curves may not be realistic/a good idea. Some form of an elasticity model would likely be more suitable.
4. (a) A 0.8 percentage point decrease.
- (b) A 0.125% fall.
5. The correct terminology would be the stock return increased by 3 percentage points, a 20% increase in the return on the stock.
6. (a) 20%
- (b)  $\approx 18.2321557\%$
7. (a) \$ 40134.84
- \$ 45935.80
- (b)  $\% \Delta \text{salary} = 100(.027)(5) = 13.5\%$
- (c) 7.0642847% error
8. The intercept indicates that, with no sales tax, the proportionate growth in employment would be .043 units. The slope indicates that for every unit increase in sales tax, we would expect the proportionate growth in employment to decrease by .78 units.
9. (a) See below
- (b) The most notable difference is the convexity/concavity of the functions. A linear model would have constant marginal returns to yield with respect to fertilizer while the given relationship displays diminishing marginal returns.



10. (a) It's not of much interest by itself. It suggests a class with 0 students would expect a test score of 45.6, which doesn't make sense or have any real meaning.

(b)

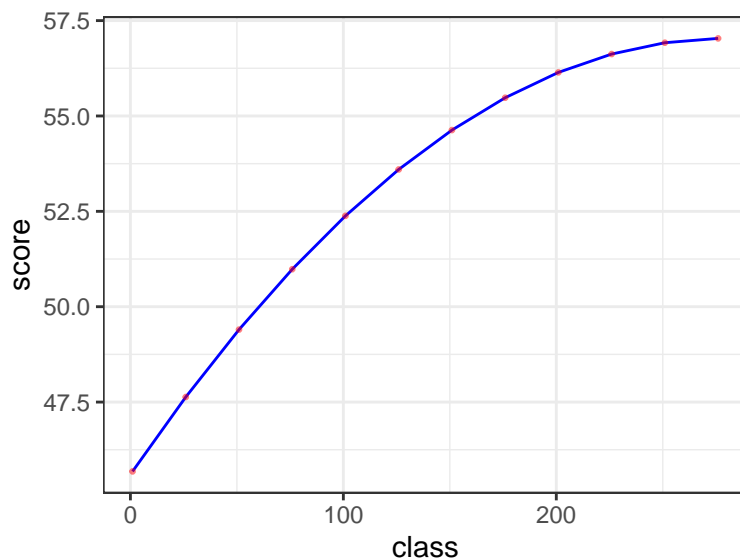
$$\frac{\partial \text{score}}{\partial \text{class}} = .082 - .000294 \cdot \text{class} = 0$$

$$\rightarrow \text{class}^* = \frac{.082}{.000294} \approx 279 \text{ students}$$

The highest achievable test score is about 57.

(c) See below

- (d) No, this equation may give an idea about what one can expect for **score** given **class**, but it's unrealistic to expect exact results.



11. (a)

$$\begin{aligned}\bar{y} &= \frac{y_1 + y_2}{2} = \frac{\beta_0 + \beta_1 x_1 + \beta_0 + \beta_1 x_2}{2} \\ &= \frac{2\beta_0 + \beta_1(x_1 + x_2)}{2} \\ &= \beta_0 + \beta_1 \frac{(x_1 + x_2)}{2} = \beta_0 + \beta_1 \bar{x}\end{aligned}$$

(b)

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^n \beta_0 + \beta_1 x_i}{n} \\ &= \frac{n\beta_0 + \beta_1 \sum_{i=1}^n x_i}{n} \\ &= \beta_0 + \beta_1 \frac{n\bar{x}}{n} = \beta_0 + \beta_1 \bar{x}\end{aligned}$$

12. (a)

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \left( \sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^n x_i \right) \\ &= \frac{1}{n} (n_1 \bar{x}_1 + n_2 \bar{x}_2) \\ &= \frac{n_1}{n} \bar{x}_1 + \frac{n_2}{n} \bar{x}_2 \\ &= w_1 \bar{x}_1 + w_2 \bar{x}_2\end{aligned}$$

(b) Yes, they represent the relative portions of the sample space in  $i = 1, \dots, n_1$  and  $i = n_1 + 1, \dots, n$ .

(c) The case in part (a) applies to all cases for  $g \in \mathbb{Z}^+$ .

13. (a) No, take the sample  $\{x_1, x_2\} = \{2, 2\}$ . Then,

$$\sum_{i=1}^n \frac{1}{x_i} = \frac{1}{2} + \frac{1}{2} = 1$$

while

$$\frac{1}{\sum_{i=1}^n x_i} = \frac{1}{2+2} = \frac{1}{4}$$

(b) No, see part (a) where  $x_i = 2 \forall i$ .

## Math Review B

### Notes

#### Experiments

An **experiment** is a procedure that can theoretically be conducted an infinite number of times and has a well-defined set of outcomes.

A **random variable** is a variable that takes on numerical values and has an outcome determined by an experiment.

## Variables

A **Bernoulli** (or **binary**) **random variable** is a random variable that can only take on the values zero and one.

A **discrete random variable** is one that takes on only a finite or countably infinite number of values. A Bernoulli random variable is the simplest example of a discrete random variable.

A **continuous random variable** is a random variable that takes on any real value with *zero* probability.

## Density Functions

A **probability density function (pdf)** summarizes the information concerning the possible outcomes of a random variable and the corresponding probabilities. A pdf of a random variable  $X$  is generally denoted  $f(x)$  or  $f_x \equiv P(X = x)$ .

A **cumulative distribution function (cdf)** is a function that describes the cumulative probability that a random variable's value is less than (or equal to, if continuous) a given value. A cdf of a random variable  $X$  is generally denoted  $F(x)$  or  $F_x \equiv P(X \leq x)$

Cumulative Distribution Function Properties:

1. For any number  $c$ ,  $P(X > c) = 1 - F(c)$
2. For any numbers  $a < b$ ,  $P(a < X \leq b) = F(b) - F(a)$
3. For a continuous random variable  $X$ ,  $P(X \geq c) = P(X > c)$
4. For a continuous random variable  $X$ ,  $P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$

## Independence

Let  $X$  and  $Y$  be discrete random variables. Then,  $(X, Y)$  have a **joint distribution**, which is fully described by the **joint probability density function** of  $(X, Y)$ :

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

Two random variables  $X$  and  $Y$  are said to be **independent** if, and only if,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

or

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all  $x$  and  $y$ . The pdfs  $f_X$  and  $f_Y$  are often called the **marginal probability density functions** to distinguish them from the joint pdf  $f_{X,Y}$ .

Beyond the case of two random variables, the same concept applies. Random variables  $X_1, X_2, \dots, X_n$  are **independent random variables** if, and only if, their joint pdf is the product of the individual pdfs for any  $(x_1, x_2, \dots, x_n)$ . This definition of independence holds for both continuous and discrete random variables.

Given independent outcomes with 'success' rate  $\theta$ , the pdf ( $X \sim \text{Binomial}(n, \theta)$ ) is equal to  $f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$  where  $\binom{n}{x} = {}^nC_x = \frac{n!}{x!(n-x)!}$ .

## Conditional Distributions

The **conditional distribution** of  $Y$  given  $X$  is summarized by the **conditional probability density function**, defined by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

for all values of  $x$  such that  $f_X(x) > 0$ . The interpretation of the conditional probability density function is

$$f_{Y|X}(y|x) = P(Y = y|X = x)$$

## Expected Values

If  $X$  is a random variable, the **expected value** (or **expectation**) of  $X$ , denoted  $E[X]$  and sometimes  $\mu_X$  or simply  $\mu$ , is a weighted average of all possible values of  $X$ . The weights are determined by the probability distribution function. Sometimes, the expected value is called the *population mean*, especially when we want to emphasize that  $X$  represents some variable in a population. For a discrete random variable  $X$  that takes on values  $\{x_1, \dots, x_n\}$ ,

$$E[X] = x_1 f(x_1) + \dots + x_n f(x_n) = \sum_{i=1}^n x_i f(x_i)$$

If  $X$  is a continuous random variable, then  $E[X]$  is defined through an integral as

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

which we assume is well-defined.

### Expected Values Properties

1. For a constant  $c$ ,  $E[c] = c$
2. For any constants  $a$  and  $b$ ,  $E[aX + b] = aE[X] + b$
3. If  $\{a_1, a_2, \dots, a_n\}$  are constants and  $\{X_1, X_2, \dots, X_n\}$  are random variables, then

$$E[a_1 X_1 + a_2 X_2 + \dots + a_n X_n] = E\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i E[X_i]$$

For  $X \sim \text{Binomial}(n, \theta)$ , we can rewrite  $X$  as  $Y_1 + \dots + Y_n$ , where each  $Y_i \sim \text{Bernoulli}(\theta)$ . Then,

$$E[X] = \sum_{i=1}^n E[Y_i] = \sum_{i=1}^n \theta = n\theta$$

## Variance

For a random variable  $X$ ,

$$\text{Var}(x) = E[(X - \mu)^2]$$

The **variance** tells us the expected distance from  $X$  to its mean and is sometimes denoted  $\sigma_x^2$  or just  $\sigma^2$ . For a Bernoulli random variable  $X$

$$\sigma_x^2 = E[X^2] - E[X]^2 = \theta - \theta^2 = \theta(1 - \theta)$$

### Variance Properties

1.  $\text{Var}(x) = 0$  if, and only if, there is a constant  $c$  such that  $P(X = c) = 1$ , in which case  $E[X] = c$ . In other words, this first property says that the variance of any constant is zero and if a random variable has zero variance, then it is essentially constant.
2. For any constants  $a$  and  $b$ ,  $\text{Var}(aX + b) = a^2 \text{Var}(x)$
3. For any constants  $a$  and  $b$ ,

$$\text{Var}(aX + bY) = a^2 \text{Var}(x) + 2ab \text{Cov}(X, Y) + b^2 \text{Var}(Y)$$

4. If  $\{X_1, \dots, X_n\}$  are pairwise uncorrelated random variables and  $a_i : i = 1, \dots, n$  are constants then

$$\text{Var}(a_1 X_1 + \dots + a_n X_n) = a_1^2 \text{Var}(X_1) + \dots + a_n^2 \text{Var}(X_n) = \text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$$



## Standard Deviation

The **standard deviation** of a random variable, denoted  $\text{sd}(X)$ , is simply the positive square root of the variance:  $\text{sd}(X) = +\sqrt{\text{Var}(X)}$ . The standard deviation is sometimes denoted  $\sigma_X$ , or simply  $\sigma$ , when the random variable is understood.

### Standard Deviation Properties

1. For any constant  $c$ ,  $\text{sd}(c) = 0$ .
2. For any constants  $a$  and  $b$ ,

$$\text{sd}(aX + b) = |a|\text{sd}(X) = |a|\sigma_X$$

## Standardized Random Variables

If  $X$  is a random variable, we can redefine a random variable

$$Z \equiv \frac{X - \mu}{\sigma},$$

which we can write as  $Z = aX + b$ , where  $a \equiv (1/\sigma)$  and  $b \equiv -(\mu/\sigma)$ . Then,

$$E[Z] = aE[X] + b = (\mu/\sigma) - (\mu/\sigma) = 0$$

and

$$\text{Var}(Z) = a^2\text{Var}(X) = 1$$

Thus, the random variable  $Z$  has a mean of zero and a variance (and therefore a standard deviation) equal to one. This procedure is sometimes known as standardizing the random variable  $X$ , and  $Z$  is called a **standardized random variable**.

## Skewness and Kurtosis

We can use the standardized version of a random variable to define other features of the distribution of a random variable. These features are described by using what are called *higher order moments*. For example, the third moment of the standardized random variable  $Z$  is used to determine whether a distribution is symmetric about its mean. We can write

$$E[Z^3] = \frac{E[(X - \mu)^3]}{\sigma^3}$$

Generally,  $\frac{E[(X - \mu)^3]}{\sigma^3}$  is viewed as a measure of **skewness** in the distribution of  $X$ . If  $X$  has a symmetric distribution about  $\mu$ , then  $Z$  has a symmetric distribution about zero. That means the density of  $Z$  at any two points  $z$  and  $-z$  is the same.

It also can be informative to compute the fourth moment of  $Z$

$$E[Z^4] = \frac{E[(X - \mu)^4]}{\sigma^4}$$

The fourth moment  $E[Z^4]$  is called a measure of **kurtosis** in the distribution of  $X$ . Generally, larger values mean that the tails in the distribution of  $X$  are thicker.

## Covariance and Correlation

The **covariance** between two random variables  $X$  and  $Y$ , sometimes called the *population covariance* to emphasize that it concerns the relationship between two variables describing a population, is defined as the expected value of the product  $(X - \mu_X)(Y - \mu_Y)$ :

$$\text{Cov}(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)]$$

which is sometimes denoted  $\sigma_{XY}$ . If  $\sigma_{XY} > 0$ , then, on average, when  $X$  is above its mean,  $Y$  is also above its mean. If  $\sigma_{XY} < 0$ , then, on average, when  $X$  is above its mean,  $Y$  is below its mean. Note that

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] = E[(X - \mu_X)Y] \\ &= E[X(Y - \mu_Y)] = E[XY] - E[X]E[Y]\end{aligned}$$

Covariance measures the amount of linear dependence between two random variables. A positive covariance indicates that two random variables move in the same direction, while a negative covariance indicates they move in opposite directions.

#### Covariance Properties

1. If  $X$  and  $Y$  are independent, then  $\text{Cov}(X, Y) = 0$

This property stems from the fact that  $E[XY] = E[X]E[Y]$  when  $X$  and  $Y$  are independent. It is important to remember that the converse of is not true: zero covariance between  $X$  and  $Y$  does not imply that  $X$  and  $Y$  are independent.

2. For any constants  $a_1, b_1, a_2$ , and  $b_2$ ,

$$\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X, Y)$$

3. From the **Cauchy-Schwartz inequality**:

$$|\text{Cov}(X, Y)| \leq \text{sd}(X)\text{sd}(Y)$$

The fact that the covariance depends on units of measurement is a deficiency that is overcome by the **correlation coefficient** between  $X$  and  $Y$ :

$$\text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

the correlation coefficient between  $X$  and  $Y$  is sometimes denoted  $\rho_{XY}$  (and is sometimes called the *population correlation*).

#### Correlation Properties

1.  $-1 \leq \text{Corr}(X, Y) \leq 1$

2. For any constants  $a_1, b_1, a_2$ , and  $b_2$ ,

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = \text{Corr}(X, Y)$$

if  $a_1a_2 > 0$  or

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = -\text{Corr}(X, Y)$$

if  $a_1a_2 < 0$ .

#### **Conditional Expectation**

The **conditional expectation** of a random variable is the expected or average value of one random variable, called the dependent or explained variable, that depends on the values of one or more other variables, called the independent or explanatory variables. When  $Y$  is a discrete random variable

$$E[Y|x] = \sum_{i=1}^n y_i f_{Y|X}(y_i|x)$$

When  $Y$  is a continuous random variable

$$E[Y|x] = \int_{-\infty}^{\infty} y_i f_{Y|X}(y_i|x) dy$$

### Conditional Expectation Properties

1.  $E[c(X)|X] = c(X)$ , for any function  $c(X)$ .
2. For any functions  $a(X)$  and  $b(X)$ ,

$$E[a(X)Y + b(X)|X] = a(X)E[Y|X] + b(X)$$

3. If  $X$  and  $Y$  are independent,  $E[Y|X] = E[Y]$ .
4. From the **law of iterated expectations**  $E[E[Y|X]] = E[Y]$ .
5. From a more general version of the law of iterated expectation  $E[Y|X] = E[E[Y|X, Z]|X]$ .
6. If  $E[Y|X] = E[Y]$ , then  $\text{Cov}(X, Y) = 0$ .
7. If  $E[Y^2] < \infty$  and  $E[g(X)^2] < \infty$  for some function  $g$ , then  $E[[Y - E[Y|X]]^2|X] \leq E[[Y - g(X)]^2|X]$  and  $E[[Y - E[Y|X]]^2] \leq E[[Y - g(X)]^2]$ . This property is very useful in predicting or forecasting contexts. The first inequality says that, if we measure prediction inaccuracy as the expected squared prediction error, conditional on  $X$ , then the conditional mean is better than any other function of  $X$  for predicting  $Y$ . The conditional mean also minimizes the unconditional expected squared prediction error.

### **Conditional Variance**

Given random variables  $X$  and  $Y$ , the variance of  $Y$ , conditional on  $X = x$ , is simply the variance associated with the conditional distribution of  $Y$ , given  $X = x : E[(Y - E[Y|x])^2|x]$ . The formula can be rewritten as

$$\text{Var}(Y|X = x) = E[Y^2|x] - E[Y|x]^2$$

### Conditional Variance Properties

1. If  $X$  and  $Y$  are independent, then  $\text{Var}(Y|X) = \text{Var}(Y)$

### **Normal Distribution**

A **normal random variable** is a continuous random variable that can take on any value. Its probability density function has the familiar bell-shaped graph and can be written as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x - \mu)^2/2\sigma^2], \quad -\infty < x < \infty$$

We say that  $X$  has a **normal distribution** with expected value  $\mu$  and variance  $\sigma^2$ , written as  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Because the normal distribution is symmetric about  $\mu$ ,  $\mu$  is also the median of  $X$ . The normal distribution is also sometimes called the Gaussian distribution after Carl Friedrich Gauss.

One special case of the normal distribution is the **standard normal distribution** where the mean is zero and the variance is unity. If a random variable  $Z$  has a  $\text{Normal}(0,1)$  distribution, then we say it has a standard normal distribution, and its pdf is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \quad -\infty < z < \infty$$

### Normal Distribution Properties

1. If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$
2. If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
3. If  $X$  and  $Y$  are jointly normally distributed, then they are independent if, and only if,  $\text{Cov}(X, Y) = 0$
4. Any linear combination of independent, identically distributed normal random variables has a normal distribution.

## Chi-Square Distribution

The chi-square distribution is obtained directly from independent, standard normal random variables. Let  $Z_i, i = 1, 2, \dots, n$  be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the  $Z_i$ :

$$X = \sum_{i=1}^n Z_i^2$$

Then,  $X$  has what is known as a **chi-square distribution** with  $n$  **degrees of freedom**. We write this as  $X \sim \chi_n^2$  where the expected value of  $X$  is  $n$  and the variance of  $X$  is  $2n$ .

## t Distribution

A **t distribution** is obtained from a standard normal and a chi-square random variable. Let  $Z$  have a standard normal distribution and let  $X$  have a chi-square distribution with  $n$  degrees of freedom. Further, assume that  $Z$  and  $X$  are independent. Then, the random variable

$$T = \frac{Z}{\sqrt{X/n}}$$

has a  $t$  distribution with  $n$  degrees of freedom. This is denoted by  $T \sim t_n$  where  $n$  comes from the degrees of freedom of the chi-square random variable in the denominator. The pdf of the  $t$  distribution has a shape similar to that of the standard normal distribution (maintaining a zero expected value), except that it is more spread out (with a variance of  $n/(n-2)$ ) and therefore has more area in the tails. As the degrees of freedom gets large, the  $t$  distribution approaches the standard normal distribution.

## F Distribution

To define an F random variable, let  $X_1 \sim \chi_{k_1}^2$ , and  $X_2 \sim \chi_{k_2}^2$  and assume that  $X_1$  and  $X_2$  are independent. Then, the random variable

$$F = \frac{X_1/k_1}{X_2/k_2}$$

has an **F distribution** with  $(k_1, k_2)$  degrees of freedom. This is denoted as  $F \sim F_{k_1, k_2}$

## Exercises

1. His or her eventual SAT score is viewed as a random variable because his or her score is a variable that takes on numerical values and has an outcome determined by an experiment (the test). The test score is stochastic as it can change from day-to-day or depending on other various circumstances/conditions.
2. (a)  $P(X \leq 6) = 0.6914625$   
(b)  $P(X > 4) = 1 - P(X \leq 4) = 0.6914625$   
(c)  $P(|X - 5| > 1) = P([X - 5 > 1] \text{ or } [X - 5 < -1]) = 1 - P(4 < X < 6) = 1 - (P(X < 6) - P(X < 4)) = 0.6170751$
3. (a) 0.0009766 or 0.0976562%  
(b) 4.0722656 mutual funds  
(c)  $P(\text{At Least One}) = 1 - P(\text{None}) = 1 - (1 - .5^{10})^{4170} = 0.9829951$   
Similarly, this equals  $1 - \binom{4170}{0}(.5^{10})^0(1 - .5^{10})^{4170} = 0.9829951$   
(d)  $P(X \geq 5) = 1 - P(X < 4) = 0.3852852$
4.  $P(X \geq .6) = 1 - P(X < .6) = F(.6) = 1.512$

5. (a)  $P(\text{At least one}) = 1 - P(\text{None}) = 1 - \binom{12}{0}(.2)^0(.8)^{12} = 0.9312805$   
 (b)  $P(X \geq 2) = 1 - P(X < 2) = 1 - \binom{12}{1}(.2)^1(.8)^{11} = 0.7251221$
6.  $E[X] = \int_0^3 \frac{x^2}{9} x dx = \frac{1}{9} \int_0^3 x^3 dx = \frac{1}{9} \left[ \frac{x^4}{4} \right]_0^3 = \frac{81}{36} = \frac{9}{4}$
7.  $E[\text{Made FTs}] = .74 * 8 = 5.92$
8.  $E[GPA] = 3.5(\frac{2}{9}) + 3(\frac{7}{9}) = \frac{28}{9} \approx 3.11$
9.  $E[\text{salary}] = 52.3 \times 1000 = \$52300$   
 $\sigma_{\text{salary}} = |1000| \times 14.6 = \$14600$
10. (a)  $E[GPA|SAT = 800] = .70 + .002(800) = 2.3$   
 $E[GPA|SAT = 1400] = .70 + .002(1400) = 3.5$  The difference in expected GPAs is fairly large, but the difference in SAT scores is also rather large. I don't feel these estimates are entirely unreasonable.  
 (b)  $E[GPA] = E[E[GPA|SAT]] = E[.70 + .002(1100)] = .70 + .002(1100) = 2.9$   
 (c) No, we don't know any particular student's GPA given his or her SAT score. The provided formula only allows us to derive an expected GPA given an SAT score.
11. (a)  $E[X] = 1/2(-1) + 1/2(1) = 0$   
 $E[X^2] = 1/2(-1)^2 + 1/2(1)^2 = 1$   
 (b)  $E[X] = 1/2(1) + 1/2(2) = 3/2$   
 $E[1/X] = 1/2(1) + 1/2(1/2) = 3/4$   
 (c) From part(a),  $E[X^2] = 1 \neq (E[X])^2 = 0^2 = 0$   
 From part(b),  $E[1/X] = 3/4 \neq (1/E[X]) = 2/3$   
 (d)

$$E[F] = E \left[ \frac{X_1/k_1}{X_2/k_2} \right]$$

Because  $k_1$  and  $k_2$  are constants,

$$= \frac{k_2}{k_1} E \left[ \frac{X_1}{X_2} \right]$$

Using the fact that  $X_1$  and  $X_2$  are assumed independent

$$= \frac{k_2}{k_1} E[X_1] E[X_2^{-1}]$$

Using the fact that  $X_1$  is a chi-square random variable (and thus has a mean of  $k_1$ )

$$= \frac{k_2}{k_1} k_1 E[X_2^{-1}] = k_2 E[X_2^{-1}] = E \left[ \frac{k_2}{X_2} \right] = E \left[ \frac{1}{X_2/k_2} \right]$$

As we showed in parts (a-c),  $E \left[ \frac{1}{X_2/k_2} \right]$  has a nonlinear 'internal' function, and thus, we cannot conclude that  $E[F] = 1$ .

# Math Review C

## Notes

### Populations, Parameters, and Random Sampling

A **population** is any well-defined group of subjects, which could be individuals, firms, cities, or many other possibilities.

A **random sample** is a sample obtained by sampling randomly from the specified population. In particular, no unit is more likely to be selected than any other unit, and each draw is independent of all other draws.

When  $\{Y_1, \dots, Y_n\}$  is a random sample from the density  $f(y; \theta)$ , we also say that the  $Y_i$  are **independent, identically distributed** (or **i.i.d.**) random variables from  $f(y; \theta)$ .

### Estimators

An **estimator** is a rule for combining data to produce a numerical value for a population parameter; the form of the rule does not depend on the particular sample obtained. More generally, an estimator  $W$  of a parameter  $\theta$  can be expressed as an abstract mathematical formula:

$$W = h(Y_1, Y_2, \dots, Y_n)$$

for some known function  $h$  of the random variables  $Y_1, Y_2, \dots, Y_n$

### Unbiasedness

An estimator,  $W$  of  $\theta$ , is an **unbiased estimator** if

$$E[W] = \theta$$

for all possible values of  $\theta$ .

