The Classical Model

Gauss-Markov Theorem,
Specification, Endogeneity

Properties of Least Squares Estimators

- Here's the model:

$$\boxed{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i}$$

- For the case with 1 regressor and 1 constant, I showed some conditions under which the OLS estimator of the parameters of this model is unbiased, and I gave its variance.
- I asserted that unbiasedness goes through with more regressors.
- I asserted that the variance of the estimated parameters can be calculated with more regressors.
- It turns out that the OLS estimator is BLUE.
  - There is a set of 6 assumptions, called the **Classical Assumptions**. If they are satisfied, then the ordinary least squares estimators is "best" among all linear estimators.
  - "best" means **minimum variance** in a particular class of estimators.

The Classical Assumptions

1. The regression model is **linear in the coefficients**, **correctly specified**, and has an **additive error term**.
2. The error term has **zero population mean**: $E(\varepsilon_i) = 0$.
3. All independent variables are **uncorrelated with the error term**: $Cov(X_i,\varepsilon_i) = 0$ for each independent variable X_i .
4. Errors are uncorrelated across observations: $Cov(\varepsilon_i,\varepsilon_j) = 0$ for two observations *i* and *j* (no **serial correlation**).
5. The error term has constant variance: $Var(\varepsilon_i) = \sigma^2$ for every *i* (no **heteroskedasticity**).
6. No independent variable is a **perfect linear function** of any other independent variable (no **perfect multi-collinearity**).
7. The error terms are normally distributed. *We'll consider all the others, and see what we get. Then, we'll add this one.*

Assumption 1: Linearity, Correct Specification, and Additive Error

- We've already discussed linearity in the coefficients.
- Remember that we wrote the regression model as:

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i \qquad (1)$$

- Assumption 1 says three things:

1. *The regression model has an additive error term.* That is, we can write the regression model like (1)
   - That's the additive error part. This is not very restrictive: we can **always** write the regression model this way if we **define** the error as $\varepsilon_i = Y_i - E[Y_i \mid X_i]$
2. *The regression is linear in parameters.* That is, $E[Y_i \mid X_i]$ is **really** linear in parameters
   - example: $E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i + \beta_2 (X_i)^2$
3. *The regression is correctly specified.* That is, we have the correct **functional form** for $E[Y_i \mid X_i]$,
   - $E[Y_i \mid X_i]$ is not only linear in parameters, but we have all the right *X's* on the right hand side, we squared them if they should be squared, we took logarithms if they should be in logarithms, etc.

Assumption 2: $E(\varepsilon_i)=0$

- This is a pretty weak assumption
- All it says is that there is no expected error in the regression function.
- If we **expected** a particular error value (e.g., if $E(\varepsilon_i) = 5$), then (part of) the error term would be predictable, and we could just add that to the regression model.
- Example: suppose $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ and $E(\varepsilon_i) = 5$. Then

  $E(Y_i|X_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i + 5$

  We could just define a new intercept $\beta_0^* = \beta_0 + 5$ and a new error term $\varepsilon_i^* = \varepsilon_i - 5$. Then we have a new regression model:

  $$Y_i = \beta_0^* + \beta_1 X_i + \varepsilon_i^*$$

  that satisfies Assumption 2.

Assumption 3: $Cov(X_i,\varepsilon_i) = 0$

- We need assumption 3 to be satisfied for **all** the independent variables X_i
- When assumption 3 is satisfied, we say X_i is **exogenous.**
- When assumption 3 is violated, we say X_i is **endogenous.**
- Why is endogeneity a problem?
- Remember we can't observe ε_i.
- If $Cov(X_i,\varepsilon_i) \neq 0$ and X_i is in our model, then OLS attributes variation in Y_i to X_i that is really due to ε_i varying with X_i
- That is, Y moves around when ε moves around. Our estimator **should not** "explain" this variation in Y using X, because it is due to the error, not due to X.
- But if $Cov(X,\varepsilon) \neq 0$, then when ε moves around, so does X. We see X and Y moving together, and the least squares estimator therefore "explains" some of this variation in Y using X. *But really the variation comes from* ε.
- Consequently, we get a biased estimate of the coefficient on X, i.e. β, because it measures the effect of X **and** ε on Y.
- How do we know that assumption 3 is satisfied? We rely on **economic theory** (and some common sense) to tell us that our independent variables are exogenous (there are also some tests available, but they are not very convincing).

Unbiasedness

- If Assumptions 1 – 3 are satisfied, then the least squares estimator of the regression coefficients is **unbiased**.
- Suppose we have the simple linear regression: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ then we can write the least squares estimator of β_1 as:

$$\hat{\beta}_1 = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_i \left(X_i - \overline{X}\right)^2} = \frac{\sum_i \left(X_i - \overline{X}\right)\left(\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 - \beta_1 \overline{X} - \overline{\varepsilon}\right)}{\sum_i \left(X_i - \overline{X}\right)^2}$$

$$= \frac{\sum_i \left(X_i - \overline{X}\right)\left(\beta_1\left(X_i - \overline{X}\right) + \varepsilon_i - \overline{\varepsilon}\right)}{\sum_i \left(X_i - \overline{X}\right)^2} = \frac{\beta_1 \sum_i \left(X_i - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2} + \frac{\sum_i \left(X_i - \overline{X}\right)\left(\varepsilon_i - \overline{\varepsilon}\right)}{\sum_i \left(X_i - \overline{X}\right)^2}$$

$$= \beta_1 + \frac{\sum_i \left(X_i - \overline{X}\right)\left(\varepsilon_i - \overline{\varepsilon}\right)}{\sum_i \left(X_i - \overline{X}\right)^2} \Rightarrow E\left(\hat{\beta}_1\right) = \beta_1$$

Assumptions 4,5: $Cov(\varepsilon_i,\varepsilon_j) = 0$ and $Var(\varepsilon_i) = \sigma^2$

- If these assumptions are violated, we say the errors are **serially correlated** (violation of A4) and/or **heteroskedastic** (violation of A5).
- The least squares estimator is **unbiased** even if these assumptions are violated.
- But, it turns out there are **more efficient** estimators than least squares if the errors are heteroskedastic and/or serially correlated.
- Example of serial correlation: if $\varepsilon_1 > 0$, then ε_2 is more likely to be positive also.
  - usually a problem in **time-series** applications, e.g., where we model an economic variable (GDP, stock price, etc.) over time.
  - Example: Suppose there is a big negative shock ($\varepsilon < 0$) to GDP this year (e.g., oil prices rise). If this triggers a recession, we're likely to see another negative shock next year.
- Example of heteroskedasticity: $Var(\varepsilon)$ depends on some variable Z (which may or may not be in the model)
  - Example: suppose we regress income (Y) on education (X). Although people with more education to have higher income on average, they also (as a group) have more variability in their earnings. That is, some people with PhD's get good jobs & earn a lot, but some are "over-qualified" for everything except Starbucks. So we see that highly educated people have a high variance of earnings, as well as a high mean. In contrast, almost everyone that flunks out of high school earns very little. (low education, low variance of earnings) ... **draw a picture of a more efficient estimator**

Assumption 6: No perfect collinearity

- This is really a technical assumption.
- With perfect collinearity, one (or more) independent variables is a perfect linear function of others.
- Perfect collinearity is a problem, because the least squares estimator cannot separately attribute variation in *Y* to the independent variables.
  - Example: suppose we regress weight on height measured in meters and height measured in centimeters. How could we decide which regressor to attribute the changing weight to?
- The solution is just to exclude one of the (redundant) variables from the model.

The Gauss-Markov Theorem (GMT)

- GMT: when the classical Assumptions 1-6 are satisfied, then the least squares estimator $\hat{\beta}_j$ has the smallest variance of all **linear unbiased estimators** of β_j, for $j = 0,1,2,...,k$.
- This is a pretty powerful theorem.
- Sometimes we say the least squares estimator is
  BLUE: **B**est **L**inear **U**nbiased **E**stimator
  where "best" means most efficient. This is just a convenient way of remembering the Gauss-Markov Theorem (GMT).
- The GMT is a big reason we like least squares so much.
- What do we mean by a "linear" estimator? One that is **linear in y** (take a look at the formula for the least squares estimator). Linear estimators are easy to compute. Linearity also makes statistical properties simple if the errors are normal: linear functions of normal variables are normal.
- Is linearity restrictive? No, as it turns out. The proof is complex, but you can show that any nonlinear estimator is biased for **some** error probability distribution. So the real "bite" of the GMT is unbiasedness, not linearity.