The distribution of an estimator is often called its **sampling distribution**, because this distribution describes the likelihood of various outcomes of  $W$  across different random samples.

If  $W$  is an estimator of  $\theta$ , its **bias** is defined

$$\text{Bias}(W) \equiv E[W] - \theta$$

The **sample average** is an unbiased estimator of the population variance and is defined as

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

### Proof of Unbiasedness

$$\begin{aligned} E[\bar{Y}] &= E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^n Y_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[Y_i] = \frac{1}{n} (n\mu) = \mu \end{aligned}$$

The **sample variance** is an unbiased estimator of the population variance and is defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Proof of Unbiasedness

$$\begin{aligned} E[S^2] &= E \left[ \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \right] \\ &= \frac{1}{n-1} E \left[ \sum_{i=1}^n Y_i^2 - 2 \sum_{i=1}^n Y_i \bar{Y} + \sum_{i=1}^n \bar{Y}^2 \right] \\ &= \frac{1}{n-1} E \left[ \sum_{i=1}^n Y_i^2 - 2\bar{Y} \sum_{i=1}^n Y_i + \bar{Y} \sum_{i=1}^n \bar{Y} \right] \\ &= \frac{1}{n-1} E \left[ \sum_{i=1}^n Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2 \right] \\ &= \frac{1}{n-1} \left\{ \sum_{i=1}^n E[Y_i^2] - nE[\bar{Y}^2] \right\} \end{aligned}$$

Using the facts that  $\text{Var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$  and  $\sigma_Y^2 = E[Y_i^2] - E[Y_i]^2$ ,

$$\begin{aligned} &= \frac{1}{n-1} \left\{ \sum_{i=1}^n (\sigma_Y^2 + E[Y_i]^2) - n \left( \frac{\sigma_Y^2}{n} + E[\bar{Y}]^2 \right) \right\} \\ &= \frac{1}{n-1} \left\{ n\sigma_Y^2 + n\mu_Y^2 - \sigma_Y^2 - n\mu_Y^2 \right\} \\ &= \frac{1}{n-1} [(n-1)\sigma_Y^2] = \sigma_Y^2 \end{aligned}$$

The **sample covariance** is defined as

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

and is an unbiased and consistent estimator of  $\sigma_{XY}$

The **sample correlation coefficient** is defined as

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

and is a consistent but biased estimator of  $\rho_{XY}$ . Because  $S_{XY}$ ,  $S_X$ , and  $S_Y$  are consistent for the corresponding population parameter,  $R_{XY}$  is a consistent estimator of the population correlation,  $\rho_{XY}$ . However,  $R_{XY}$  is a biased estimator for two reasons. First,  $S_X$  and  $S_Y$  are biased estimators of  $\sigma_X$  and  $\sigma_Y$ , respectively. Second,  $R_{XY}$  is a ratio of estimators, so it would not be unbiased, even if  $S_X$  and  $S_Y$  were.

## Efficiency

The variance of an estimator is often called its **sampling variance** because it is the variance associated with a sampling distribution. Remember, the sampling variance is not a random variable; it is a constant, but it might be unknown.

An estimator,  $W_1$ , is **efficient** relative to another estimator,  $W_2$ , when  $\text{Var}(W_1) \leq \text{Var}(W_2)$  for all  $\theta$ , with strict inequality for at least one value of  $\theta$ .

One way to compare estimators that are not necessarily unbiased is to compute the **mean squared error (MSE)** of the estimators. If  $W$  is an estimator of  $\theta$ , then the MSE of  $W$  is defined as

$$\text{MSE}(W) = E[(W - \theta)^2]$$

The MSE measures how far, on average, the estimator is away from  $\theta$ . It can be shown that  $\text{MSE}(W) = \text{Var}(W) + [\text{Bias}(W)]^2$ , so that  $\text{MSE}(W)$  depends on the variance and bias (if any is present). This allows us to compare two estimators when one or both are biased.

## Consistency

Let  $W_n$  be an estimator of  $\theta$  based on a sample  $Y_1, Y_2, \dots, Y_n$  of size  $n$ . Then,  $W_n$  is a **consistent estimator** of  $\theta$  if for every  $\epsilon > 0$ ,

$$P(|W_n - \theta| > \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

When  $W_n$  is consistent, we also say that  $\theta$  is the probability limit of  $W_n$ , written as  $\text{plim}(W_n) = \theta$ . Unlike unbiasedness—which is a feature of an estimator for a given sample size—consistency involves the behavior of the sampling distribution of the estimator as the sample size  $n$  gets large.

Asymptotic Unbiasedness  $\leftarrow$  Consistency + Bounded Variance

Consider an estimator  $W_n$  for a parameter  $\theta$ . Asymptotic unbiasedness means that the bias of the estimator goes to zero as  $n \rightarrow \infty$ , which means that the expected value of the estimator converges to the true value of the parameter. Consistency is a stronger condition than this; it requires the estimator (not just its expected value) to converge to the true value of the parameter (with convergence interpreted in various ways). Since there is generally some non-zero variance in the estimator, it will not generally be equal to (or converge to) its expected value. Assuming the variance of the estimator is bounded, consistency ensures asymptotic unbiasedness, but asymptotic unbiasedness is not enough to get consistency. To put it another way, under some mild conditions, asymptotic unbiasedness is a necessary but not sufficient condition for consistency.

Asymptotic Unbiasedness + Vanishing Variance  $\rightarrow$  Consistency

If you have an asymptotically unbiased estimator, and its variance converges to zero, this is sufficient to give weak consistency. (This follows from Markov's inequality, which ensures that convergence in mean-square implies convergence in probability). Intuitively, this reflects the fact that a vanishing variance means that the sequence of random variables is converging closer and closer to the expected value, and if the expected value converges to the true parameter (as it does under asymptotic unbiasedness) then the random variable is converging to the true parameter.

More simply, unbiased estimators are not necessarily consistent, but those whose variances shrink to zero as the sample size grows are *consistent*. For example, the sample variance and standard deviation formulas without **Bessel's correction** are biased estimators; however, they are also consistent because they converge in probability toward their population values as  $n \rightarrow \infty$ .

The **law of large numbers (LLN)** is a theorem that states the average from a random sample converges in probability to the population average. It also holds for stationary and weakly dependent time series. This result comes from the fact that  $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$ , which approaches 0 as  $n \rightarrow \infty$ .



## Plim Properties

1. If  $\theta$  is a parameter and  $\gamma = g(\theta)$  is a newly-defined parameter for some continuous function  $g(\theta)$ . If  $\text{plim}(W_n) = \theta$ , then the estimator of  $\gamma$ ,  $G_n = g(W_n)$ , has a plim defined by

$$\text{plim}(G_n) = \gamma$$

This is often stated as

$$\text{plim}(g(W_n)) = g(\text{plim}(W_n))$$

2. If  $\text{plim}(T_n) = \alpha$  and  $\text{plim}(U_n) = \beta$ , then

(a)  $\text{plim}(T_n + U_n) = \alpha + \beta$

(b)  $\text{plim}(T_n U_n) = \alpha\beta$

(c)  $\text{plim}(T_n/U_n) = \alpha/\beta$ , if  $\beta \neq 0$

## Asymptotic Normality

Let  $\{Z_n : n = 1, 2, \dots\}$  be a sequence of random variables, such that for numbers  $z$ ,

$$P(Z_n \leq z) \rightarrow \Phi(z) \text{ as } n \rightarrow \infty$$

where  $\Phi(z)$  is the standard normal cumulative distribution function. Then,  $Z_n$  is said to have an **asymptotic standard normal distribution**. This is sometimes written as  $Z_n \overset{a}{\sim} \mathcal{N}(0, 1)$ , where the “a” stands for “asymptotically” or “approximately.”

The **central limit theorem (CLT)** states that the average from a random sample (and many other estimators that depend on the sample mean) for any population (with finite variance), when standardized, has an asymptotic standard normal distribution. More formally, for a random sample  $\{Y_1, Y_2, \dots, Y_n\}$  with a mean  $\mu$  and a variance  $\sigma^2$ . Then,

$$Z_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

has an asymptotic standard normal distribution. Note that  $Z_n$  is the standardized version of  $\bar{Y}_n$ :  $E[\bar{Y}_n] = \mu$  has been subtracted off and divided by  $\text{sd}(\bar{Y}_n) = \sigma/\sqrt{n}$ .

## Maximum Likelihood

The **maximum likelihood estimator** of  $\theta$ , call it  $W$ , is the value of  $\theta$  that maximizes the **likelihood function**

$$L(\theta; Y_1, Y_2, \dots, Y_n) = f(Y_1; \theta)f(Y_2; \theta) \dots f(Y_n; \theta)$$

which equals  $P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$  in the discrete case. Usually, it is more convenient to work with the **log-likelihood function**, which is obtained by taking the natural log of the likelihood function:

$$\mathcal{L}(\theta) = \ln(L(\theta; Y_1, Y_2, \dots, Y_n)) = \sum_{i=1}^n \ln[f(Y_i, \theta)] = \sum_{i=1}^n \ell(\theta; X_i)$$

where we use the fact that the log of the product is the sum of the logs.

## Least Squares

A least squares estimator is an estimator of a parameter that minimizes the sum of squared differences. That is, an estimator,  $W$  is a least squares estimator if it minimizes

$$\sum_{i=1}^n (W - \theta)^2$$

It should be noted that the principles of least squares, method of moments, and maximum likelihood often result in the same estimator. In other cases, the estimators are similar but not identical.

## Confidence Intervals

A **confidence interval** is a rule used to construct a random interval so that a certain percentage of all data sets, determined by the confidence level, yields an interval that contains the population value. Thus, a 95% confidence interval of an estimator will contain the true population value 95% of the time. Theoretically, for the sample average the 95% confidence interval can be constructed as follows:

$$P\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$
$$\rightarrow CI_{95} = [\bar{y} - 1.96(\sigma/\sqrt{n}), \bar{y} + 1.96(\sigma/\sqrt{n})]$$

where  $\mu$  is the hypothesized population mean. In practice, however,  $\sigma$  is unknown and must be estimated with  $s$ . Unfortunately, this does not preserve the 95% level of confidence because  $s$  depends on the particular sample. In other words, the random interval  $[\bar{Y} \pm 1.96(S/\sqrt{n})]$  no longer contains  $\mu$  with probability .95 because the constant  $\sigma$  has been replaced with the random variable  $S$ . Thus, rather than using the standard normal distribution, we must rely on the  $t$  distribution. The  $t$  distribution arises from the fact that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

The denominator  $S/\sqrt{n}$  is an estimate of the  $\text{sd}(\bar{Y})$ . In general, these estimators of **sampling standard deviations** are referred to as **standard errors**.

## Hypothesis Testing

A **Type I error** is an error in which a true null hypothesis is rejected.

A **Type II error** is an error in which one fails to reject a false null hypothesis.

A **significance level** is the probability of committing a Type I error, which is generally denoted

$$\alpha = P(\text{Reject } H_0 | H_0)$$

A  **$p$ -value** is the *largest* significance level at which we could carry out a test and still fail to reject the null hypothesis. Formally,

$$p\text{-value} = P(T > t | H_0) = 1 - \Phi(t)$$

where  $\Phi(\cdot)$  is the standard normal cdf.

## Exercises

1. (a)

$$E[\bar{Y}] = E\left[\frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)\right] = \frac{1}{4}\{E[Y_1] + E[Y_2] + E[Y_3] + E[Y_4]\} = \frac{1}{4}(4\mu) = \mu$$

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)\right) = \frac{1}{16} \sum_{i=1}^4 \text{Var}(Y_i) = \frac{1}{16} \sum_{i=1}^4 \sigma^2 = \frac{\sigma^2}{4}$$

(b)

$$\begin{aligned} E[W] &= E\left[\frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4\right] \\ &= \frac{1}{8}E[Y_1] + \frac{1}{8}E[Y_2] + \frac{1}{4}E[Y_3] + \frac{1}{2}E[Y_4] \\ &= \frac{1}{8}\mu + \frac{1}{8}\mu + \frac{1}{4}\mu + \frac{1}{2}\mu = \mu \end{aligned}$$

$$\begin{aligned}\text{Var}(W) &= \text{Var}\left(\frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4\right) \\ &= \frac{1}{64}\text{Var}(Y_1) + \frac{1}{64}\text{Var}(Y_2) + \frac{1}{16}\text{Var}(Y_3) + \frac{1}{4}\text{Var}(Y_4) = \frac{11}{32}\sigma^2\end{aligned}$$

(c) I prefer  $\bar{Y}$  over  $W$  because  $\bar{Y}$  is a more efficient unbiased estimator.

*Note:* Parts (a and b) make use of the fact that the sample is iid and thus the variance of the sums of the variables is equal to the sum of the variances of the variables.

2. (a)  $\sum_{i=1}^n a_i = 1$

(b)

$$\text{Var}(W_a) = \text{Var}\left(\sum_{i=1}^n a_i Y_i\right) = \sum_{i=1}^n \text{Var}(a_i Y_i) = a_1^2 \sigma^2 + \cdots + a_n^2 \sigma^2 = \sigma^2 \sum_{i=1}^n a_i^2$$

$$\min_{a_1, \dots, a_n} \text{Var}(W_a) = a_1^2 \sigma^2 + \cdots + a_n^2 \sigma^2 \text{ s.t. } a_1 + \cdots + a_n = 1$$

$$\mathcal{L} = a_1^2 \sigma^2 + \cdots + a_n^2 \sigma^2 + \lambda(1 - a_1 - \cdots - a_n)$$

$$\frac{\partial \mathcal{L}}{\partial a_1} = 2a_1 \sigma^2 - \lambda = 0$$

...

$$\frac{\partial \mathcal{L}}{\partial a_n} = 2a_n \sigma^2 - \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - a_1 - \cdots - a_n = 0$$

$$\rightarrow a_1 = \cdots = a_n \text{ and } \sum_{i=1}^n a_i = 1$$

$$\rightarrow \sum_{i=1}^n a_i = 1$$

$$\rightarrow a_1^* = \cdots = a_n^* = \frac{1}{n}$$

3. (a)  $E[W_1] = E\left[\frac{n-1}{n}(\bar{Y})\right] = \frac{n-1}{n}E[\bar{Y}] = \frac{n-1}{n}(\mu) \neq \mu$

$$\text{bias}(W_1) = E[W_1] - \mu = \frac{n-1}{n}(\mu) - \mu = \frac{-\mu}{n}$$

$$E[W_2] = E\left[\frac{\bar{Y}}{2}\right] = \frac{1}{2}E[\bar{Y}] = \frac{\mu}{2}$$

$$\text{bias}(W_2) = E[W_2] - \mu = \frac{\mu}{2} - \mu = \frac{-\mu}{2}$$

One important difference is that the bias of  $W_1$  converges to 0 as  $n \rightarrow \infty$ , while the bias of  $W_2$  is constant.

(b)  $\text{plim}(W_1) = \text{plim}\left(\frac{n-1}{n}(\bar{Y})\right) = \text{plim}\left(\frac{n-1}{n}\right) \text{plim}(\bar{Y}) = 1 \cdot \mu = \mu$

$$\text{plim}(W_2) = \text{plim}\left(\frac{\bar{Y}}{2}\right) = \text{plim}(\bar{Y})/\text{plim}(2) = \frac{\mu}{2}$$

$W_1$  is consistent.

(c)

$$\begin{aligned}\text{Var}(W_1) &= \text{Var}\left(\frac{n-1}{n}(\bar{Y})\right) = \left(\frac{n-1}{n}\right)^2 \text{Var}(\bar{Y}) = \left(\frac{(n-1)^2\sigma^2}{n^3}\right) \\ \text{Var}(W_2) &= \text{Var}\left(\frac{\bar{Y}}{2}\right) = \frac{1}{4}\text{Var}(\bar{Y}) = \frac{\sigma^2}{4n}\end{aligned}$$

(d) While  $\bar{Y}$  is unbiased regardless of the value of  $\mu$ , when  $\mu$  is "close" to zero, the bias of  $W_1$  is also close to zero (especially for large samples). Thus, it may be worthwhile to consider  $W_1$  over  $\bar{Y}$  if  $W_1$  is efficient relative to  $\bar{Y}$ . We know  $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$  and  $\text{Var}(W_1) = \left(\frac{(n-1)^2\sigma^2}{n^3}\right)$ . Using these calculations to evaluate when  $W_1$  is efficient relative to  $\bar{Y}$ :

$$\begin{aligned}\left(\frac{(n-1)^2\sigma^2}{n^3}\right) &\leq \frac{\sigma^2}{n} \\ \left(\frac{n-1}{n}\right)^2 &\leq 1\end{aligned}$$

which holds for all positive values of  $n$ . Thus,  $W_1$  is efficient relative to  $\bar{Y}$  and, given the small amount of bias, it may be a better estimator of  $\mu$ .

4. (a)  $E[Z] = E[E[Z|X]] = E\left[E\left[\frac{Y}{\bar{X}}|X\right]\right] = E\left[\frac{1}{\bar{X}}E[Y|X]\right] = E\left[\frac{1}{\bar{X}}\theta X\right] = E[\theta] = \theta$
- (b)  $E[W_1] = E\left[n^{-1}\sum_{i=1}^n(Y_i/X_i)\right] = n^{-1}\sum_{i=1}^n E[(Y_i/X_i)] = n^{-1}\sum_{i=1}^n \theta = n^{-1}(n\theta) = \theta$
- (c) In general, the average of the ratios,  $Y_i/X_i$ , is not the ratio of the averages  $\bar{Y}/\bar{X}$ .

$$\begin{aligned}E[W_2|X_1, \dots, X_n] &= E\left[\frac{\bar{Y}}{\bar{X}}|X_1, \dots, X_n\right] \\ &= \frac{1}{\bar{X}}E[\bar{Y}|X_1, \dots, X_n] \\ &= \frac{1}{\bar{X}}E\left[n^{-1}\sum_{i=1}^n Y_i|X_1, \dots, X_n\right] \\ &= \frac{1}{n\bar{X}}\sum_{i=1}^n E[Y_i|X_1, \dots, X_n] \\ &= \frac{1}{n\bar{X}}(n\theta\bar{X}) = \theta\end{aligned}$$

(d)  $W_1 = n^{-1}\sum_{i=1}^n(Y_i/X_i) = 0.4179674$

$$W_2 = \frac{\bar{Y}}{\bar{X}} = 0.4180968$$

Yes, they are similar.

5. (a)  $G$  is not an unbiased estimator of  $\gamma$  because  $G$  has a nonlinear relationship with  $\bar{Y}$ . As we concluded in Math Review B, the expected value of the ratio is not the ratio of the expected value.
- (b)  $\text{plim}(G) = \text{plim}\left(\frac{\bar{Y}}{1-\bar{Y}}\right) = \text{plim}(\bar{Y})/\text{plim}(1-\bar{Y}) = \theta/(\text{plim}(1) - \text{plim}(\bar{Y})) = \frac{\theta}{1-\theta} = \gamma$

6. (a)

$$H_0 : \mu = 0$$

(b)

$$H_1 : \mu < 0$$

$$(c) \ t = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{-32.8 - 0}{466.4/\sqrt{900}} \approx -2.109777$$

$$p \approx 0.0175767$$

We reject the null hypothesis at the 5% level but fail to reject  $H_0$  at the 1% level.

- (d) We've already shown there is a statistically significant difference at the 5% level but not at the 1% level. On the other hand, I would struggle to argue there is a practical significance when there is only a 32.8 ounce difference in alcohol consumption over an entire year.
- (e) This analysis implicitly assumes all other factors that affect liquor consumption have remained the same. Factors such as income, or changes in price due to transportation costs, are assumed constant over the two years.

7. (a)  $CI_{95} = [-0.0096847, 0.4896847]$

(b)

$$H_0 : \mu = 0$$

$$H_1 : \mu > 0$$

$$(c) \ t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = 2.0615954$$

For a one-sided test, we would reject  $H_0$  at the 5% level but fail to reject  $H_0$  at the 1% level.

(d)  $p = 0.0291614$

8. (a)  $\bar{Y} = \frac{188}{429} \approx 0.4382284$

(b)

$$sd(\bar{Y}) = \frac{\sigma_\theta}{\sqrt{n}} = \sqrt{\frac{\theta(1-\theta)}{n}}$$

$$(c) \ t = \frac{\bar{Y} - .5}{se(\bar{Y})} = \frac{\bar{Y} - .5}{\sqrt{\bar{Y}(1-\bar{Y})/n}} = -2.5786184$$

$$p = 0.0051263$$

Thus, we reject  $H_0$  at the 1% level.

9. (a)  $E[X] = 200(.65) = 130$

(b)  $sd(X) = \sqrt{|200| \times .65(1 - .65)} = 6.7453688$

$$(c) \ t = \frac{(115/200) - .65}{\sqrt{(.65)(.35)/200}} = -2.2237479$$

$$p = 0.0136449$$

- (d) The value calculated in part(c) is a  $p$ -value which is the probability of rejecting a true null hypothesis. In the previous part, we would reject the dictator's claim at the 5

10.  $CI_{95} = \left[ .394 - 1.96\sqrt{(.394)(1 - .394)/419}, .394 + 1.96\sqrt{(.394)(1 - .394)/419} \right] = [0.3472121, 0.4407879]$

Based on his average up to the strike, there is not very strong evidence against  $\theta = .400$ , as this value is well within the 95% confidence interval.

11.  $t = \frac{\bar{y} - 0}{s/\sqrt{n}} = \frac{.132 - 0}{1.27/20} \approx 2.0787402$

The difference is statistically greater than zero at the 5% level but not at the 1% level.

# Chapter 1

## Notes

### Data Types

**Nonexperimental data** are not accumulated through controlled experiments on individuals, firms, or segments of the economy. Non-experimental data are sometimes called **observational data**, or **retrospective data**, to emphasize the fact that the researcher is a passive collector of the data.

**Experimental data** are often collected in laboratory environments in the natural sciences, but they are more difficult to obtain in the social sciences. Although some social experiments can be devised, it is often impossible, prohibitively expensive, or morally repugnant to conduct the kinds of controlled experiments that would be needed to address economic issues.

An **empirical analysis** uses data to test a theory or to estimate a relationship.

A **cross-sectional data set** consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time. An important feature of cross-sectional data is that we can often assume that they have been obtained by **random sampling** from the underlying population. Sometimes, however, the random sampling assumption is not appropriate for a variety of reasons such as respondents' willingness to answer or sampling from units that are large relative to the population. Nevertheless, random sampling is often assumed with cross-sectional data. The analysis of cross-sectional data is closely aligned with the applied microeconomics fields, such as labor economics, state and local public finance, industrial organization, urban economics, demography, and health economics.

A **time series data set** consists of observations on a variable or several variables over time. Unlike the arrangement of cross-sectional data, the chronological ordering of observations in a time series conveys potentially important information. A key feature of time series data that makes them more difficult to analyze than cross-sectional data is that economic observations can rarely, if ever, be assumed to be independent across time.

A **pooled cross section** is a data configuration where independent cross sections, usually collected at different points in time, are combined to produce a single data set.

A **panel data** (or **longitudinal data**) **set** consists of a time series for each cross-sectional member in the data set. The key feature of panel data that distinguishes them from a pooled cross section is that the *same* cross-sectional units are followed over a given time period.

### Causality

In econometrics, we're often concerned about finding a **causal effect**, or A **ceteris paribus** (meaning all other relevant factors are held fixed) change in one variable that has an effect on another variable.

## Exercises

### Problems

1. (a) I would randomly assign (that is without other factors that affect student performance in mind) fourth grade students to varying class sizes and compare students' performances across the various groups.  
(b) I might expect a negative correlation between class size and test score because generally, larger classes have less funding per student and students in larger classes receive less individualized instruction. There are many additional factors positively correlated with student performance and negatively correlated with class size.  
(c) No, causality can only be established when ceteris-paribus is satisfied; however, this is not the case as some of the confounding factors that wouldn't be controlled for are listed in part(b).

2.
  - (a) All else equal, do job training programs improve worker productivity?
  - (b) No. For one, perhaps a firm that requires job training because it has less-skilled workers and wants to increase its production efficiency. There are many other factors like this on the firm side that make me believe that a firm's decision to train its workers will be independent of worker characteristics. Additionally, perhaps the firm does not require but offers job training. It's likely individuals who actually do the job training have different characteristics to those who do not. Some factors may include innate ability, intelligence, and motivation.
  - (c) The quality of the equipment.
  - (d) No, causality can only be established when ceteris-paribus is satisfied; however, this is not the case as some of the confounding factors that wouldn't be controlled for (such as the quality of the equipment).
3. No. Again, ceteris-paribus has not been established. Many other confounding factors correlated with "work" and "study" would not be controlled.
4.
  - (a) Ideally, panel data that contains corporate tax rates and GSP.
  - (b) Theoretically, it'd be possible to do a controlled experiment, but it would not be ethical. It would require randomly assigning individuals to varying levels of corporate tax rates and measuring the GSP within these groups.
  - (c) If other factors impacting GSP growth and tax rates are sufficiently controlled for, then yes. Otherwise, such correlational analysis will likely be biased and not convincing.

## Computer Exercises

C1)

```
#i
mean(wage1$educ)

## [1] 12.56274

min(wage1$educ)

## [1] 0

max(wage1$educ)

## [1] 18

#ii
mean(wage1$wage)

## [1] 5.896103

#It seems low for the year 2013

#iii
#Consumer Price Index (CPI) for the years 1976 and 2013.
cpi_1976 <- 55.6
cpi_2013 <- 230.280

#iv
mean(wage1$wage)*cpi_2013/cpi_1976

## [1] 24.42005
```

```
#Yes, the average wage now seems more reasonable
```

```
#v
```

```
#women
```

```
sum(wage1$female)
```

```
## [1] 252
```

```
#men
```

```
sum(wage1$female==0)
```

```
## [1] 274
```

```
rm(cpi_1976, cpi_2008)
```

```
## Warning in rm(cpi_1976, cpi_2008): object 'cpi_2008' not found
```

```
gc()
```

```
##          used (Mb) gc trigger  (Mb) max used   (Mb)
```

```
## Ncells 1102320 58.9   2291854 122.4  2291854 122.4
```

```
## Vcells 1951547 14.9   8388608  64.0   2699321  20.6
```

```
C2)
```

```
#i
```

```
#number of women
```

```
sum(bwght$male==0)
```

```
## [1] 665
```

```
#number of women who report smoking during pregnancy
```

```
sum(bwght$male==0 & bwght$cigs>0)
```

```
## [1] 112
```

```
#ii and iii
```

```
mean(bwght$cigs)
```

```
## [1] 2.087176
```

```
#No, this average includes males, a more descriptive average may be:
```

```
mean(bwght$cigs[bwght$male==0])
```

```
## [1] 2.090226
```

```
#for all women and
```

```
mean(bwght$cigs[bwght$male==0 & bwght$cigs>0])
```

```
## [1] 12.41071
```

```
#for those that reported smoking
```

```
#iv
```

```
mean(bwght$fatheduc, na.rm = T)
```

```
## [1] 13.18624
```

```
#There are only 1192 observations because there are 196 missing values for fatheduc
```

```
#v
```

```
mean(bwght$faminc) * 1000
```



```
## [1] 29026.66
sd(bwght$faminc) * 1000

## [1] 18739.28
C3)

#i
min(meap01$math4)

## [1] 0
max(meap01$math4)

## [1] 100
#Without knowing much about the data, the range seems to make sense covering all 100% of pass rates

#ii
sum(meap01$math4 == 100)

## [1] 38
#iii
sum(meap01$math4 == 50)

## [1] 17
#iv
#math
mean(meap01$math4)

## [1] 71.909
#reading
mean(meap01$read4)

## [1] 60.06188
#The reading test seems harder to pass

#v
cor(meap01$math4, meap01$read4)

## [1] 0.8427281
#Those with higher pass rates on one exam tend to have higher pass rates on the other
#In other words, pass rates on the math and reading tests are highly correlated

#vi
mean(meap01$exppp)

## [1] 5194.865
sd(meap01$exppp)

## [1] 1091.89
#vii
#actual
500*100/5500
```

```
## [1] 9.090909
```

```
#ln approx  
100 * (log(6000)-log(5500))
```

```
## [1] 8.701138
```

C4)

```
#i (reporting proportion instead of fraction)  
mean(jtrain2$train)
```

```
## [1] 0.4157303
```

```
#i  
#receiving training  
mean(jtrain2$re78[jtrain2$train==1])
```

```
## [1] 6.349145
```

```
#not receiving training  
mean(jtrain2$re78[jtrain2$train==0])
```

```
## [1] 4.554802
```

```
#The difference does appear economically large
```

```
#iii  
#receiving training  
mean(jtrain2$unem78[jtrain2$train==1])
```

```
## [1] 0.2432432
```

```
#not receiving training  
mean(jtrain2$unem78[jtrain2$train==0])
```

```
## [1] 0.3538462
```

```
#The proportion of unemployed who are not receiving training is about 11% larger
```

```
#Yes, the training seems effective, establishing ceteris paribus would make the results more convincing
```

C5)

```
#i  
#min  
min(fertil2$children)
```

```
## [1] 0
```

```
#max  
max(fertil2$children)
```

```
## [1] 13
```

```
#mean  
mean(fertil2$children)
```

```
## [1] 2.267828
```

```
#ii  
sum(fertil2$children>0) / length(fertil2$children)
```

```
## [1] 0.7404265
```

```
#iii  
#for those who have electricity  
mean(fertil2$children[fertil2$electric==1 & !is.na(fertil2$electric)])
```

```
## [1] 1.898527
```

```
#for those who do not have electricity  
mean(fertil2$children[fertil2$electric==0 & !is.na(fertil2$electric)])
```

```
## [1] 2.327729
```

```
#Those without electricity have more children on average (for this sample)
```

```
#iv  
#No, ceteris paribus is not established
```

C6)

```
county_murders <- as_tibble(countymurders) %>%  
  filter(year == 1996)
```

```
#i  
n_distinct(county_murders$countyid)
```

```
## [1] 2197
```

```
#number with 0 murders  
n_distinct(county_murders$countyid[county_murders$murders==0])
```

```
## [1] 1051
```

```
#percent with 0 murders  
100 * n_distinct(county_murders$countyid[county_murders$murders==0]) / n_distinct(county_murders$countyid)
```

```
## [1] 47.83796
```

```
#ii  
#max number of murders  
max(county_murders$murders)
```

```
## [1] 1403
```

```
#max number of executions  
max(county_murders$execs)
```

```
## [1] 3
```

```
#mean number of executions  
mean(county_murders$execs)
```

```
## [1] 0.01593081
```

```
#iii  
cor(county_murders$murders, county_murders$execs)
```

```
## [1] 0.2095042
```

```
#iv  
#No, I would suspect the positive correlation may be due to executions resulting from murders  
#and other crimes, which I suspect is correlated with the number of murders  
rm(county_murders)  
gc()
```

```
##          used (Mb) gc trigger  (Mb) max used   (Mb)
## Ncells 1128928 60.3      2291854 122.4   2291854 122.4
## Vcells 2700104 20.7      8388608  64.0   5309419  40.6
```

C7)

```
#i
#percent who report abusing alcohol
100 * mean(alcohol$abuse)
```

```
## [1] 9.916514
```

```
#employment rate
100 * mean(alcohol$employ)
```

```
## [1] 89.81877
```

```
#ii
100 * mean(alcohol$employ[alcohol$abuse==1])
```

```
## [1] 87.26899
```

```
#iii
100 * mean(alcohol$employ[alcohol$abuse==0])
```

```
## [1] 90.09946
```

```
#iv
#No, ceteris paribus is not established
```

C8)

```
#i
#assuming each obs/row corresponds to a unique student
NROW(econmath)
```

```
## [1] 856
```

```
#ii
#for those who took econ in high school
mean(econmath$score[econmath$econhs==1])
```

```
## [1] 72.07593
```

```
#for those who did not take econ in high school
mean(econmath$score[econmath$econhs==0])
```

```
## [1] 72.90792
```

```
#iii
#No, it simply tells us the average scores for those who did and did not take econ in hs
#It may signify some level of correlation but not causality

#iv
#Randomly assigning individuals to two groups (one who does take econ in high school
#and one that does not), then performing part(ii) can be used to obtain a good causal estimate
```

## Chapter 2

### Notes

A **simple linear regression model** is a model relating a dependent variable to one independent variable and generally takes the form

$$y = \beta_0 + \beta_1 x + u$$

It is also called the **two-variable linear regression model** or **bivariate linear regression model** because it relates the two variables  $x$  and  $y$ .

The variable  $u$ , called the **error term** or **disturbance** in the relationship, represents factors other than  $x$  that affect  $y$ .

If the other factors in  $u$  are held fixed, so that the change in  $u$  is zero,  $\Delta u = 0$ , then  $x$  has a linear effect on  $y$ :

$$\Delta y = \beta_1 \Delta x \text{ if } \Delta u = 0$$

As long as the intercept  $\beta_0$  is included in the equation, nothing is lost by assuming that the average value of  $u$  in the population is zero. Mathematically,

$$E[u] = 0$$

One of the most important assumptions in econometrics is the **zero conditional mean assumption**. This crucial assumption says that the average value of the unobservables ( $u$ ) is the same across all slices of the population determined by the value of  $x$  and that the common average *is* necessarily equal to the average of  $u$  over the entire population. Thus, under this assumption (and the zero unconditional mean assumption), we write

$$E[u|x] = E[u] = 0$$

Using the zero conditional mean assumption and taking the expected value of  $y$  given  $x$  from the simple linear regression model gives

$$E[y|x] = \beta_0 + \beta_1 x,$$

This is called the **population regression function (PRF)** and shows that  $E[y|x]$  is a linear function of  $x$ . The PRF gives us a relationship between the average level of  $y$  at different levels of  $x$ . It is important to understand that the PRF tells us how the average value of  $y$  changes with  $x$ ; it does not say that  $y$  equals  $\beta_0 + \beta_1 x$  for all units in the population.

### OLS Estimators Derivation (SLR)

Let  $\{(x_i, y_i) : i = 1, \dots, n\}$  denote a random sample of size  $n$  from the population. Then, we can write

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

for each  $i$  where  $u_i$  is the error term for each observation  $i$  because it contains all factors affecting  $y_i$  other than  $x_i$ . Because there are two unknown parameters to estimate  $(\beta_0, \beta_1)$ , we might hope to obtain good estimators, which we will denote  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . Thus, we may want to find estimators that minimize the sum of squared residuals

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

where  $\hat{y}_i$  represents fitted values from an estimated model using  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . Formally,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Hence, we can rewrite

$$SSR = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

To minimize the sum of squared residuals, we take two first order conditions:

FOC 1:

$$\begin{aligned} \frac{\partial SSR}{\partial \beta_0} &= \sum_{i=1}^n -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ &\rightarrow \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ &\rightarrow n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} = 0 \\ &\rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ &\rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

*Note:* The significance of this FOC is that, using least squares criteria, our line of best fit runs through the sample means  $\bar{y}$  and  $\bar{x}$ .

FOC 2:

$$\begin{aligned} \frac{\partial SSR}{\partial \beta_1} &= \sum_{i=1}^n -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ &\rightarrow \sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ &\rightarrow \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0 \\ &\rightarrow \sum_{i=1}^n x_i y_i = \sum_{i=1}^n (\hat{\beta}_0 x_i + \hat{\beta}_1 x_i^2) \\ &\rightarrow \sum_{i=1}^n x_i y_i = \sum_{i=1}^n ((\bar{y} - \hat{\beta}_1 \bar{x})x_i + \hat{\beta}_1 x_i^2) \end{aligned}$$

$$\begin{aligned}
& \rightarrow \sum_{i=1}^n x_i y_i = \sum_{i=1}^n \left( \bar{y} x_i - \hat{\beta}_1 \bar{x} x_i + \hat{\beta}_1 x_i^2 \right) \\
& \rightarrow \sum_{i=1}^n x_i y_i = n \bar{x} \bar{y} + \hat{\beta}_1 \sum_{i=1}^n (x_i^2 - \bar{x} x_i) \\
& \rightarrow \sum_{i=1}^n (x_i y_i) - n \bar{x} \bar{y} = \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}) \\
& \rightarrow \sum_{i=1}^n (x_i y_i - \bar{x} y_i) = \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}) \\
& \rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \hat{\rho}_{xy} \cdot \frac{\hat{\sigma}_y}{\hat{\sigma}_x}
\end{aligned}$$

Note that the FOCs are analogous to the zero mean assumptions (using residuals instead of the error term):

FOC 1 can be restated as

$$E[\hat{u}] = 0 \rightarrow E[y - \hat{y}] = E[y - \hat{\beta}_0 - \hat{\beta}_1 x] = n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

And FOC 2 can be restated as

$$\begin{aligned}
E[\hat{u}|x] = E[\hat{u}] \rightarrow \hat{u} \perp x \rightarrow \text{Cov}(x, \hat{u}) = 0 \rightarrow E[x\hat{u}] - E[x]E[\hat{u}] = E[x\hat{u}] = 0 \\
\rightarrow n^{-1} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0
\end{aligned}$$

## OLS Regression

The estimators  $(\hat{\beta}_0, \hat{\beta}_1)$  are called the **ordinary least squares (OLS)** estimators of  $\beta_0$  and  $\beta_1$ . A **fitted value** for  $y$  when  $x = x_i$  is defined as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The **residual** for observation  $i$  is the difference between the actual  $y_i$  and its fitted value:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Thus, we defined the **OLS regression line** or **sample regression function (SRF)** as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

## Properties of OLS Statistics

1. The sum, and therefore the sample average of the OLS residuals, is zero, such that

$$\sum_{i=1}^n \hat{u}_i = 0$$

This result follows from FOC 1.

2. The sample covariance between the regressors and the OLS residuals is zero. This follows from FOC2, which can be written in terms of the residuals as

$$\sum_{i=1}^n x_i \hat{u}_i = 0$$

3. The point  $(\bar{x}, \bar{y})$  is always on the OLS regression line. In other words, if we take the OLS regression line and plug in  $\bar{x}$  for  $x$ , then the predicted value is  $\bar{y}$ . This is exactly what the derivation from FOC 1 showed us.

## Sum of Squares

We can view OLS as decomposing each  $y_i$  into two parts, a fitted value and a residual. The fitted values and residuals are uncorrelated in the sample.

The **total sum of squares (SST)** can be defined as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

and is a measure of the variation in the  $y_i$  of the given sample. If we divided SST by  $n - 1$ , we obtain the sample variance of  $y$ .

The **explained sum of squares (SSE)** can be defined as

$$SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

and is a measure of the variation in the  $\hat{y}_i$ .

The **residual sum of squares (SSR)** can be defined as

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{u}_i^2$$

and measures the sample variation in the  $\hat{u}_i$ .

In total,

$$SST = SSE + SSR$$

## Proof

$$\begin{aligned} SSE + SSR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (\hat{y}_i^2 - 2\bar{y}\hat{y}_i + \bar{y}^2) + \sum_{i=1}^n (y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2) \\ &= \left[ \sum_{i=1}^n \hat{y}_i^2 - 2\bar{y} \sum_{i=1}^n \hat{y}_i + \sum_{i=1}^n \bar{y}^2 \right] + \left[ \sum_{i=1}^n y_i^2 - 2 \sum_{i=1}^n y_i \hat{y}_i + \sum_{i=1}^n \hat{y}_i^2 \right] \\ &= \left[ \sum_{i=1}^n y_i^2 - 2n\bar{y}^2 + \sum_{i=1}^n \bar{y}^2 \right] + \left[ \sum_{i=1}^n \hat{y}_i^2 - 2 \sum_{i=1}^n y_i \hat{y}_i + \sum_{i=1}^n \hat{y}_i^2 \right] \\ &= \left[ \sum_{i=1}^n y_i^2 - 2\bar{y} \sum_{i=1}^n y_i + \sum_{i=1}^n \bar{y}^2 \right] + \left[ \sum_{i=1}^n (2\hat{y}_i^2 - 2y_i\hat{y}_i) \right] \\ &= \left[ \sum_{i=1}^n (y_i^2 - 2\bar{y}y_i + \bar{y}^2) \right] + \left[ \sum_{i=1}^n 2\hat{y}_i(\hat{y}_i - y_i) \right] \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2 \sum_{i=1}^n \hat{y}_i(\hat{u}_i) \\ &= SST - 2 \sum_{i=1}^n \hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= SST - 2 \left[ \hat{\beta}_0 \sum_{i=1}^n \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^n x_i \hat{u}_i \right] = SST \end{aligned}$$



The **R-squared** of the regression, sometimes called the **coefficient of determination**, is defined as

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

$R^2$  is the ratio of the explained variation compared to the total variation; thus, it is interpreted as the fraction of the sample variation in  $y$  that is explained by  $x$ .  $R^2$  is equal to the square of the sample correlation coefficient between  $y_i$  and  $\hat{y}_i$ :

$$R^2 = \frac{[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}$$

An important fact about  $R^2$  is that it never decreases, and it usually increases, when another independent variable is added to a regression and the same set of observations is used for both regressions. The fact that  $R^2$  never decreases when any variable is added to a regression makes it a poor tool for deciding whether particular variables should be added to a model; instead, one should focus on whether an explanatory variable has a nonzero partial effect on  $y$  in the *population* to decide whether to include an explanatory variable in a model.

## Nonlinear Models

In a linear regression model, changing the units of the variables has a multiplicative effect on the slope and intercepts but has not effect on the  $R^2$ .

In a nonlinear regression model that incorporates natural logarithms, changing the units has no effect on the slope but does change the intercept. Take for example,  $\ln(y) = \beta_0 + \beta_1 \ln(x) + u$ . If we multiply  $y$  by some constant  $c_1$  and  $x$  by some constant  $c_2$ , the equation becomes

$$\begin{aligned} \ln(c_1 y) &= \beta_0 + \beta_1 \ln(c_2 x) + u \\ \rightarrow \ln(c_1) + \ln(y) &= \beta_0 + \beta_1 [\ln(c_2) + \ln(x)] + u \\ \rightarrow \ln(y) &= \alpha + \beta_1 \ln(x) + u \end{aligned}$$

where  $\alpha = \beta_0 - \ln(c_1) + \beta_1 \ln(c_2)$ .

## Summary of Functional Forms Involving Natural Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of $\beta_1$
Level-level (linear model)	$y$	$x$	$\Delta y = \beta_1 \Delta x$
Level-log	$y$	$\ln(x)$	$\Delta y = (\beta_1/100)\% \Delta x$
Log-level (semi-elasticity model)	$\ln(y)$	$x$	$\% \Delta y = (100\beta_1) \Delta x$
Log-log (constant elasticity model)	$\ln(y)$	$\ln(x)$	$\% \Delta y = \beta_1 \% \Delta x$

## Gauss-Markov Assumptions

- Assumption 1: The population model is linear in parameters such that

$$y_i = \alpha + \beta_1 x_{i,1} + \dots + \beta_m x_{i,m} + u_i$$

- Assumption 2: The sample data  $\{x_{i,1}, \dots, x_{i,m}, y_i\}$  is a random sample.
- Assumption 3: The error term has a zero conditional mean such that

$$E[u|x_{i,1}, \dots, x_{i,m}] = 0$$

For a random sample, this assumption implies that  $E[u_i|x_i] = 0 \forall i = 1, 2, \dots, n$

- Assumption 4: The error term is homoskedastic such that

$$Var(u|x_{i,1}, \dots, x_{i,m}) = \sigma_u^2$$

- Assumption 5: None of the regressors exhibit perfect collinearity with one another.
- Assumption 6: There exists no serial correlation between the error terms such that

$$\text{cov}(u_i, u_j) = 0 \quad \forall i \neq j$$

Note: Assumption 2 is sufficient to satisfy assumption 6 in the case of cross-sectional data (see below for proof).

Under this set of assumptions, the **Gauss-Markov Theorem** states that the OLS estimator is BLUE (conditional on the sample values of the explanatory variable(s)).

Proof that Random Sampling Implies Zero Serial Correlation with Cross-Sectional Data Under assumption 2, we know the sample data is independent and identically distributed (i.i.d), meaning drawing one observation does not make drawing another observation any more or less likely and that the observations come from the same distribution. Thus, for some observation  $i \neq j$ ,

$$\text{cov}(y_i, y_j) = E[y_i y_j] - E[y_i]E[y_j] = E[y_i]E[y_j] - E[y_i]E[y_j] = 0$$

Similarly, by our first assumption that  $y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{i,m} x_{i,m} + u_i$

$$\begin{aligned} 0 &= \text{cov}(y_i, y_j) = \text{cov}(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{i,m} x_{i,m} + u_i, \beta_0 + \beta_1 x_{j,1} + \dots + \beta_{j,m} x_{j,m} + u_j) \\ &= \sum_{s=1}^m \sum_{t=1}^m \beta_s \beta_t \text{cov}(x_{i,s}, x_{j,t}) + \sum_{s=1}^m \beta_s \text{cov}(x_{i,s}, u_j) + \sum_{t=1}^m \beta_t \text{cov}(x_{j,t}, u_i) + \text{cov}(u_i, u_j) \end{aligned}$$

Using the fact that the sample is i.i.d.,

$$\sum_{s=1}^m \sum_{t=1}^m \beta_s \beta_t \text{cov}(x_{i,s}, x_{j,t}) = 0$$

Also, by the exogeneity assumption,

$$\sum_{s=1}^m \beta_s \text{cov}(x_{i,s}, u_j) + \sum_{t=1}^m \beta_t \text{cov}(x_{j,t}, u_i) = 0$$

Leaving us with

$$\text{cov}(u_i, u_j) = 0$$

### Unbiasedness of OLS Estimators (SLR)

$$\begin{aligned} E[\hat{\beta}_1 | x] &= E \left[ \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n x_i (x_i - \bar{x})} \middle| x \right] \\ &= E \left[ \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + u_i) (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \middle| x \right] \\ &= E \left[ \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n x_i (x_i - \bar{x}) + \sum_{i=1}^n u_i (x_i - \bar{x})}{SST_x} \middle| x \right] \\ &= E \left[ \frac{\beta_0 (n\bar{x} - n\bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n u_i (x_i - \bar{x})}{SST_x} \middle| x \right] \\ &= E \left[ \beta_1 + \frac{\sum_{i=1}^n u_i (x_i - \bar{x})}{SST_x} \middle| x \right] \\ &= \beta_1 + \frac{E[\sum_{i=1}^n u_i (x_i - \bar{x}) | x]}{SST_x} \\ &= \beta_1 + \frac{\sum_{i=1}^n E[u_i (x_i - \bar{x}) | x]}{SST_x} \end{aligned}$$

$$= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E[u_i|x]}{SST_x} = \beta_1$$

where we have used the fact that the expected value of each  $u_i$  (conditional on  $\{x_1, x_2, \dots, x_n\}$ ) is zero under the second and third Gauss-Markov assumptions. Because unbiasedness holds for any outcome on  $\{x_1, x_2, \dots, x_n\}$ , unbiasedness also holds without conditioning on  $\{x_1, x_2, \dots, x_n\}$ .

$$\begin{aligned} E[\hat{\beta}_0|x] &= E[\bar{y} - \hat{\beta}_1 \bar{x}|x] \\ &= E[\bar{y}|x] - E[\hat{\beta}_1 \bar{x}|x] \\ &= E\left[n^{-1} \sum_{i=1}^n (\beta_0 + \beta_1 x_i + u_i) \middle| x\right] - \bar{x} E[\hat{\beta}_1|x] \\ &= E[\beta_0 + \beta_1 \bar{x}|x] - \beta_1 \bar{x} \\ &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \end{aligned}$$

Unbiasedness generally fails if any of our first three assumptions fail. This means that it is important to think about the veracity of each assumption for a particular application. The first assumption requires that  $y$  and  $x$  be linearly related, with an additive disturbance. This can certainly fail. But we also know that  $y$  and  $x$  can be chosen to yield interesting nonlinear relationships. Random sampling can fail in a cross section when samples are not representative of the underlying population; in fact, some data sets are constructed by intentionally oversampling different parts of the population. Finally, the third assumption is possibly the most important of the Gauss-Markov assumptions. Using simple regression when  $u$  contains factors affecting  $y$  that are also correlated with  $x$  can result in spurious correlation: that is, we find a relationship between  $y$  and  $x$  that is really due to other unobserved factors that affect  $y$  and also happen to be correlated with  $x$ . In addition to omitted variables, there are other reasons for  $x$  to be correlated with  $u$  in the simple regression model.