Violating the Classical Assumptions

- We know that when these six assumptions are satisfied, the least squares estimator is BLUE
- We almost always use least squares to estimate linear regression models
- So in a particular application, we'd like to know whether or not the classical assumptions are satisfied
  - if they're not, then there is usually a "better" estimator available
- For the remainder of the semester, we'll talk about
  - **what happens** when the classical assumptions are violated
  - **how to test** for violations
  - **what to do about it** if we find a violation
- We'll deal with the assumptions one-by-one, starting with Assumption 1.

Violating Assumption 1

- Assumption 1 of the CLRM is that the regression function:
  - is **linear in the coefficients**
  - is **correctly specified**
  - has an **additive error term**
- As we said already, the "additive error" part is a pretty weak assumption – nothing really to worry about here
- We rely on economic theory (and maybe common sense) to tell us that the regression is (or is not) linear in the coefficients
  - if it's not, we can estimate a **nonlinear regression** model; we'll see some examples later in the semester, time permitting
- So we'll begin by talking about **specification errors**: when our regression model is **incorrectly specified**

Specification

- Every time we write down a regression model (and estimate it!) we make some important choices:
  - what independent variables belong in the model?
  - what functional form should the regression function take (i.e., logarithms, quadratic, cubic, etc.)?
  - what kind of distribution should the errors have?
- Usually, we look to economic theory (and some common sense!) to guide us in making these decisions.
- The particular model that we decide to estimate is the culmination of these choices: we call it a **specification**
  - a regression specification consists of the model's independent variables, the functional form, and an assumed error distribution

Specification Error

- It is convenient to think of there being a **right** answer to each of the questions on the preceding slide
- That is, a **correct specification**
- Sometimes we call it the **data generating process (DGP)**
  - the DGP is the **true** (unknown) model that “generates” the data we observe
- The DGP is a population concept: we never observe it
- One way to think about regression analysis is that we want to learn about the DGP
- A regression model that differs from the DGP is an **incorrect specification**
- An incorrect specification arises if we make an **incorrect choice** of:
  - independent variables to include in the model
  - functional form
  - error distribution
- We call an incorrect specification a **specification error**

What is functional form, and why does it matter?

- As always, our regression model is:

  $$Y_i = E[Y_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}] + \varepsilon_i$$

- Having chosen the independent variables $X_{1i}, X_{2i}, \ldots, X_{ki}$ that will be included in the model, we need to decide on a **shape** for the regression function $E[Y_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}]$
  - should it pass through the origin? or should it have a non-zero intercept? should the intercept be the same for all observations? or are there distinct groups of observations (e.g., men/women, before/after a policy, etc.) that have a separate intercept?
  - do you think the relationship between X_{ji} and Y_i is a straight line? a curve? is it monotone? should the slope be the same for every observation? or are there distinct groups of observations that have separate slopes?
- A **functional form** is a mathematical specification of the regression function $E[Y_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}]$ that we **choose** in response to these questions.
- Different functional forms may give very different “answers” about the marginal effects of X on Y – and typically different predictions too (draw some pictures)
- As always, let economic theory and common sense be your guide.

Should the Model Include an Intercept?

- The short answer: **yes**, **always**.
- Why?
- Even if theory tells you that the regression function should pass through the origin (i.e., theory tells you that when all the *X* are zero, then *Y* is zero also) **it is better to estimate a zero intercept than to force the intercept to be zero.**
- Why is this better? **In case the theory is wrong.**
- If leave out the intercept, but the **true** intercept (i.e., of the DGP) turns out **not** to be zero, you can really screw up your estimates of the slopes (**draw a picture**)
  - For example, firm output depends on inputs, zero gets zero. But, are there threshold input levels to get positive output?

The Linear Functional Form

- The simplest functional form arises when the independent variables enter **linearly:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- Why might you choose this form?
  - if theory tells you that the **marginal effect** of *X* on *Y* is a **constant:** i.e., the same at every level of *X*

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 \quad \text{and} \quad \frac{\partial Y_i}{\partial X_{2i}} = \beta_2$$

    equivalently, the regression function is a straight line (has a constant slope).
  - if theory tells you that the **elasticity** of *Y* with respect to *X* is **not a constant**:

$$\eta_{Y,X_1} = \frac{\partial Y_i}{\partial X_{1i}} \frac{X_{1i}}{Y_i} = \beta_1 \frac{X_{1i}}{Y_i} \quad \text{and} \quad \eta_{Y,X_2} = \frac{\partial Y_i}{\partial X_{2i}} \frac{X_{2i}}{Y_i} = \beta_2 \frac{X_{2i}}{Y_i}$$

  - if you don’t know what else to do (but a polynomial is better)

The Polynomial Functional Form

- A flexible alternative to the linear functional form is a **polynomial**: one or more independent variables are raised to powers other than one, e.g.,

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_3 X_{2i} + \varepsilon_i \qquad \text{(model 1)}$$

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_3 (X_{1i})^3 + \beta_4 X_{2i} + \varepsilon_i \qquad \text{(model 2)}$$

- Why would you choose a polynomial functional form?
  - if theory tells you that the **marginal effect** of *X* on *Y* is a **not constant:** i.e., it changes with the level of *X* (the regression function is a curve)

    $$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 + 2\beta_2 X_{2i} \quad \text{and} \quad \frac{\partial Y_i}{\partial X_{2i}} = \beta_3 \qquad \text{(model 1)}$$

  - the **elasticities** are also not constant:

    $$\eta_{Y,X_1} = \frac{\partial Y_i}{\partial X_{1i}} \frac{X_{1i}}{Y_i} = \left(\beta_1 + 2\beta_2 X_{1i}\right)\frac{X_{1i}}{Y_i} \quad \text{and} \quad \eta_{Y,X_2} = \frac{\partial Y_i}{\partial X_{2i}} \frac{X_{2i}}{Y_i} = \beta_3 \frac{X_{2i}}{Y_i}$$

  - if you don’t know what else to do (polynomial specifications are **very flexible – draw some pictures**).
- Caveat: difficult to interpret the individual regression coefficients

The Double-Log Functional Form

- Maybe the most common specification that is linear in parameters but non-linear in the variables:

  $\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \varepsilon_i$

  where "ln" is the natural logarithm
- Why might you choose this form?
  - if theory tells you that the **marginal effect** of *X* on *Y* is a **not constant:** i.e., changes with the level of *X* (the regression function is a curve)

    $$\frac{\partial Y_i}{\partial X_{1i}} = \frac{\partial Y_i}{\partial \ln Y_i} \frac{\partial \ln Y_i}{\partial \ln X_{1i}} \frac{\partial \ln X_{1i}}{\partial X_{1i}} = \frac{1}{\partial \ln Y_i / \partial Y_i} \frac{\partial \ln Y_i}{\partial \ln X_{1i}} \frac{\partial \ln X_{1i}}{\partial X_{1i}} = \beta_1 \frac{Y_i}{X_{1i}}$$
  - if theory tells you that the **elasticity** of *Y* with respect to *X* **is constant (this is the main reason for using this form):**

    $$\eta_{Y,X_1} = \frac{\partial Y_i}{\partial X_{1i}} \frac{X_{1i}}{Y_i} = \beta_1 \frac{Y_i}{X_{1i}} \frac{X_{1i}}{Y_i} = \beta_1$$
- In the double-log model, the *β's* measure elasticities: the % change in *Y* for a 1% change in *X* (holding other independent variables constant)
- Implies a smooth but nonlinear relationship between *X* and *Y*
- Don't forget ln is only defined for positive numbers! You need to make sure that *X* and *Y* are not zero & not negative.

The Semi-log Functional Form

- There are two versions of this one:

$$\ln Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad \text{(model 1)}$$

$$Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad \text{(model 2)}$$

- That is, some but not all variables are in natural logarithms
- We use this kind of functional form very often – a very common application is when the variable being logged has a very **skewed distribution**: taking the logarithm compresses extreme values
- Neither the marginal effect nor the elasticity is constant
- The coefficients in model 1 have a very useful interpretation: β_1 measures the **percentage** change in Y_i for a **one unit** change in X_i
  - very common in instances where changes in Y occur on a percentage basis, i.e., where we want to model **growth rates**
  - *example: in model 1, if Y is a person's salary and* X_1 *is years of education, then* β_1 *measures the % increase in salary associated with acquiring one more year of education*