### Sampling Variance of OLS Estimators (SLR)

Using a previous derivation,

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x}$$

Thus,

$$\begin{aligned} \text{Var}(\hat{\beta}_1|x) &= \text{Var}\left(\beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x} \middle| x\right) \\ &= \frac{1}{SST_x^2} \text{Var}\left(\sum_{i=1}^n u_i(x_i - \bar{x}) \middle| x\right) \\ &= \frac{1}{SST_x^2} \sum_{i=1}^n \text{Var}(u_i(x_i - \bar{x})|x) \\ &= \frac{1}{SST_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(u_i|x) \\ &= \frac{1}{SST_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma_u^2 \\ &= \frac{\sigma_u^2}{SST_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

$$= \frac{\sigma_u^2}{SST_x}$$

Again, using a previous derivation,

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&\rightarrow \text{Var}(\hat{\beta}_0 | x) = \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x} | x) \\
&= \text{Var}(\bar{y} | x) + \text{Var}(\hat{\beta}_1 \bar{x} | x) - 2\text{Cov}(\bar{y}, \hat{\beta}_1 \bar{x} | x) \\
&= \text{Var}(\beta_0 + \beta_1 \bar{x} + \bar{u} | x) + \bar{x}^2 \text{Var}(\hat{\beta}_1 | x) - \frac{2\bar{x}}{n} \text{Cov}\left(\sum_{i=1}^n y_i, \hat{\beta}_1 \middle| x\right) \\
&= \text{Var}(\beta_1 \bar{x} | x) + \text{Var}(\bar{u} | x) + 2\text{Cov}(\beta_1 \bar{x}, \bar{u} | x) + \bar{x}^2 \frac{\sigma_u^2}{SST_x} - \frac{2\bar{x}}{n} \text{Cov}\left(\sum_{i=1}^n y_i, \beta_1 + \frac{\sum_{i=1}^n u_i (x_i - \bar{x})}{SST_x} \middle| x\right) \\
&= 0 + \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n u_i \middle| x\right) + 2 \cdot 0 + \frac{\sigma_u^2 \bar{x}^2}{SST_x} - \frac{2\bar{x}}{n SST_x} \text{Cov}\left(\sum_{i=1}^n y_i, \sum_{i=1}^n u_i (x_i - \bar{x}) \middle| x\right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(u_i | x) + \frac{\sigma_u^2 \bar{x}^2}{SST_x} - \frac{2\bar{x} \sum_{i=1}^n (x_i - \bar{x})}{n SST_x} \text{Cov}\left(\sum_{i=1}^n y_i, \sum_{i=1}^n u_i \middle| x\right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \sigma_u^2 + \frac{\sigma_u^2 \bar{x}^2}{SST_x} - \frac{2\bar{x}(0)}{n SST_x} \text{Cov}\left(\sum_{i=1}^n y_i, \sum_{i=1}^n u_i \middle| x\right) \\
&= \frac{\sigma_u^2}{n} + \frac{\sigma_u^2 \bar{x}^2}{SST_x} \\
&= \frac{SST_x \sigma_u^2 + n \sigma_u^2 \bar{x}^2}{n SST_x} \\
&= \frac{\sigma_u^2 (SST_x + n \bar{x}^2)}{n SST_x} \\
&= \frac{\sigma_u^2 (\sum_{i=1}^n (x_i - \bar{x})^2 + n \bar{x}^2)}{n SST_x} \\
&= \frac{\sigma_u^2 (\sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2\bar{x}x_i) + n \bar{x}^2)}{n SST_x} \\
&= \frac{\sigma_u^2 (\sum_{i=1}^n (x_i^2) + n \bar{x}^2 - 2n \bar{x}^2 + n \bar{x}^2)}{n SST_x} \\
&= \frac{\sigma_u^2 / n \sum_{i=1}^n x_i^2}{SST_x}
\end{aligned}$$

Generally, we are interested in  $\text{Var}(\hat{\beta}_1)$ . To summarize how this variance depends on the error variance,  $\sigma_u^2$ , and the total variation in  $\{x_1, x_2, \dots, x_n\}$   $SST_x$ :

1. The larger the error variance, the larger is  $\text{Var}(\hat{\beta}_1)$ . This makes sense because more variation in the unobservables affecting  $y$  makes it more difficult to precisely estimate  $\beta_1$ .
2. More variability in the independent variable is preferred: as the variability in the  $x_i$  increases, the variance of  $\hat{\beta}_1$  decreases. This also makes intuitive sense because the more spread out the sample of independent variables is, the easier it is to trace out the relationship between  $E[y|x]$  and  $x$  (i.e., it becomes easier to estimate  $\beta_1$ ). If there is little variation in the  $x_i$ , then it can be hard to pinpoint how  $E[y|x]$  varies with  $x$ . As the sample size increases, so does the total variation in the  $x_i$ . Therefore, a larger sample size results in a smaller variance for  $\hat{\beta}_1$ .

These formulas allow us to isolate the factors that contribute to  $\text{Var}(\hat{\beta}_0)$  and  $\text{Var}(\hat{\beta}_1)$ . But these formulas are unknown, except in the extremely rare case that  $\sigma_u^2$  is known. Considering the  $u_i$  are unobserved, an unbiased estimator of  $\sigma_u^2$  is

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-2},$$

where the  $n-2$  is the degrees of freedom in the OLS residuals due to the two OLS first order conditions:

$$\sum_{i=1}^n \hat{u}_i = 0, \sum_{i=1}^n x_i \hat{u}_i = 0$$

Naturally, the estimators of  $\text{Var}(\hat{\beta}_0)$  and  $\text{Var}(\hat{\beta}_1)$  become

$$\hat{\sigma}_{\beta_0}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2 / n(n-2) + \sum_{i=1}^n x_i^2}{SST_x}$$

and

$$\hat{\sigma}_{\beta_1}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2 / (n-2)}{SST_x}$$

Additionally, the natural estimator of  $\sigma_u$  is

$$\hat{\sigma}_u = \sqrt{\hat{\sigma}_u^2},$$

which is called the **standard error of the regression (SER)**.

Although  $\hat{\sigma}_u^2$  is an unbiased and consistent estimator of  $\sigma_u^2$ ,  $\hat{\sigma}_u$  is consistent but biased.

All together the **standard errors of the coefficients** are

$$\text{se}(\hat{\beta}_0) = \sqrt{\frac{\sum_{i=1}^n \hat{u}_i^2 / n(n-2) + \sum_{i=1}^n x_i^2}{SST_x}}$$

and

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\sum_{i=1}^n \hat{u}_i^2 / (n-2)}{SST_x}}$$

#### A Note on the Conditioning on $x$

In addition to restricting the relationship between  $u$  and  $x$  in the population, the zero conditional mean assumption—coupled with the random sampling assumption—allows for a convenient technical simplification. In particular, we can derive the statistical properties of the OLS estimators as conditional on the values of the  $x_i$  in our sample. Technically, in statistical derivations, conditioning on the sample values of the independent variable is the same as treating the  $x_i$  as fixed in repeated samples, which we think of as follows. We first choose  $n$  sample values for  $\{x_1, x_2, \dots, x_n\}$ . Given these values, we then obtain a sample on  $y$ . Next, another sample of  $y$  is obtained, using the same values for  $\{x_1, x_2, \dots, x_n\}$ . Then another sample of  $y$  is obtained, again using the same  $\{x_1, x_2, \dots, x_n\}$ . And so on. While the fixed-in-repeated-samples scenario is not very realistic in non-experimental contexts, random sampling, where individuals are chosen randomly, is representative of how most data sets are obtained for empirical analysis in social sciences. Once we assume that  $E[u|x] = 0$ , and we have random sampling, nothing is lost in derivations by treating the  $x_i$  as nonrandom. The danger is that the fixed-in-repeated-samples assumption always implies that  $u_i$  and  $x_i$  are independent.

#### **Regression through the Origin**

In some cases, we may wish to impose the restriction that, when  $x = 0$ ,  $E[y] = 0$ . In these cases we perform **regression through the origin**, which takes the form

$$\tilde{y} = \tilde{\beta}_1 x,$$

where the tildes are used to distinguish this problem from the much more common problem of estimating an intercept along with a slope. We still rely on the method of OLS to obtain the slope estimate. Thus, we must solve for the  $\tilde{\beta}_1$  that minimizes

$$SSR = \sum_{i=1}^n (y_i - \tilde{\beta}_1 x_i)^2$$

Taking the partial derivative with respect to  $\tilde{\beta}_1$ , we obtain

$$\frac{\partial SSR}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^n x_i (y_i - \tilde{\beta}_1 x_i) = 0$$

Solving for  $\tilde{\beta}_1$ :

$$\begin{aligned} \sum_{i=1}^n x_i (y_i - \tilde{\beta}_1 x_i) &= 0 \\ \rightarrow \sum_{i=1}^n x_i y_i - \sum_{i=1}^n \tilde{\beta}_1 x_i^2 &= 0 \\ \rightarrow \tilde{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \\ \rightarrow \tilde{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \end{aligned}$$

Obtaining an estimate of  $\beta_1$  using regression through the origin is not done very often in applied work, and for good reason: if the intercept is 0, then  $\tilde{\beta}_1$  is a biased estimator of  $\beta_1$ . In cases where it is appropriate, the  $R$ -squared is computed as

$$1 - \frac{\sum_{i=1}^n (y_i - \tilde{\beta}_1 x_i)^2}{\sum_{i=1}^n y_i^2} = 1 - \frac{SSR}{SST}$$

Note: If we only regress on a constant, (i.e., we set the slope to zero and estimate an intercept only), the intercept that minimizes the sum of squared deviations is  $\bar{y}$ .

## Regression on a Binary Explanatory Variable

Simple regression can also be applied to the case where  $x$  is a **binary variable**, often called a **dummy variable** in the context of regression analysis. When the explanatory variable is binary, the PRF takes two forms (under the exogeneity assumption):

$$E[y|x] = E[\beta_0 + \beta_1 x + u|x] = \beta_0 + \beta_1 x$$

1.  $E[y|x = 0] = \beta_0$
2.  $E[y|x = 1] = \beta_0 + \beta_1$

It follows that

$$\beta_1 = E[y|x = 1] - E[y|x = 0]$$

In cases where we hope to study the effect of an intervention or new policy, the idea of counterfactuals or potential outcomes are used. Define, the **control group** as those not subject to the intervention or new policy act and the **treatment group** as those subject to the intervention. Then, the causal (or treatment) effect of the intervention for unit  $i$  is

$$te_i = y_i(1) - y_i(0),$$

the difference between the two potential outcomes. A noteworthy item about  $te_i$  is it is not observed for any unit  $i$  because it depends on both counterfactuals. We cannot hope to estimate  $te_i$  for each unit  $i$ . Instead, the focus is typically on the **average treatment effect (ATE)**, also called the **average causal effect**

(**ACE**). The ATE is simply the average of the treatment effects across the entire population. We can write the ATE parameter as

$$\tau_{ate} = E[te_i] = E[y_i(1) - y_i(0)] = E[y_i(1)] - E[y_i(0)]$$

For each unit  $i$  let  $x_i$  be the program participation status—a binary variable. Then the observed outcome,  $y_i$ , can be written as

$$y_i = (1 - x_i)y_i(0) + x_iy_i(1) = y_i(0) + [y_i(1) - y_i(0)]x_i$$

Imposing a usually unrealistic assumption of a constant treatment effect

$$\rightarrow y_i = y_i(0) + \tau x_i$$

Rewriting  $y_i(0)$  as  $\beta_0 + u_i$  and  $\tau$  as  $\beta_1$ ,

$$\rightarrow y_i = \beta_0 + \tau x_i + u_i$$

If  $x_i \perp u_i$ , then  $\tau$  is the unbiased estimator of the treatment effect. The assumption that  $x_i \perp u_i$  is the same as  $x_i \perp y_i(0)$ . This assumption can be guaranteed only under **random assignment**, whereby units are assigned to the treatment and control groups using a randomization mechanism that ignores any features of the individual units. Random assignment is the hallmark of a **randomized controlled trial (RCT)**, which is considered the gold standard for determining whether medical interventions have causal effects.

## Exercises

### Problems

1. (a) Among the many factors contained in  $u$  are marriage status, desire to have kids, health, and wealth. It's likely that at least wealth is correlated with level of education.  
 (b) No. Since it's likely  $u$  is correlated with *kids* and *educ*, a simple regression analysis will likely give a biased estimate of  $\beta_1$ .
2. Rewriting the model as  $y = \beta_0 + \beta_1x + u + \alpha_0 - \alpha_0$ , define a new error term  $e = u - \alpha_0$  and a new intercept  $\gamma_0 = \alpha_0 + \beta_0$ . The model becomes  $y = \gamma_0 + \beta_1x + e$ . It follows that  $E[e] = E[u - \alpha_0] = E[u] - E[\alpha_0] = \alpha_0 - \alpha_0 = 0$ .

3.

a)