Specification Error: choosing the wrong functional form

- As always, you should choose a functional form based on **economic theory** and **common sense**
- You should **avoid** choosing functional form based on model fit
  - i.e., don't simply choose a functional form that gives the highest R^2, or adjusted-R^2
- Why?
- **Most importantly:** because it ignores economic theory and common sense
- Note you **cannot** compare R^2 or adjusted-R^2 across specifications with different functional forms for the dependent variables (e.g., Y vs. ln$Y$)
  - Changing the dependent variable changes TSS. This makes the comparison meaningless.
- the best-fitting functional form might be "wrong" (i.e., different from the DGP)
  - just because it fits well in this sample, doesn't mean it's "right" – it might fit badly in another sample (**draw some pictures**)
  - For example, consider earnings regressed on age and age-squared, using only young people. It might predict negative earnings for old people!

Dummy Variables

- A **dummy variable** is a variable that takes value 0 or 1
- Usually, we use dummy variables to indicate the presence or absence of a characteristic
- Dummy variable examples:
  - M_i = 1 if person *i* is a man
    M_i = 0 if person *i* is a woman
  - U_i = 1 if person *i* is a member of a union
    U_i = 0 if person *i* is not a member of a union
  - X_i = 1 if firm *i* operates in an export market
    X_i = 0 if firm *i* does not operate in an export market
- We use dummy variables **all the time** to allow different groups of observations to have different slopes and/or intercepts

Intercept Dummies

- The most common use of dummy variables is to allow different regression intercepts for different groups of observations
- Example:

  $W_i = \beta_0 + \beta_1 ED_i + \beta_2 F_i + \varepsilon_i$

  where

  W_i is person *i*'s hourly wage
  ED_i is person *i*'s education (years)
  F_i = 1 if person *i* is female
  F_i = 0 if person *i* is male

  then for females ($F_i = 1$), the regression model is:

  $W_i = \beta_0 + \beta_1 ED_i + \beta_2 + \varepsilon_i$ (intercept is $\beta_0 + \beta_2$)

  and for males ($F_i = 0$), the regression model is

  $W_i = \beta_0 + \beta_1 ED_i + \varepsilon_i$ (intercept is β_0)
- Notice that the **slope** of the two regression models is the same, but the intercept differs (**draw a picture)**
  - the model says that education has the same marginal effect on men's and women's wages, but that for a given level of education, the **average wage of men and women is different** differs by β_2 dollars
- What if the dependent variable was lnW_i?
- Other examples?

The Dummy Variable Trap

- Notice that in the previous example, we didn't include a second dummy variable for being a man, i.e.,

  $$W_i = \beta_0 + \beta_1 ED_i + \beta_2 F_i + \beta_3 M_i + \varepsilon_i$$

  where

  W_i is person i's hourly wage
  ED_i is person i's education (years)
  $F_i = 1$ if person i is female
  $F_i = 0$ if person i is not female (is male)
  $M_i = 1$ if person i is male
  $M_i = 0$ if person i is not male (is female)
- The reason: this model violates Assumption 6 of the CLRM (no perfect collinearity) because $M_i = 1 – F_i$, i.e., M_i is an **exact** linear function of F_i.
  - M_i is redundant – it contains **no information** that isn't already in F_i
  - there is no way to distinguish between the effect of M_i and F_i on W_i
- **WE ALWAYS HAVE ONE LESS DUMMY VARIABLE THAN CONDITIONS (CATEGORIES)!**
  - if you violate this (fall into the "dummy variable trap"), you violate Assumption 6 of the CLRM

More than 2 categories

- Using dummy variables to indicate the absence/presence of conditions with more than 2 categories is no problem – just create more dummies (one fewer than the number of categories)
- Example: dummies for a immigrant cohort
  - COHORT takes one of eight values (1950s, 1960s, 1970s, 1980s, 1990s, 2000s, temporary, and Canadian-born)
  - We could create a set of cohort dummies C:
    FIFTIES = 1 if COHORT =1, and 0 otherwise
    SIXTIES = 1 if COHORT =2, and 0 otherwise
    SEVENTIES= 1 if COHORT =3, and 0 otherwise
  - EIGHTIES = 1 if COHORT =4, and 0 otherwise
  - NINETIES = 1 if COHORT =5, and 0 otherwise
  - NOUGHTS = 1 if COHORT =6, and 0 otherwise
  - TEMPORARIES = 1 if COHORT =7, and 0 otherwise
  - CANADIAN-BORN is the **omitted category** defined by COHORT=8
  - Our regression is $\text{EARNINGS}_\text{i} = \beta_0 + \beta_1 FIFTIES_i + \beta_2 SIXTIES_i + \ldots$
    $+\beta_7 TEMPS_i + \text{(other stuff)} + \varepsilon_i$

Slope Dummy Variables

- We can also use dummy variables to allow the **slope** of the regression function to vary across observations
- Example: our wage equation with male/female dummies
  - suppose we think the **returns to education** (the marginal effect of another year of education on wage) is different for men than for women, but that the intercepts are the same
  - we could estimate the regression model: $W_i = \beta_0 + \beta_1 ED_i + \beta_2 ED_i F_i + \varepsilon_i$
  - for **women** ($F_i = 1$) the regression model is $W_i = \beta_0 + (\beta_1 + \beta_2) ED_i + \varepsilon_i$ while for **men** ($F_i = 0$) the model is $W_i = \beta_0 + \beta_1 ED_i + \varepsilon_i$
  - **draw a picture**
- all we do is introduce a "new" independent variable $ED_i F_i$ that is the product of the ED_i variable and the F_i variable (i.e., ED_i times F_i)
  - we call this variable the **interaction** of education and gender (the "Female" dummy)
- of course, we can use dummies to allow the slopes AND the intercept to vary ....
- do another example, and draw some pictures

Choosing the Independent Variables

- We’ll begin our discussion of specification errors by talking about the choice of which independent variables to include in the model
- There are several kinds of errors we can make:
  - we can leave out (omit) one or more “important” independent variables – ones that **should** be in the model
  - we can include “unimportant” (irrelevant) independent variables – ones that **should not** be in the model
  - we can abuse the tools we have and choose a specification that fits the data well, or confirms what we hoped to find **without relying on theory (or common sense) to specify the model**
- We’ll discuss the consequences of each of these kinds of specification error today, and one “omnibus” test we can use to detect possible specification error

Omitted Variables

- Suppose the **true DGP** is:

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

  but we incorrectly estimate the regression model:

  $$Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^*$$

  - *example: Y is earnings, X_1 is education, and X_2 is "work ethic" – we don't observe a person's work ethic in the data, so we can't include it in the regression model*
- That is, we **omit** the variable X_2 from our model
- What is the consequence of this?
- Does it mess up our estimates of β_0 and β_1?
  - it definitely messes up our **interpretation** of β_1. With X_2 in the model, β_1 measures the marginal effect of X_1 on Y **holding X_2 constant**. We can't hold X_2 constant if it's not in the model.
  - Our estimated regression coefficients may be **biased**
  - The estimated β_1 thus measures the marginal effect of X_1 on Y **without holding X_2 constant**. Since X_2 is in the error term, the error term will covary with X_1 if X_2 covaries with X_1 .

Omitted Variables May Cause Bias

$$E\left[\hat{\beta}_1\right] = E\left[\frac{\sum_i \left(X_{1i} - \bar{X}_1\right)\left(Y_i - \bar{Y}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right] = E\left[\frac{\sum_i \left(X_{1i} - \bar{X}_1\right)\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i - \beta_0 - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right]$$

$$= E\left[\frac{\sum_i \left(X_{1i} - \bar{X}_1\right)\left(\beta_1\left(X_{1i} - \bar{X}_1\right) + \beta_2\left(X_{2i} - \bar{X}_2\right) + \varepsilon_i - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right]$$

$$= E\left[\frac{\beta_1 \sum_i \left(X_{1i} - \bar{X}_1\right)^2}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2} + \frac{\sum_i \left(X_{1i} - \bar{X}\right)\left(\beta_2\left(X_{2i} - \bar{X}_2\right) + \varepsilon_i - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right]$$

$$E[\hat{\beta}_1] > \beta_1 \text{ or } E[\hat{\beta}_1] < \beta_1 \quad ?$$

$$= \beta_1 + \beta_2 E\left[\frac{\sum_i \left(X_{1i} - \bar{X}\right)\left(\left(X_{2i} - \bar{X}_2\right) + \varepsilon_i - \bar{\varepsilon}\right)}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2}\right] = \beta_1 + \frac{\beta_2}{\sum_i \left(X_{1i} - \bar{X}_1\right)^2} E\left[\sum_i \left(X_{1i} - \bar{X}\right)\left(X_{2i} - \bar{X}_2\right)\right]$$

$$= \beta_1 + \frac{\beta_2}{\sum_i \left(X_{1i} - \bar{X}\ \right)^2} nCov[X_1, X_2] = \beta_1 + \beta_2 \frac{Cov[X_1, X_2]}{Var[X_1]}$$