Call:

```
lm(formula = GPA ~ ACT)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.42308	-0.14863	0.06703	0.10742	0.37912

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.56813	0.92842	0.612	0.5630
ACT	0.10220	0.03569	2.863	0.0287 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2692 on 6 degrees of freedom

Multiple R-squared: 0.5774, Adjusted R-squared: 0.507

F-statistic: 8.199 on 1 and 6 DF, p-value: 0.02868

The slope tells us there is a positive correlation between GPA and ACT. The intercept gives

us an estimate of GPA if a student scores a 0 on the ACT. GPA is predicted to be 0.510989 higher if the ACT score increases by five points.

b)

Fitted values = ( 2.7143, 3.0209, 3.2253, 3.3275, 3.5319, 3.1231, 3.1231, 3.6341 )

Residuals = ( 0.0857, 0.3791, -0.2253, 0.1725, 0.0681, -0.1231, -0.4231, 0.0659 )

Sum of residuals = -0.00000000000000002775558

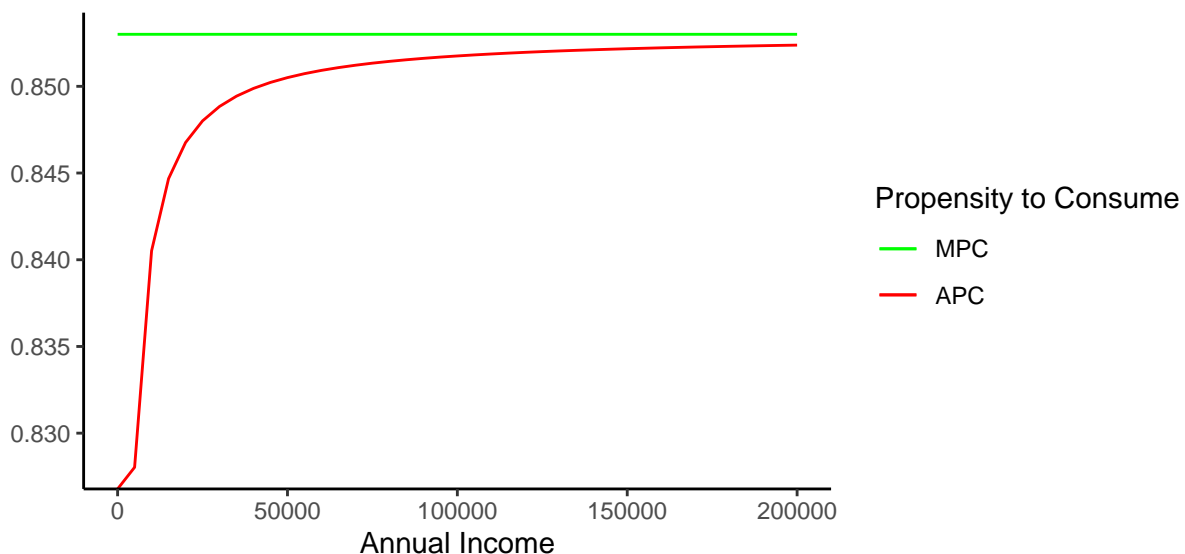
c)

Predicted value of GPA when ACT=20: 2.612088

d)

The coefficient of determination (R-squared)= 0.5774238

4. (a) The predicted birth weight when *cigs* = 0 is 119.77 ounces. When *cigs* = 20, the predicted birth weight is 109.49 ounces. This regression analysis suggests a constant loss to birth weight for every additional cigarette smoked. In this case, we would an 10.28 ounce loss in birth weight for each addition 20 cigarettes smoked.
- (b) No. Other factors correlated with *cigs* and *bwght* are not controlled for in the model. Such factors may include the health of the mother and genetics. This leads to a biased slope estimate.
- (c) Based on this model, *cigs* must equal -10.1750973 for a birth weight of 125 ounces. Obviously, this doesn't have much meaning because you can't smoke a negative number of cigarettes.
5. (a) The intercept predicts a -124.84 dollar consumption value for a family with no annual income. The slope predicts an extra 0.853 dollars of consumption for each additional dollar of annual income. The intercept offers little meaning because you can't consume a negative amount. On the other hand, the slope coefficient signifies a positive correlation between consumption and annual income. Personally, the slope coefficient seems to large to me at first glance.
- (b) The predicted consumption when family income is \$30,000 = \$ 25465.16.
- (c)





6. (a) The slope is an estimated elasticity of *price* with respect to *dist*. The estimate indicates a 0.312 percentage point increase in *price* for every percent increase in *dist*. The sign is as we'd expect. As houses move further away from the garbage incinerator, we'd expect housing prices to increase.
- (b) No. As the problem suggests, it's likely the city put the incinerator close to cheaper homes. Thus, the elasticity estimate is likely upwardly biased and overstates the impact of *dist* on *price*.
- (c) Other factors include the size of the home, the number of rooms, the quality of the surrounding area (with regards to safety, education, etc.), and the location. As suggested in part(c), these factors are likely correlated with *dist*.
7. (a)  $E[u|inc] = E[\sqrt{inc} \cdot e|inc] = \sqrt{inc} \cdot E[e|inc] = \sqrt{inc} \cdot E[e] = 0$ . Recall that  $E[e|inc] = E[e]$  because we are assuming that  $e$  is independent of *inc*.
- (b)  $\text{Var}(u|inc) = \text{Var}(\sqrt{inc} \cdot e|inc) = (\sqrt{inc})^2 \text{Var}(e|inc) = inc \text{Var}(e) = \sigma_e^2 inc$ . Recall that  $\text{Var}(e|inc) = \text{Var}(e)$  because we are assuming that  $e$  is independent of *inc*.
- (c) This notion makes sense because the greater the amount of income, the more options individuals have with that income. Some individuals may choose to spend their income while others will choose to save it. On the other hand, lower earning individuals have fewer options on how to spend their money. It's likely they have to spend a larger portion of their income on necessary consumption items and have a smaller amounts they can save.
8. (a) From the notes,

$$\begin{aligned}
 \tilde{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\
 \rightarrow E[\tilde{\beta}_1|x] &= E\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \middle| x\right] \\
 &= \frac{1}{\sum_{i=1}^n x_i^2} E\left[\sum_{i=1}^n x_i y_i \middle| x\right] \\
 &= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n E[x_i y_i | x] \\
 &= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i E[\beta_0 + \beta_1 x_i + u_i | x] \\
 &= \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n (\beta_0 x_i + \beta_1 x_i^2) \\
 &= \frac{1}{\sum_{i=1}^n x_i^2} \left( n\beta_0 \bar{x} + \sum_{i=1}^n \beta_1 x_i^2 \right) \\
 &= \frac{n\beta_0 \bar{x} + \beta_1 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2}
 \end{aligned}$$

Thus,

$$E[\tilde{\beta}_1|x, \beta_0 = 0] = \frac{\beta_1 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} = \beta_1$$

Another case in which  $\tilde{\beta}_1$  is unbiased is when  $\bar{x} = \sum_{i=1}^n x_i = 0$ .

(b)

$$\text{Var}(\tilde{\beta}_1|x) = \text{Var}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \middle| x\right)$$

$$\begin{aligned}
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \text{Var} \left( \sum_{i=1}^n x_i (\beta_0 + \beta_1 x_i + u_i) \middle| x \right) \\
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \text{Var} \left( \sum_{i=1}^n (\beta_0 x_i + \beta_1 x_i^2 + x_i u_i) \middle| x \right) \\
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n \text{Var} (\beta_0 x_i + \beta_1 x_i^2 + x_i u_i | x) \\
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \text{Var} (u_i | x) \\
&= \frac{\sigma_u^2}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \\
&= \frac{\sigma_u^2}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

(c)

$$\begin{aligned}
&\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1) \\
&\rightarrow \frac{\sigma_u^2}{\sum_{i=1}^n x_i^2} \leq \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&\rightarrow \frac{1}{\sum_{i=1}^n x_i^2} \leq \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&\rightarrow \sum_{i=1}^n x_i^2 \geq \sum_{i=1}^n (x_i - \bar{x})^2,
\end{aligned}$$

which is given to us by the hint.

- (d) Under the Gauss Markov assumptions,  $\hat{\beta}_1$  is always unbiased while  $\tilde{\beta}_1$  requires the additional constraint that  $y = 0$  when  $x = 0$  to be unbiased. As we just proved,  $\tilde{\beta}_1$  is efficient relative to  $\hat{\beta}_1$ . Thus, in the case that both estimators are unbiased,  $\tilde{\beta}_1$  is preferred.

9. (a) As the problem suggests, using our previous derivation of  $\hat{\beta}_1$ ,

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n ((c_2 x_i - c_2 \bar{x}) c_1 y_i)}{\sum_{i=1}^n (c_2 x_i - c_2 \bar{x})^2} = \frac{c_1 c_2 \sum_{i=1}^n ((x_i - \bar{x}) y_i)}{c_2^2 \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{c_1}{c_2} \hat{\beta}_1$$

We've previously derived  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ . Plugging in the new values and solving for  $\tilde{\beta}_0$ ,

$$\tilde{\beta}_0 = c_1 \bar{y} - \frac{c_1}{c_2} \hat{\beta}_1 (c_2 \bar{x}) = c_1 (\bar{y} - \hat{\beta}_1 \bar{x}) = c_1 \hat{\beta}_0$$

- (b) Again using our previous derivation of  $\hat{\beta}_1$ ,

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (((c_2 + x_i) - (c_2 + \bar{x}))(c_1 + y_i))}{\sum_{i=1}^n ((c_2 + x_i) - (c_2 + \bar{x}))^2} = \frac{\sum_{i=1}^n ((x_i - \bar{x})(c_1 + y_i))}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{c_1 \sum_{i=1}^n (x_i - \bar{x}) + \sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Using  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$  and plugging in the new values,

$$\tilde{\beta}_0 = (c_1 + \bar{y}) - \hat{\beta}_1 (c_2 + \bar{x}) = (\bar{y} - \hat{\beta}_1 \bar{x}) + c_1 - c_2 \hat{\beta}_1 = \hat{\beta}_0 + c_1 - c_2 \hat{\beta}_1$$

- (c) Changing the model from  $\ln(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$  to  $\ln(c_1 y_i) = \tilde{\beta}_0 + \tilde{\beta}_1 x_i$ , the new model can be rewritten as

$$\ln(c_1) + \ln(y_i) = \tilde{\beta}_0 + \tilde{\beta}_1 x_i$$

Trivially, by subtracting  $\ln(c_1)$  from both sides, we see  $\tilde{\beta}_0 = \hat{\beta}_0 + \ln(c_1)$  and  $\tilde{\beta}_1 = \hat{\beta}_1$ .

- (d) Changing the model from  $y_i = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_i)$  to  $y_i = \tilde{\beta}_0 + \tilde{\beta}_1 \ln(c_2 x_i)$ , the new model can be rewritten as

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 (\ln(c_2) + \ln(x_i))$$

Again, it's trivial to see that  $\tilde{\beta}_0 = \hat{\beta}_0 - \tilde{\beta}_1 \ln(c_2) = \hat{\beta}_0 - \hat{\beta}_1 \ln(c_2)$  and  $\tilde{\beta}_1 = \hat{\beta}_1$ .

10. (a)

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{1}{SST_x} \sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + u_i) \\ &= \frac{1}{SST_x} \left[ \beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n x_i (x_i - \bar{x}) + \sum_{i=1}^n u_i (x_i - \bar{x}) \right] \\ &= \frac{1}{SST_x} \left[ \beta_0 (n\bar{x} - n\bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n u_i (x_i - \bar{x}) \right] \\ &= \beta_1 + \frac{1}{SST_x} \sum_{i=1}^n u_i (x_i - \bar{x}) \\ &= \beta_1 + \sum_{i=1}^n \frac{d_i u_i}{SST_x} \\ &= \beta_1 + \sum_{i=1}^n w_i u_i \end{aligned}$$

(b)

$$\begin{aligned} E \left[ (\hat{\beta}_1 - \beta_1) \cdot \bar{u} \mid x \right] &= E \left[ \left( \beta_1 + \sum_{i=1}^n w_i u_i - \beta_1 \right) \cdot \bar{u} \mid x \right] \\ &= E \left[ \sum_{i=1}^n \bar{u} w_i u_i \mid x \right] \\ &= \sum_{i=1}^n E [\bar{u} w_i u_i \mid x] \\ &= \sum_{i=1}^n w_i E [\bar{u} u_i \mid x] \\ &= \sum_{i=1}^n w_i E [\bar{u} u_i \mid x] \end{aligned}$$

Because  $u_i$  and  $u_j$  are pairwise uncorrelated for all  $i \neq j$ ,  $E[u_i u_j] = E[u_i] E[u_j] = 0 \cdot 0 = 0$ . Thus,

$$E[\bar{u} u_i \mid x] = \frac{1}{n} E[u_i^2 \mid x] = \frac{1}{n} \left\{ E[u_i^2 \mid x] - E[u_i \mid x]^2 \right\} = \frac{1}{n} \left\{ \text{Var}(u_i \mid x) \right\} = \frac{\sigma_u^2}{n}. \text{ Hence,}$$

$$\sum_{i=1}^n w_i E[\bar{u} u_i \mid x] = \frac{\sigma_u^2}{n} \sum_{i=1}^n w_i = \frac{\sigma_u^2}{n} \cdot 0 = 0$$

(c)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + \beta_1 \bar{x} + \bar{u} - \hat{\beta}_1 \bar{x} = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1) \bar{x}$$

(d)

$$\begin{aligned} \text{Var}(\hat{\beta}_0 | x) &= \text{Var}(\beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1) \bar{x} | x) \\ &= \text{Var}(\bar{u} - \hat{\beta}_1 \bar{x} | x) \\ &= \text{Var}(\bar{u} | x) + \bar{x}^2 \text{Var}(\hat{\beta}_1 | x) - 2\bar{x} \text{Cov}(\bar{u}, \hat{\beta}_1 | x) \\ &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n u_i \middle| x\right) + \bar{x}^2 \text{Var}\left(\beta_1 + \frac{\sum_{i=1}^n u_i (x_i - \bar{x})}{SST_x} \middle| x\right) - 2\bar{x} \text{Cov}\left(\bar{u}, \beta_1 + \frac{\sum_{i=1}^n u_i (x_i - \bar{x})}{SST_x} \middle| x\right) \\ &= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n u_i \middle| x\right) + \frac{\bar{x}^2}{SST_x^2} \text{Var}\left(\sum_{i=1}^n u_i (x_i - \bar{x}) \middle| x\right) - \frac{2\bar{x}}{SST_x} \text{Cov}\left(\frac{1}{n} \sum_{i=1}^n u_i, \sum_{i=1}^n u_i (x_i - \bar{x}) \middle| x\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(u_i | x) + \frac{\bar{x}^2}{SST_x^2} \sum_{i=1}^n \text{Var}(u_i (x_i - \bar{x}) | x) - \frac{2\bar{x} \sum_{i=1}^n (x_i - \bar{x})}{nSST_x} \text{Cov}\left(\sum_{i=1}^n u_i, \sum_{i=1}^n u_i \middle| x\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma_u^2 + \frac{\bar{x}^2}{SST_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(u_i | x) - \frac{2\bar{x}(n\bar{x} - n\bar{x})}{nSST_x} \text{Cov}\left(\sum_{i=1}^n u_i, \sum_{i=1}^n u_i \middle| x\right) \\ &= \frac{\sigma_u^2}{n} + \frac{\sigma_u^2 \bar{x}^2}{SST_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 - 0 \\ &= \frac{\sigma_u^2}{n} + \frac{\sigma_u^2 \bar{x}^2}{SST_x} \end{aligned}$$

(e)

$$\begin{aligned} \frac{\sigma_u^2}{n} + \frac{\sigma_u^2 \bar{x}^2}{SST_x} &= \frac{\sigma_u^2 SST_x}{nSST_x} + \frac{n\sigma_u^2 \bar{x}^2}{nSST_x} \\ &= \frac{\sigma_u^2 SST_x + n\sigma_u^2 \bar{x}^2}{nSST_x} \\ &= \frac{\sum_{i=1}^n (\sigma_u^2 (x_i - \bar{x})^2) + n\sigma_u^2 \bar{x}^2}{nSST_x} \\ &= \frac{\sum_{i=1}^n (\sigma_u^2 x_i^2 - 2\sigma_u^2 x_i \bar{x} + \sigma_u^2 \bar{x}^2) + n\sigma_u^2 \bar{x}^2}{nSST_x} \\ &= \frac{\sigma_u^2 \sum_{i=1}^n x_i^2 - 2\sigma_u^2 \bar{x} \sum_{i=1}^n x_i + \sigma_u^2 \sum_{i=1}^n \bar{x}^2 + n\sigma_u^2 \bar{x}^2}{nSST_x} \\ &= \frac{\sigma_u^2 \sum_{i=1}^n x_i^2 - 2n\sigma_u^2 \bar{x}^2 + n\sigma_u^2 \bar{x}^2 + n\sigma_u^2 \bar{x}^2}{nSST_x} \\ &= \frac{\sigma_u^2 n^{-1} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

11. (a) I would randomly assign the students to a wide variety of *hours* and then measure the SAT results.
- (b) Two factors that are likely contained in  $u$  are intelligence and cognitive clarity (on the day of the test). It's difficult to say whether intelligence has a positive or negative correlation with *hours*, but it's likely that cognitive clarity is positively correlated with *hours* because more preparation tends to improve a student's ability to remain focused during the exam.

- (c)  $\beta_1$  should be positive.  
 (d)  $\beta_0$  is the expected *sat* for a student that spends 0 hours in the SAT preparation course.  
 12. (a)

$$\begin{aligned}\min_{b_0} SSR &= \sum_{i=1}^n (y_i - b_0)^2 \\ \frac{\partial SSR}{\partial b_0} &= \sum_{i=1}^n -2(y_i - b_0) = 0 \\ &\rightarrow \sum_{i=1}^n y_i - \sum_{i=1}^n b_0 = 0 \\ &\rightarrow n\bar{y} = nb_0 \\ &\rightarrow b_0^* = \tilde{\beta}_0 = \bar{y}\end{aligned}$$

- (b)  $\sum_{i=1}^n \tilde{u}_i = \sum_{i=1}^n (y_i - \bar{y}) = n\bar{y} - n\bar{y} = 0$   
 13. (a)  $1 - x_i$  equals 1 when  $x_i = 0$  and equals 0 otherwise. On the other hand,  $x_i$  equals one when  $x_i = 1$  and equals 0 otherwise. Thus,  $\sum_{i=1}^n (1 - x_i)$  equals the number of observations with  $x_i = 0$  and  $\sum_{i=1}^n x_i$  equals the number of observations with  $x_i = 1$ .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{n_1}{n}$$

$\bar{x}$  is the proportion of observations where  $x_i = 1$ .

(b)

$$\begin{aligned}\bar{y}_0 &= \frac{1}{n_0} \sum_{i=1}^n y_i(0) = n_0^{-1} \sum_{i=1}^n (1 - x_i)y_i \\ \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^n y_i(1) = n_1^{-1} \sum_{i=1}^n x_i y_i\end{aligned}$$

$$(c) \quad \bar{y} = \frac{n_0}{n} \bar{y}_0 + \frac{n_1}{n} \bar{y}_1 = \bar{y}_0 \left[ \frac{1}{n} \sum_{i=1}^n (1 - x_i) \right] + \bar{y}_1 \left[ \frac{1}{n} \sum_{i=1}^n x_i \right] = \bar{y}_0 \left[ \frac{1}{n} (n - n\bar{x}) \right] + \bar{y}_1 \left[ \frac{1}{n} (n\bar{x}) \right] = (1 - \bar{x}) \bar{y}_0 + \bar{x} \bar{y}_1$$

$$(d) \quad n^{-1} \sum_{i=1}^n x_i^2 - \bar{x}^2 = n^{-1} \sum_{i=1}^n x_i - \bar{x}^2 = n^{-1} (n\bar{x}) - \bar{x}^2 = \bar{x} - \bar{x}^2 = \bar{x} (1 - \bar{x})$$

(e)

$$\begin{aligned}n^{-1} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} &= \frac{n_1}{n} \bar{y}_1 - \bar{x} ((1 - \bar{x}) \bar{y}_0 + \bar{x} \bar{y}_1) \\ &= \bar{x} \bar{y}_1 - \bar{x} (\bar{y}_0 - \bar{x} \bar{y}_0 + \bar{x} \bar{y}_1) \\ &= \bar{x} \bar{y}_1 - \bar{x} \bar{y}_0 + \bar{x}^2 \bar{y}_0 - \bar{x}^2 \bar{y}_1 \\ &= \bar{x} (\bar{y}_1 - \bar{y}_0 + \bar{x} \bar{y}_0 - \bar{x} \bar{y}_1) \\ &= \bar{x} (\bar{y}_1 (1 - \bar{x}) - \bar{y}_0 (1 - \bar{x})) \\ &= \bar{x} (1 - \bar{x}) (\bar{y}_1 - \bar{y}_0)\end{aligned}$$

(f)

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2} \\
&= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \\
&= \frac{n^{-1} \{ \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \}}{n^{-1} \{ \sum_{i=1}^n x_i^2 - n\bar{x}^2 \}} \\
&= \frac{n^{-1} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{n^{-1} \sum_{i=1}^n x_i^2 - \bar{x}^2} \\
&= \frac{\bar{x}(1 - \bar{x})(\bar{y}_1 - \bar{y}_0)}{\bar{x}(1 - \bar{x})} = \bar{y}_1 - \bar{y}_0
\end{aligned}$$

$$(g) \quad \bar{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = (1 - \bar{x})\bar{y}_0 + \bar{x}\bar{y}_1 - (\bar{y}_1 - \bar{y}_0) \bar{x} = \bar{y}_0 - \bar{x}\bar{y}_0 + \bar{x}\bar{y}_1 - \bar{x}\bar{y}_1 + \bar{x}\bar{y}_0 = \bar{y}_0$$

14. As proven in the previous problem,  $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$ . Thus, it equals the sample average (proportion) of those who were employed and participated in the program minus the sample average (proportion) of those who were unemployed but did not partake in the program. Thus,  $\hat{\beta}_1$  is simply the difference in employment rates between those who participated in the program and those who did not within the given sample.

15. (a)  $E[n^{-1} \sum_{i=1}^n [y_i(1) - y_i(0)]] = n^{-1} \sum_{i=1}^n (E[y_i(1)] - E[y_i(0)]) = n^{-1} (nE[y_i(1)] - nE[y_i(0)]) = E[y_i(1)] - E[y_i(0)] = \tau_{ate}$
- (b)  $\bar{y}_0$  is the sample average of  $y$  for the observations where  $x_i = 0$ .  $\bar{y}_1$  is the sample average of  $y$  for the observations where  $x_i = 1$ .  $\bar{y}(1)$  and  $\bar{y}(0)$  are relative to the entire sample such that  $\bar{y}(\cdot) = n \cdot n^{-1} \bar{y}$ .
16. (a) The difference in means estimator is generally no longer unbiased because the ceteris paribus condition is no longer satisfied. More specifically, the decision to participate may be correlated with the independent and dependent variables, which would cause the estimator to be biased.
- (b) One example that would cause bias is if wealthier individuals tended to not participate. It's likely wealth is correlated with both *unemployment* and *program*, which would cause for a biased estimator.

17.

$$\begin{aligned}
\text{Var}(u_i | x_i) &= \text{Var}((1 - x_i)u_i(0) + x_i u_i(1) | x) \\
&= \text{Var}((1 - x_i)u_i(0) | x) + \text{Var}(x_i u_i(1) | x) \\
&= (1 - x_i)^2 \text{Var}(u_i(0) | x) + x_i^2 \text{Var}(u_i(1) | x) \\
&= (1 - x_i)^2 \sigma_0^2 + x_i^2 \sigma_1^2
\end{aligned}$$

18. (a)  $P(\text{All } x = 1 \text{ or all } x = 0) = P(\text{All } x = 1 \cup \text{all } x = 0) = P(\text{All } x = 1) + P(\text{All } x = 0) - P(\text{All } x = 1 \cap \text{all } x = 0) = \rho^n + (1 - \rho)^n - P(\emptyset) = \rho^n + (1 - \rho)^n$ . Because  $0 < \rho < 1$  (and thus  $0 < (1 - \rho) < 1$ ),

$$\lim_{n \rightarrow \infty} P(\text{All } x = 1) = \lim_{n \rightarrow \infty} \rho^n = 0$$

Likewise,

$$\lim_{n \rightarrow \infty} P(\text{All } x = 0) = \lim_{n \rightarrow \infty} (1 - \rho)^n = 0$$

(b)  $n=10$ :

$$P(\text{All } x = 1 \text{ or all } x = 0) = P(\text{All } x = 1 \cup \text{all } x = 0) = 0.5^{10} + (1 - 0.5)^{10} = 2 \cdot 0.5^{10} \approx 0.0019531$$

$n=100$

$$P(\text{All } x = 1 \text{ or all } x = 0) = P(\text{All } x = 1 \cup \text{all } x = 0) = 0.5^{100} + (1 - 0.5)^{100} = 2 \cdot 0.5^{100} \approx 0$$

(c)  $n=10$ :

$$P(\text{All } x = 1 \text{ or all } x = 0) = P(\text{All } x = 1 \cup \text{all } x = 0) = 0.9^{10} + (1 - 0.9)^{10} \approx 0$$

$n=100$

$$P(\text{All } x = 1 \text{ or all } x = 0) = P(\text{All } x = 1 \cup \text{all } x = 0) = 0.9^{100} + (1 - 0.9)^{100} \approx 0$$

## Computer Exercises

```
#i
mean(k401k$prate)
```

```
[1] 87.36291
```

```
mean(k401k$mrage)
```

```
[1] 0.7315124
```

```
#ii
lm1 <- lm(prate ~ mrage, data = k401k)
summary(lm1)
```

Call:

```
lm(formula = prate ~ mrage, data = k401k)
```

Residuals:

Min	1Q	Median	3Q	Max
-82.303	-8.184	5.178	12.712	16.807

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	83.0755	0.5633	147.48	<0.0000000000000002 ***
mrage	5.8611	0.5270	11.12	<0.0000000000000002 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.09 on 1532 degrees of freedom

Multiple R-squared: 0.0747, Adjusted R-squared: 0.0741

F-statistic: 123.7 on 1 and 1532 DF, p-value: < 0.00000000000000022

```
#iii
#The intercept suggests a participation rate of 83.07546% when the match rate is 0%
#The coefficient on mrage suggests an additional 5.861079% in participation rate for
#every additional percentage of prate
```

```
#iv
```

```
unnamed[1] + unnamed[2] * 3.5
```

```
[1] 103.5892
```

```
#The percentage exceeds 100% as this is a linear model. Clearly, this is not a reasonable
#prediction
```

```
#v
summary(lm1)["r.squared"]

$r.squared
[1] 0.0747031

#About 7.5%. Personally, I would consider this a lot for one variable, but I
#expected it to be higher
rm(lm1)
```

```
#i
mean(ceosal2$salary)
```

```
[1] 865.8644
```

```
mean(ceosal2$ceoten)
```

```
[1] 7.954802
```

```
#ii
sum(ceosal2$ceoten==0)
```

```
[1] 5
```

```
max(ceosal2$ceoten)
```

```
[1] 37
```

```
#iii
lm(log(salary) ~ ceoten, data = ceosal2)
```

```
Call:
lm(formula = log(salary) ~ ceoten, data = ceosal2)
```

```
Coefficients:
(Intercept)      ceoten
  6.505498      0.009724
```

```
#An additional year as CEO is predicted to increase salary by about 1%.
```

```
#i
lm1 <- lm(sleep ~ totwrk, data = sleep75)
summary(lm1)
```

```
Call:
lm(formula = sleep ~ totwrk, data = sleep75)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-2429.94  -240.25    4.91   250.53  1339.72
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3586.37695   38.91243   92.165 <0.0000000000000002 ***
totwrk       -0.15075    0.01674   -9.005 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 421.1 on 704 degrees of freedom
```



Multiple R-squared: 0.1033, Adjusted R-squared: 0.102  
F-statistic: 81.09 on 1 and 704 DF, p-value: < 0.00000000000000022

*#The intercept (3586.37695) is the predicted minutes of sleep for an individual with  
#0 minutes spent in paid work*

*#ii*  
unnname(lm1\$coefficients[2])\*120

[1] -18.0895

*#A loss in 18 minutes of sleep for additional hours of totwrk does not seem large*  
rm(lm1)

*#i*  
mean(wage2\$wage)

[1] 957.9455

mean(wage2\$IQ)

[1] 101.2824

sd(wage2\$IQ)

[1] 15.05264

*#ii*  
lm1 <- lm(wage ~ IQ, data = wage2)  
unnname(lm1\$coefficients[2])\*15

[1] 124.546

summary(lm1)["r.squared"]

\$r.squared  
[1] 0.09553528

*#IQ explains about 9.5% of the variation in wage*

*#iii*  
lm1 <- lm(log(wage) ~ IQ, data = wage2)  
unnname(lm1\$coefficients[2])\*15\*100

[1] 13.21073

rm(lm1)

*#i*  
*#ln(rd) = beta\_0+ beta\_1 ln(sales) + u*

*#ii*  
lm1 <- lm(lrd ~ lsales, data = rdchem)  
summary(lm1)

Call:

lm(formula = lrd ~ lsales, data = rdchem)

Residuals:

	Min	1Q	Median	3Q	Max
	-0.90406	-0.40086	-0.02178	0.40562	1.10439

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.10472	0.45277	-9.066	0.000000000427 ***
lsales	1.07573	0.06183	17.399	< 0.0000000000000002 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5294 on 30 degrees of freedom

Multiple R-squared: 0.9098, Adjusted R-squared: 0.9068

F-statistic: 302.7 on 1 and 30 DF, p-value: < 0.00000000000000022

*#The estimated elasticity suggests a 1.07573% increase in rd for every percentage increase in sales*

*#i*

*#I expected a diminishing marginal return in the pass rate for each additional dollar spent.*

*#For one, the pass rate cannot exceed 100%. Additionally, only so much can be spent on a student*

*#before no additional gains are possible. Thus, a constant marginal return seems unlikely/unreasonable.*

*#ii*

*#The percentage change in math10 for a 1% increase in expend is beta\_1/100*

*#Thus, beta\_1/10 is the percentage point change in math10 given a 10% increase in expend*

*#iii*

```
lm1 <- lm(math10 ~ lexpend, data = meap93)
```

```
summary(lm1)
```

Call:

```
lm(formula = math10 ~ lexpend, data = meap93)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.343	-7.100	-0.914	6.148	39.093

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-69.341	26.530	-2.614	0.009290 **
lexpend	11.164	3.169	3.523	0.000475 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.35 on 406 degrees of freedom

Multiple R-squared: 0.02966, Adjusted R-squared: 0.02727

F-statistic: 12.41 on 1 and 406 DF, p-value: 0.0004752

*#iv*

```
unname(lm1$coefficients[2]) / 10
```

```
[1] 1.116439
```

*#v*

```
max(meap93$math10)
```

```
[1] 66.7
```

```
#In this data set, the largest value of math10 is 66.7, which isn't particularly close to 100  
rm(lm1)
```

```
#i  
mean(charity$gift)
```

```
[1] 7.44447
```

```
mean(charity$gift > 0)
```

```
[1] 0.3999531
```

```
#ii  
mean(charity$mailsyear)
```

```
[1] 2.049555
```

```
max(charity$mailsyear)
```

```
[1] 3.5
```

```
min(charity$mailsyear)
```

```
[1] 0.25
```

```
#iii  
lm1 <- lm(gift ~ mailsyear, data = charity)  
summary(lm1)
```

Call:

```
lm(formula = gift ~ mailsyear, data = charity)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.287	-7.976	-5.976	2.687	245.999

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.0141	0.7395	2.724	0.00648 **
mailsyear	2.6495	0.3431	7.723	0.000000000000014 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.96 on 4266 degrees of freedom

Multiple R-squared: 0.01379, Adjusted R-squared: 0.01356

F-statistic: 59.65 on 1 and 4266 DF, p-value: 0.00000000000001404

```
#iv  
#The slope coefficient predicts about a 2.7 Dutch guilders increase in gift for every additional  
#mailsyear. If each mailing costs on guilder, the charity is expected to make a net gain on each  
#mailing, but this doesn't mean the charity makes a net gain on every mailing
```

```
#v  
min(lm1$fitted.values)
```

```
[1] 2.676466
```

```
#With this regression result, a negative value for mailsyear would be required to predict 0 for  
#gift, which is unfeasible
```

```
rm(lm1)
```

```
#i
```

```
x <- runif(500, 0, 10)
mean(x)
```

```
[1] 4.916788
```

```
sd(x)
```

```
[1] 2.98676
```

```
#ii
```

```
u <- runif(500, 0, 36)
#No, the sample average is not zero because all of the u are positive
sd(u)
```

```
[1] 10.10354
```

```
#iii
```

```
y <- 1 + 2 * x + u
lm1 <- lm(y ~ x)
summary(lm1)
```

```
Call:
```

```
lm(formula = y ~ x)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-18.4352	-8.2309	-0.4958	8.5050	18.8156

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	20.0587	0.8695	23.07	<0.0000000000000002 ***
x	1.7509	0.1512	11.58	<0.0000000000000002 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 10.09 on 498 degrees of freedom
```

```
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2106
```

```
F-statistic: 134.1 on 1 and 498 DF,  p-value: < 0.00000000000000022
```

```
#No, they are not equal because the unobserved factors are not included in the regression model
```

```
#iv
```

```
sum(lm1$residuals)
```

```
[1] 0.0000000000002867151
```

```
sum(lm1$residuals * x)
```

```
[1] 0.0000000000001879275
```

```
#v
```

```
sum(u)
```

```
[1] 8916.918
```

```
sum(u * x)
```

```
[1] 42733.63
```

```
#The sum of the error terms and the errors terms times x generally won't sum  
#to 0 like the residuals will
```

```
#vi
```

```
x <- runif(500, 0, 10)  
mean(x)
```

```
[1] 4.948361
```

```
sd(x)
```

```
[1] 3.005408
```

```
u <- runif(500, 0, 36)  
sd(u)
```

```
[1] 10.79501
```

```
y <- 1 + 2 * x + u  
lm2 <- lm(y ~ x)  
summary(lm2)
```

```
Call:
```

```
lm(formula = y ~ x)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-19.737	-10.014	0.388	9.447	17.858

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	21.0973	0.9280	22.73	<0.0000000000000002 ***
x	1.6829	0.1603	10.50	<0.0000000000000002 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 10.76 on 498 degrees of freedom
```

```
Multiple R-squared:  0.1812,    Adjusted R-squared:  0.1795
```

```
F-statistic: 110.2 on 1 and 498 DF,  p-value: < 0.00000000000000022
```

```
#The results are not the same because the samples are different  
rm(x,y,u,lm1,lm2)
```

```
county_murders <- as_tibble(countymurders) %>%  
  filter(year == 1996)
```

```
#i
```

```
sum(county_murders$murders == 0)
```

```
[1] 1051
```

```
sum(county_murders$execs > 0)
```

```
[1] 31
```

```
max(county_murders$execs)
```

```
[1] 3
```

```
#ii
lm1 <- lm(murders ~ execs, data = county_murders)
summary(lm1)
```

Call:

```
lm(formula = murders ~ execs, data = county_murders)
```

Residuals:

Min	1Q	Median	3Q	Max
-149.12	-5.46	-4.46	-2.46	1338.99

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.4572	0.8348	6.537	0.0000000000779 ***
execs	58.5555	5.8333	10.038	< 0.000000000000002 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.89 on 2195 degrees of freedom

Multiple R-squared: 0.04389, Adjusted R-squared: 0.04346

F-statistic: 100.8 on 1 and 2195 DF, p-value: < 0.0000000000000022

```
#iii
#The model predicts an additional 58.555 executions for every additional murder in a county.
#This does not suggest a deterrent effect of capital punishment
```

```
#iv
min(lm1$fitted.values)
```

```
[1] 5.457241
```

```
unnname(lm1$residuals[which(county_murders$execs == 0)[1]])
```

```
[1] 1.542759
```

```
#v
#A simple regression analysis not well suited for determining whether capital punishment
#has a deterrent effect on murders because there are many unobserved factors that are likely
#correlated with both murders and execs, which leads to a biased slope estimate. Also, there's
#likely reverse causality in which murders causes more execs
rm(county_murders, lm1)
```

```
#i
length(catholic$math12)
```

```
[1] 7430
```

```
mean(catholic$math12)
```

```
[1] 52.13362
```

```
sd(catholic$math12)
```

```
[1] 9.459117
```

```
mean(catholic$read12)
```

```
[1] 51.7724
```

```
sd(catholic$read12)
```

```
[1] 9.407761
```

```
#ii
```

```
lm1 <- lm(math12 ~ read12, data = catholic)
```

$$\widehat{math12} = 15.1530378 + 0.7142915read12$$

$$n = 7430, R^2 = 0.5046872$$

```
#iii
```

```
#Yes, the intercept suggests prediction of 15.15304 on math12 for a score of 0 on read12
```

```
#iv
```

```
#I'm not surprised by the coefficient. I would expect the scores to be positively correlated.
```

```
#I'm somewhat surprised by how large the coefficient of determination is.
```

```
#v
```

```
#I would counter their statement by stating it's more correlational rather than causal.
```

```
#Unobserved factors correlated with math12 and read12 are uncontrolled for, which would
```

```
#make the slope coefficient biased (if interpreted as a causal effect)
```

```
#i
```

```
length(gpa1$colGPA)
```

```
[1] 141
```

```
mean(gpa1$colGPA)
```

```
[1] 3.056738
```

```
max(gpa1$colGPA)
```

```
[1] 4
```

```
#ii
```

```
sum(gpa1$PC)
```

```
[1] 56
```

```
#iii
```

```
lm1 <- lm(colGPA ~ PC, data = gpa1)
```

```
summary(lm1)
```

```
Call:
```

```
lm(formula = colGPA ~ PC, data = gpa1)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.95893	-0.25893	0.01059	0.31059	0.84107

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.98941	0.03950	75.678	<0.0000000000000002	***
PC	0.16952	0.06268	2.704	0.0077	**

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.3642 on 139 degrees of freedom  
Multiple R-squared: 0.04999, Adjusted R-squared: 0.04315  
F-statistic: 7.314 on 1 and 139 DF, p-value: 0.007697

*#The intercept estimate predicts a colGPA of 2.98941 for a student without a PC.  
#The slope estimate predicts a 0.16952 estimate of the "treatment" effect of having a PC.  
#The slope suggests a positive relationship between colGPA and PC (as one might expect).*

*#iv*  
unnname(unlist(summary(lm1)["r.squared"]))

[1] 0.04998907

*#The r-squared suggests not much of the variation in colGPA is explained by PC*

*#v  
#No, unobserved factors that are correlated with colGPA and PC do not allow for beta\_1  
#to be interpreted as an unbiased estimate of the causal effect of owning a PC on colGPA.*

rm(lm1)



## Chapter 3

### Notes

#### Multiple Regression Analysis

**Multiple regression analysis** is generally better fit to predict causal effects over simple regression analysis because it allows us to explicitly control for many factors that are correlated with the singular independent variable (from the simple regression model) and the dependent variable.

In general, the **multiple linear regression (MLR) model** takes the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

#### OLS Estimators Derivation (MLR)

Synonymous to the case with simple linear regression analysis, using OLS to obtain estimators of the parameters  $\beta_0, \beta_1, \dots, \beta_k$  yields the following:

$$\begin{aligned} \min_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} SSR &= \sum_{i=1}^n \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right)^2 \\ \frac{\partial SSR}{\partial \hat{\beta}_0} &= \sum_{i=1}^n -2 \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \\ \frac{\partial SSR}{\partial \hat{\beta}_1} &= \sum_{i=1}^n -2 x_{i1} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \\ &\quad \dots \\ \frac{\partial SSR}{\partial \hat{\beta}_k} &= \sum_{i=1}^n -2 x_{ik} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \end{aligned}$$

These are often called the OLS **first order conditions**. As with the simple regression model, the OLS first order conditions can be obtained by the method of moments: under the assumptions that  $E[u] = 0$  and  $E[x_j u] = 0 \forall j = 1, \dots, k$ .

From the first FOC,

$$\begin{aligned} \sum_{i=1}^n \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) &= 0 \\ \rightarrow \sum_{i=1}^n \hat{\beta}_0 &= \sum_{i=1}^n \left( y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) \\ \rightarrow n \hat{\beta}_0 &= n \bar{y} - n \hat{\beta}_1 \bar{x}_1 - \cdots - n \hat{\beta}_k \bar{x}_k \\ \rightarrow \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 - \cdots - \hat{\beta}_k \bar{x}_k \end{aligned}$$

*Note:* Like in the simple linear regression case, the significance of this FOC is that, using least squares criteria, our line of best fit runs through the sample means  $\bar{y}$  and  $\bar{x}_j$  for  $j = 1, \dots, k$ .

To derive the slope estimators, start by regressing  $x_\ell$  on all of the other regressors in the model, which takes the form

$$x_\ell = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_{\ell-1} x_{\ell-1} + \gamma_{\ell+1} x_{\ell+1} + \cdots + \gamma_k x_k + r_{i\ell}$$

If we denote the residuals of the model as  $\hat{r}_{i\ell}$  and the predicted values as  $\hat{x}_{i\ell}$ , then

$$x_{i\ell} = \hat{x}_{i\ell} + \hat{r}_{i\ell}$$

Plugging this derivation into the  $(\ell + 1)^{th}$  first order condition:

$$\begin{aligned}
\frac{\partial SSR}{\partial \beta_\ell} &= \sum_{i=1}^n -2x_{i\ell} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \\
&\rightarrow \sum_{i=1}^n (\hat{x}_{i\ell} + \hat{r}_{i\ell}) \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \\
&\rightarrow \sum_{i=1}^n \hat{x}_{i\ell} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) + \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \\
&\rightarrow \sum_{i=1}^n \hat{x}_{i\ell} \hat{u}_i + \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0
\end{aligned}$$

Because  $\hat{x}_{i\ell}$  is simply a linear combination of all the other regressors in the model (i.e.,  $\hat{x}_{i\ell} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{i1} + \cdots + \hat{\gamma}_{\ell-1} x_{i\ell-1} + \hat{\gamma}_{\ell+1} x_{i\ell+1} + \cdots + \hat{\gamma}_k x_{ik}$ ), it follows that  $\sum_{i=1}^n \hat{x}_{i\ell} \hat{u}_i = 0$ .

$$\begin{aligned}
&\rightarrow \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0 \\
&\rightarrow \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_\ell x_{i\ell} \right) + \sum_{i=1}^n \hat{r}_{i\ell} \left( -\hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{\ell-1} x_{i\ell-1} - \hat{\beta}_{\ell+1} x_{i\ell+1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0
\end{aligned}$$

Because  $\hat{r}_{i\ell}$  are the residuals from regressing  $x_\ell$  on all the other regressors,  $\sum_{i=1}^n x_{ij} \hat{r}_{i\ell} = 0 \ \forall \ j \neq \ell$ .

$$\begin{aligned}
&\rightarrow \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_\ell x_{i\ell} \right) = 0 \\
&\rightarrow \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_\ell (\hat{x}_{i\ell} + \hat{r}_{i\ell}) \right) = 0 \\
&\rightarrow \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_\ell \hat{r}_{i\ell} \right) - \hat{\beta}_\ell \sum_{i=1}^n \hat{r}_{i\ell} \hat{x}_{i\ell} = 0
\end{aligned}$$

Since  $\sum_{i=1}^n \hat{r}_{i\ell} \hat{x}_{i\ell} = 0$  (This can be thought of as  $\sum_{i=1}^n \hat{r}_{i\ell} \hat{x}_{i\ell} = 0 \rightarrow \sum_{i=1}^n \hat{u}_i \hat{y}_i = 0 \rightarrow \hat{\beta}_0 \sum_{i=1}^n \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^n x_i \hat{u}_i = 0 + 0 = 0$  in the SLR case),

$$\begin{aligned}
&\rightarrow \sum_{i=1}^n \hat{r}_{i\ell} \left( y_i - \hat{\beta}_\ell \hat{r}_{i\ell} \right) = 0 \\
&\rightarrow \hat{\beta}_\ell \sum_{i=1}^n \hat{r}_{i\ell}^2 = \sum_{i=1}^n \hat{r}_{i\ell} y_i \\
&\rightarrow \hat{\beta}_\ell = \frac{\sum_{i=1}^n \hat{r}_{i\ell} y_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2}
\end{aligned}$$

## MLR Interpretation

In MLR analysis, the slope estimates  $\hat{\beta}_1, \dots, \hat{\beta}_k$  are interpreted as the **partial effects** of the corresponding explanatory variables on the dependent variable. In other words, they have a **ceteris paribus** interpretation.

## OLS Fitted Values and Residuals Properties (MLR)

*Note:* The OLS fitted values and residuals have some important properties that are immediate extensions from the bivariate case:

1. The sample average of the residuals is zero and so  $\bar{y} = \bar{\hat{y}}$
2. The sample covariance between each independent variable and the OLS residuals is zero (i.e.,  $\hat{\sigma}_{x_j \hat{u}} = 0 \forall j = 1, \dots, k$ ). Hence,

$$\hat{\sigma}_{\hat{y} \hat{u}} = 0$$

3. The point  $(\bar{x}_1, \dots, \bar{x}_k, \bar{y})$  always lies on the OLS regression line:  $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \dots + \hat{\beta}_k \bar{x}_k$

The first two properties are immediate consequences of the OLS first order conditions used to obtain the OLS estimators. Namely,

$$\sum_{i=1}^n u_i = 0; \sum_{i=1}^n x_{ij} u_i = 0,$$

where  $\sum_{i=1}^n x_{ij} u_i = 0$  implies that each regressor has a zero sample covariance with  $\hat{u}_i$ . The third property was derived earlier in the “OLS Estimators Derivation (MLR)” section.

## Partialling Out Interpretation

Earlier, we derived

$$\hat{\beta}_\ell = \frac{\sum_{i=1}^n \hat{r}_{i\ell} y_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2}$$

Let  $\ell = 1$  in the model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ . Then,  $\hat{r}_{i1}$  are the OLS residuals from a simple regression of  $x_1$  on  $x_2$ . The above derivation shows that we can then do a simple regression of  $y$  on  $\hat{r}_1$  to obtain  $\hat{\beta}_1$ . (Note that the residuals  $\hat{r}_{i1}$  have a zero sample average, and so  $\hat{\beta}_1$  is the usual slope estimate from simple regression). The interpretation of this is that the residuals  $\hat{r}_{i1}$  are the part of  $x_{i1}$  that are uncorrelated with  $x_{i2}$ . In other words,  $\hat{r}_{i1}$  is  $x_{i1}$  after the effects of  $x_{i2}$  have been partialled out. Thus,  $\hat{\beta}_1$  measures the sample relationship between  $y$  and  $x_1$  after  $x_2$  has been partialled out. This result is usually called the **Frisch-Waugh theorem**.

## Simple vs. Multiple Regression Estimates

Define a simple regression model  $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$  and a multiple regression model  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ . Then, it turns out

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1,$$

where  $\tilde{\delta}_1$  is the slope coefficient from regressing  $x_{i2}$  on  $x_{i1}$  and thus the confounding term is the partial effect of  $x_2$  on  $\hat{y}$  times  $\tilde{\delta}_1$ . The two cases in which  $\tilde{\beta}_1 = \hat{\beta}_1$  are

1. The partial effect of  $x_2$  on  $\hat{y}$  is zero in the sample (i.e.,  $\hat{\beta}_2 = 0$ )
2.  $x_1$  and  $x_2$  are uncorrelated in the sample (i.e.,  $\tilde{\delta}_1 = 0$ )

More generally, if we define a model  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$  and another model  $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \dots + \tilde{\beta}_{k-1} x_{k-1}$ , then

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j$$

for  $j = 1, \dots, k-1$ . Taking the conditional expectation of this value (on  $x$ ):

$$E[\tilde{\beta}_j | x] = E[\hat{\beta}_j | x] + E[\hat{\beta}_k \tilde{\delta}_j | x] = \beta_j + \tilde{\delta}_j E[\hat{\beta}_k | x] = \beta_j + \beta_k \tilde{\delta}_j,$$

which shows  $\tilde{\beta}_j$  is biased for  $\beta_j$  unless  $\beta_k = 0$  (meaning  $x_k$  is uncorrelated with  $y$  in the population) or  $\tilde{\delta}_j = 0$  (meaning  $x_k$  is uncorrelated with  $x_j$  in the sample).

## Unbiasedness of OLS Estimators (MLR)

Earlier, we derived

$$\begin{aligned}
 \hat{\beta}_\ell &= \frac{\sum_{i=1}^n \hat{r}_{i\ell} y_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \frac{\sum_{i=1}^n \hat{r}_{i\ell} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i)}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \frac{\beta_0 \sum_{i=1}^n \hat{r}_{i\ell} + \beta_1 \sum_{i=1}^n x_{i1} \hat{r}_{i\ell} + \cdots + \beta_k \sum_{i=1}^n x_{ik} \hat{r}_{i\ell} + \sum_{i=1}^n u_i \hat{r}_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \frac{\beta_\ell \sum_{i=1}^n x_{i\ell} \hat{r}_{i\ell} + \sum_{i=1}^n u_i \hat{r}_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \frac{\beta_\ell \sum_{i=1}^n (\hat{x}_{i\ell} + \hat{r}_{i\ell}) \hat{r}_{i\ell} + \sum_{i=1}^n u_i \hat{r}_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \frac{\beta_\ell \sum_{i=1}^n \hat{x}_{i\ell} \hat{r}_{i\ell} + \beta_\ell \sum_{i=1}^n \hat{r}_{i\ell}^2 + \sum_{i=1}^n u_i \hat{r}_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \frac{\beta_\ell \sum_{i=1}^n (\hat{\gamma}_0 + \hat{\gamma}_1 x_{i1} + \cdots + \hat{\gamma}_{\ell-1} x_{i\ell-1} + \gamma_{\ell+1} x_{i\ell+1} + \cdots + \gamma_k x_{ik}) \hat{r}_{i\ell} + \beta_\ell \sum_{i=1}^n \hat{r}_{i\ell}^2 + \sum_{i=1}^n \hat{r}_{i\ell} u_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \frac{0 + \beta_\ell \sum_{i=1}^n \hat{r}_{i\ell}^2 + \sum_{i=1}^n \hat{r}_{i\ell} u_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
 &= \beta_\ell + \frac{\sum_{i=1}^n \hat{r}_{i\ell} u_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2}
 \end{aligned}$$

Taking the expectation of this value (conditional on  $x$ ):

$$\begin{aligned}
 E[\hat{\beta}_\ell | x] &= E\left[\beta_\ell + \frac{\sum_{i=1}^n \hat{r}_{i\ell} u_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \middle| x\right] \\
 &= \beta_\ell + \frac{1}{\sum_{i=1}^n \hat{r}_{i\ell}^2} E\left[\sum_{i=1}^n \hat{r}_{i\ell} u_i \middle| x\right] \\
 &= \beta_\ell + \frac{1}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \sum_{i=1}^n E[\hat{r}_{i\ell} u_i | x] \\
 &= \beta_\ell + \frac{1}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \sum_{i=1}^n \hat{r}_{i\ell} E[u_i | x] \\
 &= \beta_\ell + \frac{1}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \sum_{i=1}^n \hat{r}_{i\ell} \cdot 0 = \beta_\ell
 \end{aligned}$$

An important implication of using OLS is that the *procedure* by which OLS estimates are obtained is unbiased; however, it's almost always not the case that the estimate obtained is unbiased. Thus, if  $\hat{\beta}_1$  is the OLS slope parameter estimator and  $b_1$  is an OLS estimate from a particular sample, one may say  $\hat{\beta}_1$  is unbiased under the G-M assumptions (i.e.,  $E[\hat{\beta}_1] = \beta_1$ ), but one should refrain from saying  $b_1$  is unbiased.

## Specifying a Model

In multiple regression analysis **inclusion of an irrelevant variable** or **overspecifying/overfitting the model** is a case in which an explanatory variable is included in a regression model that has a zero population parameter in estimating an equation by OLS. While overspecifying the model does not affect the unbiasedness of the OLS estimators, including irrelevant variables can have undesirable effects on the variances of the OLS estimators.

On the other hand, omitting a variable that actually belongs in the true (or population) model is called the problem of **excluding a relevant variable** or **underspecifying the model**. Looking back to the “Sample vs. Multiple Regression Estimates” section, define a model  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$  and another model  $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_{k-1} x_{k-1}$ , then

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j$$

for  $j = 1, \dots, k-1$ . Taking the conditional expectation of this value (on  $x$ ):

$$E[\tilde{\beta}_j | x] = E[\hat{\beta}_j | x] + E[\hat{\beta}_k \tilde{\delta}_j | x] = \beta_j + \tilde{\delta}_j E[\hat{\beta}_k | x] = \beta_j + \beta_k \tilde{\delta}_j,$$

Thus, we can derive the **omitted variable bias** as

$$\text{Bias}(\tilde{\beta}_j) = E[\tilde{\beta}_j | x] - \beta_j = \beta_j + \beta_k \tilde{\delta}_j - \beta_j = \beta_k \tilde{\delta}_j$$

Summary of Bias in  $\tilde{\beta}_j$  when  $x_k$  is Omitted

	$\text{Corr}(x_j, x_k) > 0$	$\text{Corr}(x_j, x_k) < 0$
$\beta_k > 0$	Positive Bias	Negative Bias
$\beta_k < 0$	Negative Bias	Positive Bias

More generally, it's difficult to derive the sign of omitted variable bias with multiple regressors. For example, suppose

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

satisfies the unbiasedness G-M assumptions, but we estimate the model as

$$\hat{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2$$

If  $x_1$  is correlated with  $x_3$  but  $x_2$  and  $x_3$  are uncorrelated, both  $\tilde{\beta}_1$  and  $\tilde{\beta}_2$  will both be biased unless both  $x_1$  and  $x_2$  are also uncorrelated. In general, we commonly will assume  $x_1$  and  $x_2$  are uncorrelated to derive a likely direction of biasedness; however, it should be noted that this assumption generally doesn't hold.

In estimating parameters, we say an estimator has an **upward bias** if it has positive bias and a **downward bias** if it has a negative bias. In general, we say an estimator is **biased toward zero** if the estimator is closer to zero than the true parameter.

### Sampling Variance of OLS Estimators (MLR)

Using a previous derivation,

$$\hat{\beta}_\ell = \beta_\ell + \frac{\sum_{i=1}^n \hat{r}_{i\ell} u_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2}$$

Thus,

$$\begin{aligned} \text{Var}(\hat{\beta}_\ell | x) &= \text{Var}\left(\beta_\ell + \frac{\sum_{i=1}^n \hat{r}_{i\ell} u_i}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \middle| x\right) \\ &= \frac{1}{(\sum_{i=1}^n \hat{r}_{i\ell}^2)^2} \text{Var}\left(\sum_{i=1}^n \hat{r}_{i\ell} u_i \middle| x\right) \\ &= \frac{1}{(\sum_{i=1}^n \hat{r}_{i\ell}^2)^2} \sum_{i=1}^n \text{Var}(\hat{r}_{i\ell} u_i | x) \\ &= \frac{1}{(\sum_{i=1}^n \hat{r}_{i\ell}^2)^2} \sum_{i=1}^n \hat{r}_{i\ell}^2 \text{Var}(u_i | x) \\ &= \frac{1}{(\sum_{i=1}^n \hat{r}_{i\ell}^2)^2} \sum_{i=1}^n \hat{r}_{i\ell}^2 \sigma_u^2 \end{aligned}$$

$$\begin{aligned}
&= \frac{\sigma_u^2 \sum_{i=1}^n \hat{r}_{i\ell}^2}{\left(\sum_{i=1}^n \hat{r}_{i\ell}^2\right)^2} \\
&= \frac{\sigma_u^2}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
&= \frac{\sigma_u^2}{SST_{x_\ell} - SSE_{x_\ell}} \\
&= \frac{\sigma_u^2}{SST_{x_\ell} (1 - R_\ell^2)}
\end{aligned}$$

where  $R_\ell^2$  is the  $R$ -squared from regressing  $x_\ell$  on all other independent variables.

High but imperfect correlation between two or more independent variables is called **multicollinearity**. Multicollinearity leads to large values of  $R_\ell^2$ , which in turn, leads to large values for  $\text{Var}(\hat{\beta}_\ell)$ . Worrying about high degrees of correlation among the independent variables in the sample is really no different from worrying about a small sample size as both work to increase  $\text{Var}(\hat{\beta}_\ell)$ . Thus, ideas surrounding small values for  $R_\ell^2$  and  $SST_{x_\ell}$  pertain to **micronumerosity**, or the problem of small sample size. While multicollinearity can pose an issue in statistical inference, it needn't always be cause for stress for two reasons. First, multicollinearity can point out "improper" questions in which we are asking questions that may be too subtle for the available data to answer with any precision. Additionally, regression analysis where there is a high degree of correlation between certain independent variables can be irrelevant as to how well we can estimate parameters of interest in a model. For example, suppose we hope to estimate

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

where  $x_2$  and  $x_3$  are highly correlated by  $x_1$  is not highly correlated with either. It turns out that the high level of correlation between  $x_2$  and  $x_3$  has no direct effect on  $\text{Var}(\hat{\beta}_1)$ . Thus, if  $\beta_1$  is the parameter of interest, including  $x_2$  and  $x_3$  maintains the unbiasedness of  $\hat{\beta}_1$  while imposing little to no additional variance in  $\hat{\beta}_1$ . The most common statistic for detecting multicollinearity for a particular coefficient is the **variance inflation factor (VIF)**, which equals  $\frac{1}{1-R_\ell^2}$ . Hence,  $\text{Var}(\hat{\beta}_\ell)$  can be rewritten as

$$\text{Var}(\hat{\beta}_\ell) = \frac{\sigma_u^2}{SST_\ell} \cdot \text{VIF}_\ell,$$

which shows that  $\text{VIF}_\ell$  is the factor by which  $\text{Var}(\hat{\beta}_\ell)$  is higher because  $x_\ell$  is not uncorrelated with the other explanatory variables.

Similar to the SLR case, these formulas allow us to isolate the factors that contribute to  $\text{Var}(\hat{\beta}_\ell)$ . But these formulas are unknown, except in the extremely rare case that  $\sigma_u^2$  is known. Considering the  $u_i$  are unobserved, an unbiased estimator of  $\sigma_u^2$  is

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n - k - 1},$$

where the  $n - k - 1$  is the degrees of freedom in the OLS residuals due to the  $k + 1$  OLS first order conditions:

$$\sum_{i=1}^n \hat{u}_i = 0, \sum_{i=1}^n x_1 \hat{u}_i = 0, \dots, \sum_{i=1}^n x_k \hat{u}_i = 0$$

Naturally, the estimators of  $\text{Var}(\hat{\beta}_\ell)$  is

$$\hat{\sigma}_{\beta_\ell}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2 / (n - k - 1)}{SST_{x_\ell} (1 - R_\ell^2)}$$

Additionally, the natural estimator of  $\sigma_u$  is

$$\hat{\sigma}_u = \sqrt{\hat{\sigma}_u^2},$$

which is called the **standard error of the regression (SER)**.

Although  $\hat{\sigma}_u^2$  is an unbiased and consistent estimator of  $\sigma_u^2$ ,  $\hat{\sigma}_u$  is consistent but biased.