The estimated parameter is **biased**, with bias linear in the true parameter on the left-out variable, and the covariance of the left-out variable with the included variable.

Omitted Variables

- The formula is:

$$E\left[\hat{\beta}_1\right] = \beta_1 + \frac{\beta_2}{\sum_i \left(X_{1i} - \bar{X} \ \right)^2} Cov\left[X_1, X_2\right]$$

- Uncorrelated missing regressors don’t cause bias.
- Missing regressors with zero coefficients don’t cause bias.
- **Correlated missing regressors with nonzero coefficients cause bias.**

Omitted Variable Bias

- It's easy to see why leaving X_2 out of the model biases our estimate of β_1
- Because the true model is: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
  but we estimate: $Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^*$
  we can see that the error term in the mis-specified model is: $\varepsilon_i^* = \beta_2 X_{2i} + \varepsilon_i$
- So, if X_1 and X_2 are correlated, then Assumption 3 of the CLRM is violated in the mis-specified model: X_1 is correlated with ε_i^*
  - if X_2 changes, so do ε_i^*, X_1 and Y
  - but all we can **see** is X_1 and Y changing – and we incorrectly attribute variation in Y to X_1 that is really due to X_2
- That is, β_1 measures the effect of X_1 and (some of) the effect of X_2 on Y
- Consequently, our estimate of β_1 is biased
- *Back to our example: Imagine the true $\beta_1 > 0$ and $\beta_2 > 0$ so that more educated workers earn more, and so do workers with a stronger work ethic. Imagine also that $Cov(X_1,X_2) > 0$, so that workers with a stronger work ethic also acquire more education on average. When we leave work ethic out of the model, β_1^* measures the effect of education* ***and*** *work ethic on earnings.*
- Only if we are **very lucky** and $Cov(X_1,X_2) = 0$, does leaving X_2 out of the model not bias our estimate of β_1

Is the bias positive or negative?

- We know that if $Cov(X_1,X_2) \neq 0$, and we omit X_2 from the model, our estimate of β_1 is biased:
$$E[\hat{\beta}_1] \neq \beta_1$$
- But is the bias **positive** or **negative**? That is, can we predict whether:
$$E[\hat{\beta}_1] > \beta_1 \text{ or } E[\hat{\beta}_1] < \beta_1 \quad ?$$
- In fact, you can show that: $E[\hat{\beta}_1] = \beta_1 + \beta_2\alpha_1$
where α_1 is the slope coefficient from the *auxiliary regression* of X_2 on X_1 :
$$X_{2i} = \alpha_0 + \alpha_1 X_{1i} + u_i$$
where u_i is a classical error term
- Note that α_1 has the same sign as $Cov(X_1,X_2)$
- *Back to our example: we assumed $\beta_2 > 0$ (people with a stronger work ethic earn more), and $Cov(X_1,X_2) > 0$ (people with a stronger work ethic acquire more education). Because $Cov(X_1,X_2) > 0$, we know that $\alpha_1 > 0$ also. Therefore:*
$$E[\hat{\beta}_1] = \beta_1 + \beta_2\alpha_1 > \beta_1$$

  *so we **overestimate** the effect of education on earnings – we measure the effect of having more education **and** having a stronger work ethic.*

Detecting and Correcting Omitted Variable Bias

- How do you know you’ve omitted an important variable?
  - your best guide here is common sense and economic theory
  - **before** you specify and estimate the regression model, think hard about what **should** be in the model – what common sense and economic theory tell you are important predictors of Y
  - **after** you specify the regression but **before** you estimate it, predict the sign of the regression coefficients. Then compare the actual sign of the estimated regression coefficients with the predicted signs. If any have the “wrong” sign, you may have omitted something correlated with that independent variable.
- How do you correct for omitted variable bias?
  - that’s “easy” – just add the omitted variable to your model!
  - of course you probably already would have done so if you could ... (like “work ethic” – it’s hard to measure)
  - maybe you can include a “proxy” for the omitted variable instead – something highly correlated with the omitted variable (e.g., use number of sick days at work as a proxy for work ethic, or an IQ score as a proxy for intelligence)
  - Maybe you can correct for the bias using **instruments** (more later)