All together the **standard errors of the coefficients** equal

$$\text{se}(\hat{\beta}_\ell) = \sqrt{\frac{\sum_{i=1}^n \hat{u}_i^2 / (n - k - 1)}{SST_{x_\ell} (1 - R_\ell^2)}}$$

### Gauss-Markov Theorem

Under the Gauss-Markov assumptions, using least squares criteria to estimate the parameters yields the **best linear unbiased estimators (BLUE)** of the parameters, a result known as the **Gauss-Markov Theorem**. Restricting the scope to linear unbiased estimators, we can define an estimator

$$\tilde{\beta}_\ell = \sum_{i=1}^n w_{i\ell} y_i$$

where each  $w_{i\ell}$  can be a function of the sample values of all the independent variables and

$$E[\tilde{\beta}_\ell | x] = \beta_\ell$$

$$E\left[\sum_{i=1}^n w_{i\ell} y_i \middle| x\right] = \beta_\ell$$

$$E\left[\sum_{i=1}^n w_{i\ell} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i) \middle| x\right] = \beta_\ell$$

$$\beta_0 \sum_{i=1}^n w_{i\ell} + \beta_1 \sum_{i=1}^n x_{i1} w_{i\ell} + \cdots + \beta_k \sum_{i=1}^n x_{ik} w_{i\ell} + \sum_{i=1}^n w_{i\ell} E[u_i | x] = \beta_\ell$$

$$E\left[\beta_0 \sum_{i=1}^n w_{i\ell} + \beta_1 \sum_{i=1}^n x_{i1} w_{i\ell} + \cdots + \beta_\ell \sum_{i=1}^n x_{i\ell} w_{i\ell} + \cdots + \beta_k \sum_{i=1}^n x_{ik} w_{i\ell} + \sum_{i=1}^n u_i w_{i\ell} \middle| x\right] = \beta_\ell$$

where  $\sum_{i=1}^n w_{i\ell} E[u_i | x] = 0$  by the zero conditional mean assumption. Thus, it must be the case that

$$\sum_{i=1}^n w_{i\ell} = 0$$

$$\sum_{i=1}^n w_{i\ell} x_{ij} = 0 \quad \forall j \neq \ell$$

$$\sum_{i=1}^n w_{i\ell} x_{i\ell} = 1$$

Using this fact, we can derive the variance of  $\tilde{\beta}_\ell$  as

$$\begin{aligned} \text{Var}(\tilde{\beta}_\ell | x) &= \text{Var}\left(\beta_0 \sum_{i=1}^n w_{i\ell} + \beta_1 \sum_{i=1}^n x_{i1} w_{i\ell} + \cdots + \beta_\ell \sum_{i=1}^n x_{i\ell} w_{i\ell} + \cdots + \beta_k \sum_{i=1}^n x_{ik} w_{i\ell} + \sum_{i=1}^n u_i w_{i\ell} \middle| x\right) \\ &= \text{Var}\left(\beta_\ell + \sum_{i=1}^n u_i w_{i\ell} \middle| x\right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n \text{Var}(u_i w_{i\ell} | x) \\
&= \sum_{i=1}^n w_{i\ell}^2 \text{Var}(u_i | x) \\
&= \sigma_u^2 \sum_{i=1}^n w_{i\ell}^2
\end{aligned}$$

Suppose  $w_{i\ell} = v_{i\ell} + c_{i\ell}$  where  $v_{i\ell} = \frac{\hat{r}_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2}$  (the OLS weights) and  $c_{i\ell}$  is some arbitrary difference from the least squares weights, then

$$\begin{aligned}
\text{Var}(\tilde{\beta}_\ell | x) &= \sigma_u^2 \sum_{i=1}^n w_{i\ell}^2 = \sigma_u^2 \sum_{i=1}^n (v_{i\ell} + c_{i\ell})^2 \\
&= \sigma_u^2 \sum_{i=1}^n v_{i\ell}^2 + \sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 + 2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell} \\
&= \sigma_u^2 \sum_{i=1}^n \left( \frac{\hat{r}_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \right)^2 + \sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 + 2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell} \\
&= \sigma_u^2 \frac{\sum_{i=1}^n \hat{r}_{i\ell}^2}{(\sum_{i=1}^n \hat{r}_{i\ell}^2)^2} + \sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 + 2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell} \\
&= \frac{\sigma_u^2}{\sum_{i=1}^n \hat{r}_{i\ell}^2} + \sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 + 2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell} \\
&= \text{Var}(\hat{\beta}_\ell) + \sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 + 2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell}
\end{aligned}$$

Hence, to show OLS estimators are BLUE under the G-M assumptions, it remains to show that

$$\begin{aligned}
\text{Var}(\hat{\beta}_\ell) &\leq \text{Var}(\tilde{\beta}_\ell) + \sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 + 2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell} \\
&\rightarrow 0 \leq \sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 + 2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell},
\end{aligned}$$

which just requires showing  $2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell} \geq 0$  since  $\sigma_u^2 \sum_{i=1}^n c_{i\ell}^2 \geq 0$ . Finalizing the proof,

$$\begin{aligned}
2\sigma_u^2 \sum_{i=1}^n v_{i\ell} c_{i\ell} &\geq 0 \\
&\rightarrow \sum_{i=1}^n v_{i\ell} c_{i\ell} \geq 0
\end{aligned}$$

Returning to the previous derivations that  $\sum_{i=1}^n w_{i\ell} = 0$ ,  $\sum_{i=1}^n w_{i\ell} x_{ij} = 0 \forall j \neq \ell$ , and  $\sum_{i=1}^n w_{i\ell} x_{i\ell} = 1$ ,

$$\begin{aligned}
0 &= \sum_{i=1}^n w_{i\ell} = \sum_{i=1}^n (v_{i\ell} + c_{i\ell}) = \sum_{i=1}^n v_{i\ell} + \sum_{i=1}^n c_{i\ell} = 0 + \sum_{i=1}^n c_{i\ell} \\
&\rightarrow \sum_{i=1}^n c_{i\ell} = 0
\end{aligned}$$



$$\begin{aligned}
0 &= \sum_{i=1}^n w_{i\ell} x_{ij} = \sum_{i=1}^n (v_{i\ell} + c_{i\ell}) x_{ij} = \sum_{i=1}^n v_{i\ell} x_{ij} + \sum_{i=1}^n c_{i\ell} x_{ij} = 0 + \sum_{i=1}^n c_{i\ell} x_{ij} \\
&\rightarrow \sum_{i=1}^n c_{i\ell} x_{ij} = 0 \quad \forall j \neq \ell \\
1 &= \sum_{i=1}^n w_{i\ell} x_{i\ell} = \sum_{i=1}^n (v_{i\ell} + c_{i\ell}) x_{i\ell} = \sum_{i=1}^n v_{i\ell} x_{i\ell} + \sum_{i=1}^n c_{i\ell} x_{i\ell} = 1 + \sum_{i=1}^n c_{i\ell} x_{i\ell} \\
&\rightarrow \sum_{i=1}^n c_{i\ell} x_{i\ell} = 0
\end{aligned}$$

Thus,

$$\begin{aligned}
\sum_{i=1}^n v_{i\ell} c_{i\ell} &= \sum_{i=1}^n \frac{\hat{r}_{i\ell} c_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} = \frac{\sum_{i=1}^n \hat{r}_{i\ell} c_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
&= \frac{\sum_{i=1}^n (x_{i\ell} - \hat{x}_{i\ell}) c_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
&= \frac{\sum_{i=1}^n x_{i\ell} c_{i\ell} - \sum_{i=1}^n \hat{x}_{i\ell} c_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
&= \frac{0 - \sum_{i=1}^n (\hat{\gamma}_0 + \hat{\gamma}_1 x_{i1} + \cdots + \hat{\gamma}_{\ell-1} x_{i\ell-1} + \hat{\gamma}_{\ell+1} x_{i\ell+1} + \cdots + \hat{\gamma}_k x_{ik}) c_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
&= - \frac{\hat{\gamma}_0 \sum_{i=1}^n c_{i\ell} + \hat{\gamma}_1 \sum_{i=1}^n x_{i1} c_{i\ell} + \cdots + \hat{\gamma}_{\ell-1} \sum_{i=1}^n x_{i\ell-1} c_{i\ell} + \hat{\gamma}_{\ell+1} \sum_{i=1}^n x_{i\ell+1} c_{i\ell} + \cdots + \hat{\gamma}_k \sum_{i=1}^n x_{ik} c_{i\ell}}{\sum_{i=1}^n \hat{r}_{i\ell}^2} \\
&= - \frac{\hat{\gamma}_0 \cdot 0 + \hat{\gamma}_1 \cdot 0 + \cdots + \hat{\gamma}_{\ell-1} \cdot 0 + \hat{\gamma}_{\ell+1} \cdot 0 + \cdots + \hat{\gamma}_k \cdot 0}{\sum_{i=1}^n \hat{r}_{i\ell}^2} = 0
\end{aligned}$$

$\therefore \hat{\beta}_\ell$  is the best linear unbiased estimator of the population parameter  $\beta_\ell$  because any other linear unbiased estimator has at least as much variance as the OLS estimator under the Gauss-Markov assumptions.

## Exercises

### Problems

1. (a) We would expect a negative coefficient on *hsperc* because we would expect those who do better in high school to do better in college as well. The better an individual does in college the larger is *colgpa* and the better one does in high school the smaller is *hsperc*.
- (b) 2.676
- (c) The predicted difference in *colgpa* is 0.2072. In my opinion the difference is fairly large. To others, the difference may seem insignificant.
- (d) Holding *hsperc* fixed, a difference in SAT scores of 337.8378378 leads to a predicted *colgpa* difference of .50.
2. (a) Yes, we would expect *educ* and *sibs* to be negatively correlated because more siblings may restrict the amount of education an individual can receive. To reduce predicted years of education by one year, *sibs* must increase by 10.6382979.
- (b) The coefficient on *meduc* indicates a predicted increase of 0.131 years of schooling for every additional year of schooling the individual's mother received.
- (c) The predicted difference in schooling is 1.364 years.
3. (a) If adults trade off sleep for work,  $\beta_1$  is negative.

- (b) I would guess that  $\beta_2$  is positive and  $\beta_3$  is negative.
- (c) If someone works five more hours per week, *sleep* is predicted to fall by 44.4 minutes.
- (d) The coefficient on *educ* indicates a predicted loss of 11.13 minutes in sleep per week for every additional year of education an individual receives. This tradeoff appears small.
- (e) Not particularly, the  $R^2$  indicates *totwrk*, *educ*, and *age* explain about 11.3% of the variation in *sleep*. Some other factors that might affect time spent sleeping are an individual's health, occupation, and stress levels. It goes without saying that *totwrk* is likely correlated with these other factors.
4. (a) We would expect  $\beta_5 \leq 0$  because we would expect those who attend better schools to earn a higher wage.
- (b) I expect the other slope parameters to be positive. Each of the corresponding dependent variables are indicators of the quality of a school or the quality of a given class, and I would expect those graduating from among more "talented" peers to earn higher wages.
- (c) The predicted ceteris paribus difference in salaries for schools with a median GPA different by one point is 24.8%.
- (d) The coefficient on  $\ln(\text{libvol})$  is an estimate of the elasticity of *salary* with respect *libvol* (or the number of volumes in the law school library). The coefficient indicates a predicted 0.095% increase in *salary* for a 1% increase in the number of volumes in the law school library.
- (e) Yes, I would say it's better to attend higher ranked schools. A difference in ranking of 20 is predicted to impact salary by 6.6%.
5. (a) No. Because the time spent on these activities must sum to 168 hours, changing *study* must change the time spent on at least one of the other activities.
- (b) This model violates the no perfect collinearity assumption because the variables can be expressed perfectly as a linear combination of the other independent variables. For example, *study* can be expressed as  $168 - \text{sleep} - \text{work} - \text{leisure}$ . This same concept applies to all of the explanatory variables.
- (c) Dropping one of the explanatory variables would alleviate the perfect collinearity issue and allow for a useful interpretation.
6. Conditioning on  $x$ :
- (a)  $E[\hat{\theta}_1] = E[\hat{\beta}_1 + \hat{\beta}_2] = E[\hat{\beta}_1] + E[\hat{\beta}_2] = \beta_1 + \beta_2 = \theta_1$
- (b)

$$\begin{aligned}
 \text{Var}(\hat{\theta}_1) &= \text{Var}(\hat{\beta}_1 + \hat{\beta}_2) \\
 &= \text{Var}(\hat{\beta}_1) + 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) + \text{Var}(\hat{\beta}_2) \\
 &= \text{Var}(\hat{\beta}_1) + 2\text{Corr}(\hat{\beta}_1, \hat{\beta}_2) \cdot \sqrt{\text{Var}(\hat{\beta}_1) \text{Var}(\hat{\beta}_2)} + \text{Var}(\hat{\beta}_2)
 \end{aligned}$$

7. Of the provided options, only "(ii) Omitting an important variable" can cause OLS estimators to be biased. Heteroskedasticity and multicollinearity have "negative" implications on the variance of the OLS estimators but do not cause OLS estimators to be biased.
8. Using intuition,  $\beta_1$  and  $\beta_2$  are likely positive as one would expect more trained and able workers are more productive. If *avgtrain* and *avgabil* are negatively correlated, we can derive the bias in  $\hat{\beta}_1$  as

$$\text{Bias}(\tilde{\beta}_1) = \beta_1 + \tilde{\delta}_1\beta_2 - \beta_1 = \tilde{\delta}_1\beta_2,$$

which we would expect to be negative since *avgtrain* and *avgabil* are negatively correlated (i.e.,  $\tilde{\delta}_1 < 0$ ) and  $\beta_2$  is likely positive.

9. (a)  $\beta_1$  is likely negative because more pollution in a community should lead to lower house prices on average.  $\beta_2$  is likely positive as more rooms in a house commands higher house prices on average.  $\beta_1$  is the elasticity of the median housing price in a community with respect to the amount of pollution in a community.
- (b)  $\ln(\text{nox})$  and *rooms* may be negatively correlated because people may choose to build homes with more rooms where there is less pollution. For example, families with children need more rooms and would prefer to be in an area with cleaner air. If there is negative correlation between  $\ln(\text{nox})$  and *rooms*, then  $\text{Bias}(\tilde{\beta}_1) = \beta_1 + \tilde{\delta}_1\beta_2 - \beta_1 = \tilde{\delta}_1\beta_2 < 0$ , meaning the simple regression estimator of  $\beta_1$  will produce a downward biased estimator.
- (c) Yes, this is what we would have expected since the estimated elasticity in the simple regression model is less than that of the multiple regression model. However, this does not necessarily mean that -.718 is definitely closer to the true elasticity than -1.043. It simply means that we expect the method by which the multiple regression coefficient was produced to be closer to the true elasticity.
10. (a) If  $x_1$  is highly correlated with  $x_2$  and  $x_3$  in the sample, and  $x_2$  and  $x_3$  have large partial effects on  $y$ , I would expect  $\tilde{\beta}_1$  and  $\hat{\beta}_1$  to be very different. Solely looking at the case of omitting  $x_2$ , we can derive the bias in  $\tilde{\beta}_1$  as  $\text{Bias}(\tilde{\beta}_1) = \beta_1 + \tilde{\delta}_1\beta_2 - \beta_1 = \tilde{\delta}_1\beta_2$ . A large correlation between  $x_1$  and  $x_2$  indicates  $\tilde{\delta}_1$  should be relatively large, and  $x_2$  having a large partial effect on  $y$  indicates  $\beta_2$  is large. Thus, we can conclude the bias in  $\tilde{\beta}_1$  should be large. Since we know (under the G-M assumptions) that  $\hat{\beta}_1$  is unbiased, I would expect the estimates produced by the separate estimators to be very different.
- (b) If  $x_1$  is almost uncorrelated with  $x_2$  and  $x_3$  but  $x_2$  and  $x_3$  are highly correlated, I would expect  $\tilde{\beta}_1$  to be very similar to  $\hat{\beta}_1$ . A large correlation between  $x_2$  and  $x_3$  doesn't massively impact the bias in  $\tilde{\beta}_1$ . Looking at part (a), we derived the bias in  $\tilde{\beta}_1$  as  $\tilde{\delta}_1\beta_2$  when the bias from omitting  $x_3$  is ignored. If  $x_1$  is almost uncorrelated with  $x_2$  in the population, it's likely  $x_1$  is almost uncorrelated with  $x_2$  in the sample and thus  $\tilde{\delta}_1$  should be quite small and hence the bias in  $\tilde{\beta}_1$  should be quite small. Again, since we know (under the G-M assumptions) that  $\hat{\beta}_1$  is unbiased, I would expect the estimates produced by the separate estimators to be very similar.
- (c) A large correlation between  $x_2$  and  $x_3$  means we're likely to experience a "high" level of multicollinearity in our model. Additionally, since  $x_2$  and  $x_3$  have small partial effects on  $y$ , omitting these variables from the model should not lead to a large increase in the sum of squared residuals. Altogether, we would expect  $\text{se}(\tilde{\beta}_1)$  to be smaller than  $\text{se}(\hat{\beta}_1)$ .
- (d) A small correlation between  $x_2$  and  $x_3$  means we're likely to experience a "low" level of multicollinearity in our model. Additionally, since  $x_2$  and  $x_3$  have large partial effects on  $y$ , omitting these variables from the model will likely lead to a large increase in the sum of squared residuals. Altogether, we would expect  $\text{se}(\tilde{\beta}_1)$  to be smaller than  $\text{se}(\hat{\beta}_1)$ .

11.

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \frac{\sum_{i=1}^n \hat{r}_{i1} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i)}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \frac{\beta_0 \sum_{i=1}^n \hat{r}_{i1} + \beta_1 \sum_{i=1}^n \hat{r}_{i1} x_{i1} + \beta_2 \sum_{i=1}^n \hat{r}_{i1} x_{i2} + \beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \frac{\beta_0 \cdot 0 + \beta_1 \sum_{i=1}^n \hat{r}_{i1} (\hat{r}_{i1} + \hat{x}_{i1}) + \beta_2 \cdot 0 + \beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\beta_1 \sum_{i=1}^n \hat{r}_{i1}^2 + \beta_1 \sum_{i=1}^n \hat{r}_{i1} \hat{x}_{i1} + \beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \beta_1 + \frac{\beta_1 \sum_{i=1}^n \hat{r}_{i1} (\hat{\gamma}_0 + \hat{\gamma}_1 x_{i2}) + \beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \beta_1 + \frac{\beta_1 \cdot 0 + \beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\
&= \beta_1 + \frac{\beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2}
\end{aligned}$$

Thus,

$$\begin{aligned}
E[\tilde{\beta}_1 | x] &= E \left[ \beta_1 + \frac{\beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \middle| x \right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n \hat{r}_{i1}^2} E \left[ \beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} u_i \middle| x \right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n \hat{r}_{i1}^2} \left\{ \beta_3 \sum_{i=1}^n E[\hat{r}_{i1} x_{i3} | x] + \sum_{i=1}^n E[\hat{r}_{i1} u_i | x] \right\} \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n \hat{r}_{i1}^2} \left\{ \beta_3 \sum_{i=1}^n \hat{r}_{i1} x_{i3} + \sum_{i=1}^n \hat{r}_{i1} E[u_i | x] \right\} \\
&= \beta_1 + \beta_3 \frac{\sum_{i=1}^n \hat{r}_{i1} x_{i3}}{\sum_{i=1}^n \hat{r}_{i1}^2}
\end{aligned}$$

12. (a) We must omit one of the tax share variables from the equation to avoid violating the no perfect collinearity assumption since the four shares add up to one.
- (b)  $\beta_1$  is the constant partial effect of *share<sub>P</sub>* on *growth*. By itself,  $\beta_1$  doesn't have much meaning because increasing a non-zero proportion by one has no meaning; however,  $\beta_1/100$  can be interpreted as the partial effect of increasing the share of property taxes in total tax revenue by 1 percentage point on *growth*.
13. (a)  $\tilde{\beta}_1$  is linear because it can be expressed in the form  $\tilde{\beta}_1 = \sum_{i=1}^n w_{i1} y_i$ , where  $w_{i1} = \frac{z_i - \bar{z}}{\sum_{i=1}^n (z_i - \bar{z}) x_i}$ .  $\tilde{\beta}_1$  can be narrowed down to

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_{i=1}^n (z_i - \bar{z}) y_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \\
&= \frac{\sum_{i=1}^n (z_i - \bar{z}) (\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \\
&= \frac{\beta_0 \sum_{i=1}^n (z_i - \bar{z}) + \beta_1 \sum_{i=1}^n (z_i - \bar{z}) x_i + \sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \\
&= \beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i}
\end{aligned}$$

Taking the conditional expectation:

$$\begin{aligned}
E[\tilde{\beta}_1 | x] &= E \left[ \beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \middle| x \right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \sum_{i=1}^n E[(z_i - \bar{z}) u_i | x]
\end{aligned}$$

$$\begin{aligned}
&= \beta_1 + \frac{1}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \sum_{i=1}^n (z_i - \bar{z}) E[u_i|x] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \sum_{i=1}^n (z_i - \bar{z}) \cdot 0 = \beta_1
\end{aligned}$$

(b) Using the previous derivation,

$$\begin{aligned}
\text{Var}(\tilde{\beta}_1|x) &= \text{Var}\left(\beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \middle| x\right) \\
&= \frac{1}{\left(\sum_{i=1}^n (z_i - \bar{z}) x_i\right)^2} \text{Var}\left(\sum_{i=1}^n (z_i - \bar{z}) u_i \middle| x\right) \\
&= \frac{1}{\left(\sum_{i=1}^n (z_i - \bar{z}) x_i\right)^2} \sum_{i=1}^n \text{Var}((z_i - \bar{z}) u_i | x) \\
&= \frac{1}{\left(\sum_{i=1}^n (z_i - \bar{z}) x_i\right)^2} \sum_{i=1}^n (z_i - \bar{z})^2 \text{Var}(u_i | x) \\
&= \frac{\sigma^2 \left(\sum_{i=1}^n (z_i - \bar{z})^2\right)}{\left(\sum_{i=1}^n (z_i - \bar{z}) x_i\right)^2}
\end{aligned}$$

(c) Under the G-M assumptions,

$$\text{Var}(\hat{\beta}_1|x) = \frac{\sigma^2}{SST_x}$$

Thus, we must show

$$\begin{aligned}
\frac{\sigma^2}{SST_x} &\leq \frac{\sigma^2 \left(\sum_{i=1}^n (z_i - \bar{z})^2\right)}{\left(\sum_{i=1}^n (z_i - \bar{z}) x_i\right)^2} \\
&\rightarrow \frac{1}{SST_x} \leq \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\left(\sum_{i=1}^n (z_i - \bar{z}) x_i\right)^2} \\
&\rightarrow \frac{1}{SST_x} \leq \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\left(\sum_{i=1}^n (z_i - \bar{z}) (x_i - \bar{x})\right)^2}
\end{aligned}$$

From the Cauchy-Schwartz inequality, we know that  $\left(\sum_{i=1}^n (z_i - \bar{z}) (x_i - \bar{x})\right)^2 \leq \sum_{i=1}^n (z_i - \bar{z})^2 \sum_{i=1}^n (x_i - \bar{x})^2$

$$\begin{aligned}
&\rightarrow \frac{1}{SST_x} \leq \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\sum_{i=1}^n (z_i - \bar{z})^2 \sum_{i=1}^n (x_i - \bar{x})^2} \\
&\rightarrow \frac{1}{SST_x} \leq \frac{1}{SST_x} \blacksquare
\end{aligned}$$

14. The standard error of  $\tilde{\beta}_1$  can be rewritten as

$$\text{se}(\tilde{\beta}_1) = \frac{\tilde{\sigma}}{\sqrt{SST_1}} = \sqrt{\frac{\sum_{i=1}^n \tilde{u}_i^2 / n - 2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}}$$

Similarly the standard error of  $\hat{\beta}_1$  can be rewritten as

$$\text{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_1}} \cdot \sqrt{\text{VIF}_1} = \sqrt{\frac{\sum_{i=1}^n \hat{u}_i^2 / n - 3}{(1 - R_1^2) \sum_{i=1}^n (x_{i1} - \bar{x}_1^2)}}$$

Thus, for a large sample size, the two important factors in determining the size of one standard error to the other is the sums of the squared residuals and the magnitude of  $R_1^2$ . If  $x_2$  does not have a large partial effect on  $y$  but is highly correlated with  $x_1$ , we would expect  $\text{se}(\tilde{\beta}_1) < \text{se}(\hat{\beta}_1)$ . On the other hand, if  $x_2$  has a large partial effect on  $y$  but has a small correlation with  $x_1$ , we would expect  $\text{se}(\tilde{\beta}_1) > \text{se}(\hat{\beta}_1)$ .

15. (a) There are 351 degrees of freedom in the first regression and 350 in the second regression. The SER is smaller in the second regression because the SSR is 127.721 smaller than that of the first regression, which outweighs the loss of an extra degree of freedom.
- (b) A sample correlation coefficient of about 0.487 between *years* and *rbisyr* makes sense because those who bat more runs in tend to better players, and better players tend to play in the major leagues longer. The variance inflation factor for the slope coefficients in the multiple regression is  $\frac{1}{1-0.487^2} \approx 1.3109063$ . Personally, I would say there is a moderate level of collinearity between *years* and *rbisyr*.
- (c) The additional SSR in the first regression to that of the second regression is 127.721, which outweighs the "cost" induced from the collinearity between *years* and *rbisyr*.
16. The sample size is smaller when *age* is added to the equation, which means it's no longer guaranteed that the  $R^2$  increases by adding an additional explanatory variable to the equation.
17. The estimated percentage change in *wage* from getting one more year of schooling is 9.4%.

## Computer Exercises

```
#i
#beta 2 is likely positive

#ii
#Yes, cigs and faminc are likely correlated
#They may be negatively correlated because one may expect
#families who smoke more to earn less
#On the other hand, they may be positively correlated
#since wealthier families can afford more cigarettes
```

```
#iii
summary(lm(bwght ~ cigs + faminc, data = bwght))
```

Call:

```
lm(formula = bwght ~ cigs + faminc, data = bwght)
```

Residuals:

Min	1Q	Median	3Q	Max
-96.061	-11.543	0.638	13.126	150.083

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	116.97413	1.04898	111.512	< 0.0000000000000002 ***
cigs	-0.46341	0.09158	-5.060	0.000000475 ***
faminc	0.09276	0.02919	3.178	0.00151 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.06 on 1385 degrees of freedom

Multiple R-squared: 0.0298, Adjusted R-squared: 0.0284

F-statistic: 21.27 on 2 and 1385 DF, p-value: 0.0000000007942

```
summary(lm(bwght ~ cigs, data = bwght))
```

Call:

```
lm(formula = bwght ~ cigs, data = bwght)
```

Residuals:

Min	1Q	Median	3Q	Max
-96.772	-11.772	0.297	13.228	151.228

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	119.77190	0.57234	209.267	< 0.0000000000000002 ***
cigs	-0.51377	0.09049	-5.678	0.0000000166 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.13 on 1386 degrees of freedom

Multiple R-squared: 0.02273, Adjusted R-squared: 0.02202

F-statistic: 32.24 on 1 and 1386 DF, p-value: 0.00000001662

```
lm1 <- lm(price ~ sqrft + bdrms, data = hprice1)
```

```
#i
```

```
lm1
```

Call:

```
lm(formula = price ~ sqrft + bdrms, data = hprice1)
```

Coefficients:

(Intercept)	sqrft	bdrms
-19.3150	0.1284	15.1982

```
#ii
```

```
lm1$coefficients[3]*1000
```

```
bdrms
15198.19
```

```
#iii
```

```
(lm1$coefficients[3] + lm1$coefficients[2]*140)*1000
```

```
bdrms
33179.26
```

```
#iv
```

```
100*summary(lm1)$r.squared
```

```
[1] 63.19184
```

```
#v
```

```
lm1$fitted.values[1]*1000
```

```
1
354605.2
```

```
#vi
```

```
lm1$residuals[1]*1000
```

-54605.25

*#This suggests the buyer underpaid for the house*

```
rm(lm1)
```

*#i*

```
lm(lsalary ~ lsales + lmktval, data = ceosal2)
```

Call:

```
lm(formula = lsalary ~ lsales + lmktval, data = ceosal2)
```

Coefficients:

(Intercept)	lsales	lmktval
4.6209	0.1621	0.1067

*#ii*

```
lm2 <- lm(lsalary ~ lsales + lmktval + profits, data = ceosal2)
```

*#profits cannot be in logarithmic form because of negative profits*

```
lm2
```

Call:

```
lm(formula = lsalary ~ lsales + lmktval + profits, data = ceosal2)
```

Coefficients:

(Intercept)	lsales	lmktval	profits
4.68692438	0.16136826	0.09752857	0.00003566

```
summary(lm2)$r.squared
```

```
[1] 0.2993366
```

*#iii*

```
lm3 <- lm(lsalary ~ lsales + lmktval + profits + ceoten, data = ceosal2)
```

```
lm3$coefficients[5] * 100
```

```
ceoten
```

```
1.168467
```

*#iv*

```
cor(ceosal2$lmktval, ceosal2$profits)
```

```
[1] 0.7768976
```

*#This tells us the OLS estimators may have large variances*

```
rm(lm2, lm3)
```

*#i*

*#atndrte*

```
min(attend$atndrte)
```

```
[1] 6.25
```

```
max(attend$atndrte)
```

```
[1] 100
```

```
mean(attend$atndrte)
```

```
[1] 81.70956
```



```
#priGPA  
min(attend$priGPA)
```

```
[1] 0.857
```

```
max(attend$priGPA)
```

```
[1] 3.93
```

```
mean(attend$priGPA)
```

```
[1] 2.586775
```

```
#ACT  
min(attend$ACT)
```

```
[1] 13
```

```
max(attend$ACT)
```

```
[1] 32
```

```
mean(attend$ACT)
```

```
[1] 22.51029
```

```
#ii  
lm1 <- lm(atndrte ~ priGPA + ACT, data = attend)  
lm1
```

Call:

```
lm(formula = atndrte ~ priGPA + ACT, data = attend)
```

Coefficients:

(Intercept)	priGPA	ACT
75.700	17.261	-1.717

```
#The intercept indicates a predicted attendance rate of 75.7%  
#for a student whose prior GPA is zero and ACT score is zero  
#This clearly doesn't have much meaning
```

```
#iii  
#The coefficient on priGPA indicates a predicted 17.3 percentage  
#points higher attendance rate for every point increase in prior GPA  
#The coefficient on ACT indicates a predicted 1.7 percentage  
#points lower attendance rate for every additional point on the ACT  
#The negative slope coefficient on ACT is a little surprising
```

```
#iv  
unname(lm1$coefficients[1]+lm1$coefficients[2]*3.65+lm1$coefficients[3]*20)
```

```
[1] 104.3705
```

```
sum(near(attend$priGPA,3.65,.001) & attend$ACT==20)
```

```
[1] 1
```

```
#v  
unname(lm1$coefficients[2]+lm1$coefficients[3]*-5)
```

```
[1] 25.84336
```

```
rm(lm1)
```

```
lm1 <- lm(educ ~ exper + tenure, data = wage1)
lm(lwage ~ lm1$residuals, data = wage1)
```

Call:

```
lm(formula = lwage ~ lm1$residuals, data = wage1)
```

Coefficients:

```
(Intercept)  lm1$residuals
      1.62327         0.09203
```

```
lm(lwage ~ educ + exper + tenure, data = wage1)
```

Call:

```
lm(formula = lwage ~ educ + exper + tenure, data = wage1)
```

Coefficients:

```
(Intercept)      educ      exper      tenure
    0.284360    0.092029    0.004121    0.022067
```

```
rm(lm1)
```

*#i*

```
lm1 <- lm(IQ ~ educ, data = wage2)
lm1$coefficients[2]
```

```
      educ
3.533829
```

*#ii*

```
lm2 <- lm(lwage ~ educ, data = wage2)
lm2$coefficients[2]
```

```
      educ
0.0598392
```

*#iii*

```
lm3 <- lm(lwage ~ educ + IQ, data = wage2)
lm3$coefficients[2]
```

```
      educ
0.0391199
```

```
lm3$coefficients[3]
```

```
      IQ
0.005863131
```

*#iv*

```
unname(near(lm2$coefficients[2], lm3$coefficients[2] + lm3$coefficients[3] * lm1$coefficients[2], .001))
```

```
[1] TRUE
```

```
rm(lm1, lm2, lm3)
```

*#i*

```
summary(lm(math10 ~ lexpend + lnchprg, data = meap93))
```

Call:

```
lm(formula = math10 ~ lexpend + lnchprg, data = meap93)
```

Residuals:

Min	1Q	Median	3Q	Max
-24.294	-6.172	-1.293	4.855	43.203

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-20.36075	25.07288	-0.812	0.4172
lexpend	6.22969	2.97263	2.096	0.0367 *
lnchprg	-0.30459	0.03536	-8.614	<0.0000000000000002 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.526 on 405 degrees of freedom

Multiple R-squared: 0.1799, Adjusted R-squared: 0.1759

F-statistic: 44.43 on 2 and 405 DF, p-value: < 0.00000000000000022

*#ii*

*#The intercept predicts a pass rate when log(expend) and lnchprg=0*

*#This doesn't make sense though because setting log(expend)=0*

*#is the same as setting expend=1*

*#iii*

```
lm(math10 ~ lexpend, data = meap93)
```

Call:

```
lm(formula = math10 ~ lexpend, data = meap93)
```

Coefficients:

(Intercept)	lexpend
-69.34	11.16

*#The estimated spending effect is now larger*

*#iv*

```
cor(meap93$lexpend, meap93$lnchprg)
```

```
[1] -0.1927042
```

*#v*

*#The result in part iv indicates a negative correlation between*

*#spending and the lunch program. Additionally, part (i) shows a*

*#negative correlation between the lunch program and the pass rate*

*#Thus, the simple regression coefficient likely overstates the*

*#spending effect as shown by the difference in the coefficient estimates*

*#i*

*#prpblck*

```
mean(discrim$prpblck, na.rm = T)
```

```
[1] 0.1134864
```

```
sd(discrim$prpblck, na.rm = T)
```

```
[1] 0.1824165
```

```
#income
mean(discrim$income, na.rm = T)
```

```
[1] 47053.78
```

```
sd(discrim$income, na.rm = T)
```

```
[1] 13179.29
```

```
#prpblck is a proportion
#income is measured in dollars
```

```
#ii
summary(lm(psoda ~ prpblck + income, data = discrim))
```

Call:

```
lm(formula = psoda ~ prpblck + income, data = discrim)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.29401	-0.05242	0.00333	0.04231	0.44322

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.9563196258	0.0189920097	50.354	< 0.0000000000000002 ***
prpblck	0.1149881907	0.0260006361	4.423	0.0000126 ***
income	0.0000016027	0.0000003618	4.430	0.0000122 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08611 on 398 degrees of freedom

(9 observations deleted due to missingness)

Multiple R-squared: 0.06422, Adjusted R-squared: 0.05952

F-statistic: 13.66 on 2 and 398 DF, p-value: 0.000001835

```
#The coefficient on prpblck indicates a predicted increase
#in the price of soda by about 11 cents for a 100 percentage
#point increase in the proportion of black citizens
```

```
#iii
lm(psoda ~ prpblck, data = discrim)
```

Call:

```
lm(formula = psoda ~ prpblck, data = discrim)
```

Coefficients:

(Intercept)	prpblck
1.03740	0.06493

```
#The discrimination effect is larger when you control for income
```

```
#iv
lm(lpsoda ~ prpblck + lincome, data = discrim)
```

Call:

```
lm(formula = lpsoda ~ prpblck + lincome, data = discrim)
```

Coefficients:

(Intercept)	prpblck	lincome
-0.79377	0.12158	0.07651

*#If prpblck increases by .20, the estimated percentage change in  
#psoda is 2.4316%*

*#v*

```
lm(lpsoda ~ prpblck + lincome + prppov, data = discrim)
```

Call:

```
lm(formula = lpsoda ~ prpblck + lincome + prppov, data = discrim)
```

Coefficients:

(Intercept)	prpblck	lincome	prppov
-1.46333	0.07281	0.13696	0.38036

*#The coefficient on prpblck falls*

*#vi*

```
cor(discrim$lincome, discrim$prppov, use = "pairwise.complete.obs")
```

```
[1] -0.838467
```

*#Yes, I would expect a strong negative correlation between income  
#and the proportion of impoverished citizens*

*#vii*

*#The statement displays a misunderstanding of multicollinearity  
#Adding both to the regression may cause the variances of the estimators  
#to be large, but it may also cause the variances to fall. On the other  
#hand, including both reduces the amount of bias in the estimators*

*#i*

```
summary(lm(gift ~ mailsyear + giftlast + propresp, data = charity))
```

Call:

```
lm(formula = gift ~ mailsyear + giftlast + propresp, data = charity)
```

Residuals:

Min	1Q	Median	3Q	Max
-52.893	-7.050	-3.650	1.397	241.206

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.551518	0.803034	-5.668	0.0000000154094 ***
mailsyear	2.166259	0.331927	6.526	0.0000000000753 ***
giftlast	0.005927	0.001432	4.138	0.0000357699351 ***
propresp	15.358605	0.874539	17.562	< 0.0000000000000002 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.43 on 4264 degrees of freedom

Multiple R-squared: 0.08336, Adjusted R-squared: 0.08271

F-statistic: 129.3 on 3 and 4264 DF, p-value: < 0.00000000000000022

```
summary(lm(gift ~ mailsyear, data = charity))
```

Call:

```
lm(formula = gift ~ mailsyear, data = charity)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.287	-7.976	-5.976	2.687	245.999

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.0141	0.7395	2.724	0.00648 **
mailsyear	2.6495	0.3431	7.723	0.000000000000014 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.96 on 4266 degrees of freedom

Multiple R-squared: 0.01379, Adjusted R-squared: 0.01356

F-statistic: 59.65 on 1 and 4266 DF, p-value: 0.00000000000001404

*#The R-squared increases from about .01 to about .08*

*#ii*

*#The multiple regression equation predicts one more mailing per year*

*#increases gifts by 2.17 guilders.*

*#The simple regression equation predicts one more mailing per year*

*#increases gifts by 2.65 guilders, a larger prediction.*

*#iii*

*#Because propresp is a proportion and thus a 100 percentage point*

*#increase offers little meaning (unless starting at zero), we can*

*#instead interpret the coefficient as a predicted increase in gifts*

*#by .15358605 guilders for every percentage point increase in propresp*

*#iv*

```
lm(gift ~ mailsyear + giftlast + propresp + avggift, data = charity)
```

Call:

```
lm(formula = gift ~ mailsyear + giftlast + propresp + avggift,  
    data = charity)
```

Coefficients:

(Intercept)	mailsyear	giftlast	propresp	avggift
-7.3278	1.2012	-0.2609	16.2046	0.5269

*#The coefficient on mailsyear falls from 2.166259 to 1.2012*

*#v*

*#The coefficient on giftlast falls from 0.005927 to -0.2609*

*#Clearly, controlling for the average gift size, it's apparent that*

*#the current gift amount and the most recent gift amount are negatively*

*#correlated*

## Chapter 4

### Notes

#### Classical Linear Model

In addition to the Gauss-Markov assumptions, there is an additional assumption commonly made called the **normality assumption**. Under the normality assumption, we assume the population error  $u$  is *independent* of the explanatory variables  $x_1, x_2, \dots, x_k$  and is normally-distributed with zero mean and variance  $\sigma_u^2$ :  $u \sim \mathcal{N}(0, \sigma_u^2)$ . It should be obvious that this assumption is stronger than any of the previous assumptions and that the exogeneity and homoskedasticity assumptions hold by the independence assumptions.

All together, the normality and Gauss-Markov assumptions are called the **classical linear model (CLM) assumptions**. Additionally, a model under these six assumptions are referred to as the **classical linear model**. Under the CLM assumptions, the OLS estimators  $\hat{\beta}_0, \dots, \hat{\beta}_k$  are the **minimum variance unbiased estimators**, meaning OLS estimators have the smallest variance among all unbiased estimators (not just linear unbiased estimators as in the G-M theorem). More succinctly, the population assumptions of CLM can be summarized by

$$y|x \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k, \sigma_u^2)$$

The argument justifying the normal distribution for the errors usually invokes the CLT (because  $u$  is the sum of many different unobserved factors affecting  $y$ ) to suggest that  $u$  has an approximate normal distribution. This argument has some merit, but it is not without weaknesses. The two main weaknesses are that the factors in  $u$  can have very different distributions in the population, which may lead to “poor” normal distributions depending on how many factors appear in  $u$  and how different their distributions are, and if  $u$  is a complicated function of the unobserved factors, then the CLT argument doesn’t really apply. Sometimes, using a transformation, such as taking the  $\ln$ , yields a distribution that is closer to normal, but in other cases, normality cannot be reasonably assumed. In any application, whether normality of  $u$  can be assumed is really an empirical matter. Nevertheless, we’ll see later that nonnormality of the errors is not a serious problem with large sample sizes.

The important implication of the CLM assumptions is the following theorem: Under the CLM assumptions,

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma_u^2)$$

Therefore, standardizing  $\hat{\beta}_j$ ,

$$\frac{\hat{\beta}_j - \beta_j}{\text{sd}(\hat{\beta}_j)} \sim \mathcal{N}(0, 1)$$

#### Hypothesis Testing about a Single Population Parameter: The $t$ Test

Directly following the previous theorem, under the CLM assumptions:

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1}$$

This theorem ultimately allows for testing hypotheses about population parameters. In econometrics, if we are interested in whether an explanatory variable has an effect on the dependent variable, we state the **null hypothesis** as

$$H_0 : \beta_j = 0$$

and the corresponding **t statistic** or **t ratio** of  $\hat{\beta}_j$  as

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

In hypothesis testing, one must also come up with a relevant **alternative hypothesis**. A **one-sided alternative** to  $B_j = 0$  takes the form

$$H_1 : \beta_j > 0$$

or

$$H_1 : \beta_j < 0$$

To decide on a **rejection rule**, one must decide on a **significance level** or a probability of committing a type I error (rejecting a true null hypothesis). After deciding on a rejection rule, one can obtain a **critical value**,  $c$ , from the  $t$  distribution (or the standard normal distribution for a sufficiently large sample size). For example, if our alternative hypothesis is  $\beta_j > 0$ , we reject the null hypothesis if

$$t_{\hat{\beta}_j} > c$$

and *fail to reject* the null hypothesis if

$$t_{\hat{\beta}_j} \leq c$$

Similarly, a **two-sided alternative** to  $B_j = 0$  takes the form

$$H_1 : \beta_j \neq 0$$

in which case we reject the null in favor of the alternative hypothesis if

$$|t_{\hat{\beta}_j}| > c$$

Generally, if we reject the null in favor of the alternative hypothesis (for testing against  $\beta_j = 0$ ) we say that  $x_j$  is **statistically significant** or **statistically different from 0** at the specified significance level. Otherwise, we say  $x_j$  is **statistically insignificant**.

Often, it's useful to compute a **p-value**, which provides the *smallest* significance level at which the null hypothesis would be rejected. In other words, the p-value is the probability of observing a t statistic as extreme as we did if the null hypothesis is true and is equal to

$$P(|T| > |t|),$$

where  $T$  denotes a t distributed random variable with  $n - k - 1$  degrees of freedom and  $t$  denotes the numerical value of the test statistic. It's important to note that these statements for the p-value hold for two-sided tests. For one-sided tests the p-value equals

$$P(T > |t|) = P(|T| > |t|) / 2$$

The CLM assumptions also make it simple to construct a **confidence interval**, which is a rule used to construct a random interval so that confidence level percentage of all data sets yields an interval that contains the population value. Confidence intervals are still dependent on the underlying assumptions, and, if these assumptions hold, take the form

$$CI = \left[ \hat{\beta}_j - c \cdot se(\hat{\beta}_j), \hat{\beta}_j + c \cdot se(\hat{\beta}_j) \right]$$

### Hypothesis Testing about a Single Linear Combination of the Parameters

It's often the case where we want to compare the coefficients on the explanatory variables. In this case, we can write the hypotheses

$$H_0 : \beta_j = \beta_\ell$$

$$H_1 : \beta_j \neq \beta_\ell$$

Thus, we can construct the t statistic as

$$t = \frac{\hat{\beta}_j - \hat{\beta}_\ell}{se(\hat{\beta}_j - \hat{\beta}_\ell)}$$



The numerator is trivial to compute; however, the standard error in the denominator is somewhat difficult to compute. We know

$$\text{Var}(\hat{\beta}_j - \hat{\beta}_\ell) = \text{Var}(\hat{\beta}_j) + \text{Var}(\hat{\beta}_\ell) - 2\text{Cov}(\hat{\beta}_j, \hat{\beta}_\ell)$$

The standard deviation of  $\hat{\beta}_j - \hat{\beta}_\ell$  is just the square root of the above equation and  $\left[\text{se}(\hat{\beta}_j)\right]^2$  and  $\left[\text{se}(\hat{\beta}_\ell)\right]^2$  are unbiased estimators of  $\text{Var}(\hat{\beta}_j)$  and  $\text{Var}(\hat{\beta}_\ell)$ . Thus we have,

$$\text{se}(\hat{\beta}_j - \hat{\beta}_\ell) = \sqrt{\left[\text{se}(\hat{\beta}_j)\right]^2 + \left[\text{se}(\hat{\beta}_\ell)\right]^2 - 2\hat{\sigma}_{\hat{\beta}_j, \hat{\beta}_\ell}}$$

An alternative to the approach involves rewriting the model. If our model starts as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_\ell x_\ell + \cdots + \beta_k x_k + u \\ \rightarrow y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_\ell x_\ell + \cdots + \beta_k x_k + u + \beta_\ell x_j - \beta_\ell x_j \\ \rightarrow y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j - \beta_\ell x_j + \cdots + \beta_\ell x_\ell + \beta_\ell x_j + \cdots + \beta_k x_k + u \\ \rightarrow y &= \beta_0 + \beta_1 x_1 + \cdots + (\beta_j - \beta_\ell) x_j + \cdots + \beta_\ell (x_\ell + x_j) + \cdots + \beta_k x_k + u \\ \rightarrow y &= \beta_0 + \beta_1 x_1 + \cdots + (\beta_j - \beta_\ell) x_j + \cdots + \beta_\ell (x_j + x_\ell) + \cdots + \beta_k x_k + u \end{aligned}$$

Creating a new variable  $x_j + x_\ell$ , it's easy to estimate the coefficient on  $x_j$  and evaluate the desired t statistic.

### Hypothesis Testing Multiple Linear Restrictions: The $F$ Test

In many cases, we are interested in whether a group of variables has no effect on the dependent variable. In these cases, we state the hypotheses as

$$H_0 : \beta_{k-q+1} = 0, \dots, \beta_k = 0$$

$$H_1 : H_0 \text{ is not true}$$

The null hypothesis consists of  $q$  **exclusion restrictions**. This exemplifies **multiple restrictions** because we are putting more than one restriction on the parameters. Thus, such a test in which multiple restrictions are tested is called a **multiple hypotheses test** or a **joint hypotheses test**. In multiple hypotheses tests, the alternative holds if at least one of  $\beta_{k-q+1}, \dots, \beta_k$  is different from zero.

In the context of multiple hypothesis testing, we define two models. One model is the **restricted model**, which takes the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$$

And the other model is the **unrestricted model**, which takes the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Using these two models, we can construct the **F statistic**, which is defined by

$$F \equiv \frac{(\text{SSR}_r - \text{SSR}_{ur})/q}{\text{SSR}_{ur}/(n - k - 1)} = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)},$$

where the  $ur$  denotes the unrestricted model and the  $r$  denotes the restricted model. The second equality is called the **R-squared form of the F statistic**.

It should immediately be obvious that  $F \geq 0$  because  $\text{SSR}_{ur} \leq \text{SSR}_r$ . The easiest way to remember where the SSRs appear is to think of  $F$  as measuring the relative increase in SSR when moving from the unrestricted to the restricted model.

The difference in SSRs in the numerator of  $F$  is divided by  $q$ , which is the number of restrictions imposed in moving from the unrestricted to the restricted model ( $q$  independent variables are dropped). Thus,  $q = \text{numerator df} = df_r - df_{ur}$

The SSR in the denominator of  $F$  is divided by the degrees of freedom in the unrestricted model which equals  $n - k - 1 = \text{denominator df} = df_{ur}$ . It should be noted that the denominator of  $F$  is simply the unbiased estimator of  $\sigma_u^2$  in the unrestricted model.

Under the null hypothesis and the CLM assumptions,  $F$  is distributed as an  $F$  random variable with  $(q, n - k - 1)$  degrees of freedom. This can be written as

$$F \sim F_{q, n-k-1}$$

Similar to the case of the  $t$  test, after a significance level is decided on, we reject  $H_0$  in favor of  $H_1$  if

$$F > c$$

and fail to reject  $H_0$  otherwise. If  $H_0$  is rejected, then we say that  $x_{k-q+1}, \dots, x_k$  are **jointly statistically significant** at the appropriate significance level. If the null is not rejected, then the variables are **jointly insignificant**, which often justifies dropping them from the model.

One should be careful in interpreting joint hypotheses using  $t$  statistics on individual parameters. In many cases, using the  $F$  statistic and individually evaluating  $t$  statistics yields different results. This generally happens when there is multicollinearity that prevents uncovering statistically significant partial effects of individual explanatory variables. The  $F$  statistic tests whether explanatory variables are *jointly* significant, and multicollinearity between explanatory variables is much less relevant for testing this hypothesis. Conversely, it is also possible that, in a group of several explanatory variables, one variable has a significant  $t$  statistic but the group of variables is jointly insignificant at the usual significance levels. Generally, such cases occur when a bunch of insignificant variables are grouped with a significant variable, leading to joint insignificance.