Including Irrelevant Variables

- What happens if we include an independent variable in our regression that doesn't belong there?
- Suppose the true DGP is:

  $$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

  but we the model we estimate is:

  $$Y_i = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \varepsilon_i^*$$
- We see that the error in the mis-specified model is $\varepsilon_i^* = \varepsilon_i - \beta_2^* X_{2i}$
- If X_2 is **really** irrelevant, then $\beta_2^* = 0$ and everything is ok:
  - our estimates of β_0 and β_1 will be unbiased
  - so will our estimate of β_2 (we expect it to be zero)
- The only cost of including an irrelevant independent variable is that we lose a degree of freedom:
  - we should expect adjusted Rsquared (Rbar-squared) to decrease
  - we get less precise estimates of all the other regression coefficients (standard errors get bigger, *t*-stats are closer to zero so we're less likely to reject any hypothesis)

Data Mining

- At the end of the day, it is up to the econometrician to decide what independent variables to include in the model
- There is a temptation to choose the model that fits “best,” or that tells you what you want to hear
- Resist this temptation! Let economic theory and common sense be your guide!
- An example of what **not** to do: data mining
  - we could estimate lots and lots and lots of “candidate” regression specifications, and choose the one whose results we like most
  - the problem: you’ll discard the specifications where [1] coefficients have the “wrong” sign (not wrong from a theoretical standpoint, but “wrong” in the sense that you don’t like the result), and/or [2] *t*-stats are small (you discard specifications where variables you care about are statistically insignificant)
  - so you end up with a regression model where the coefficients you care about have big *t* stats and the “right” sign.
  - How confident can you really be of your result if you had to throw away lots of “candidate” regressions? Do we **really** learn anything about the DGP?
    - do your coefficients **really** have the right sign?
    - are the *t* stats **really** big?

Other examples of bad practice

- There are lots of other ways to “cheat” when choosing a specification
- **Stepwise regression**: start with a list of “candidate” independent variables. Add them one at a time to the regression, and only keep them in the model if R^2 increases by a pre-specified amount.
  - **problems:** ignores economic theory (and common sense); if the independent variables are highly correlated, it can produce garbage
- **Sequential Specification Search**: start with a “big” model (one that includes lots of independent variables). Sequentially add or drop independent variables until you end up with a “good” regression (one that you like)
  - **problems:** like data mining, given that you had to throw away lots of “bad” regressions, how confident can you be of your results? If you drop variables because they have low *t* stats, you introduce omitted variable bias into other coefficients if they are correlated with the dropped variable.

The RESET test

- A common omnibus test for specification error is Ramsey's Regression Specification Error Test (RESET). It works as follows.

1. Estimate the regression you want to test. Suppose it has *k* independent variables. Call this model 1.
2. Compute the predicted values (*Y*-hat) from model 1.
3. Regress *Y* on the *k* independent variables **and** on the square of *Y*-hat, the cube of *Y*-hat, etc. (an *M*th-order polynomial in *Y*-hat. You choose *M*) Call this model 2.
4. Compare the results of the two regressions using an *F*-test. The test statistic is:

$$F = \frac{RSS_1 - RSS_2 / M}{RSS_2 /(n-k-M-1)} \sim F_{M,n-k-M-1}$$

5. If the test statistic is "big" (bigger than a critical value) then reject the null hypothesis of correct specification

- *Intuition: if the model is correctly specified, then all those functions of Y-hat shouldn't help to explain Y (their estimated coefficients should be statistically insignificant). The test statistic is "big" if RSS (the part of variation in Y that we don't explain) is much bigger in model 1 than model 2 – meaning all those functions of Y-hat helped a lot to explain variation in Y.*
- Problem: if you reject the null of correct specification, the test doesn't tell you what the specification error is, or how to fix it.
- This test detects *heteroskedasticity*, which may be caused by misspecification.