Multiple Regression Properties

Tyler Ransom

Univ of Oklahoma

Jan 31, 2019

Today's plan

1. Review reading topics

    1.1 Expected Value of OLS Estimators

    - Omitted variable bias

    1.2 Variance of OLS Estimators

    - Multicollinearity

2. In-class activity: Practice with regression properties

Expected Value of OLS Estimators

Assumptions

- As with simple regression, have assumptions under which OLS is unbiased

- Similar to simple regression, but some slight differences:

1. Linear in Parameters

2. Random Sampling

3. No Perfect Collinearity

4. $E(u|x_1, \ldots, x_k) = 0$

1. Linear in Parameters; 2. Random Sampling

- We assume that the model in the population can be written as

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

  where the β_j's are the population parameters and u is the unobserved error

- Recall that we can take apply nonlinear functions to any of the variables

- Linear-in-parameters assumption is not so limiting

- Random sampling is an assumption made about the data collection process

3. No Perfect Collinearity

- What is collinearity?

- In the sample (and population), **none of the x's can be constant**

- **None of the x's can be exact *linear* relationships of any other x's**

- This is the multi-dimensional analog of "$var(x) > 0$"

- e.g. if x_1 is an exact linear function of x_2 and x_3 in the sample:

    - we say the model suffers from **perfect collinearity**

What causes perfect collinearity?

- #1 cause: user error (specify model that has perfectly collinear variables)
- Other cause: bad luck in drawing the sample
- Also: need $N \geq K + 1$ (N is sample size, K is # of slope β's)
- **How to fix perfect collinearity problem?**
- Exclude one of the offending variables
  - e.g. if x_1 and x_2 are perfectly collinear, drop one of them
- R will usually do this for you

Correlation among the x's

- "No Perfect Collinearity" does *not* mean the x's have to be uncorrelated
  - in the population or the sample
- Nor does it say they cannot be "highly" correlated
- It simply rules out $corr(x_1, x_2) = \pm 1$
- We'll talk in a moment about the ramifications of $corr(x_1, x_2) \approx 1$
- OLS gives us *ceteris paribus* effects precisely when x's are correlated

4. $E(u|x_1, x_2, ..., x_k) = 0$ for all $x_1, ..., x_k$

- Remember, the real assumption is $E(u|x_1, x_2, ..., x_k) = E(u)$
- If u is correlated with any of the x's, the assumption is violated
- This is usually a good way to think about the problem
- When assumption holds, we say x's are **exogenous explanatory variables**
- If x_j is correlated with u, say x_j is an **endogenous explanatory variable**

Example: class size and test scores

- Suppose, for a standardized test score,

$$score = \beta_0 + \beta_1 classsize + \beta_2 income + u$$

- Even for same income, families differ in interest/concern about education
- Family support and student motivation are in u
- Are these correlated with class size even though we have included income? Probably.
- For observational data, always a risk that x's are correlated with u
  - "Correlation is not causation"

Unbiasedness of OLS

- Under previous 4 assumptions the OLS estimators are unbiased:

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, 2, \ldots, k$$

  for any values of the β_j.

- We won't prove this, but see Appendix 3A in book if interested

- Often the hope is that if our focus is on, say, x_1, we can:
  - **include enough other variables** in $x_2, \ldots, x_k$ ...
  - to make the zero conditional mean assumption **close to true**

Inclusion of irrelevant variables

- What happens if we include x's that don't affect y?
- OLS is still unbiased ($E(\hat{\beta}_j) = \beta_j$ is still true if $\beta_j = 0$)
- But including it does come at a cost:
  - The variance of the β_j's might go up

Omitted variable bias

- What happens if we *exclude* x's that *do* affect y?

- Also known as **underspecifying the model**

- OLS is biased, because Assumption 4 fails

- Hence the term **omitted variable bias**

- Why would we exclude an important x?

    - Because we can't collect data on it (e.g. cognitive ability, personality, etc.)

How bad is it?

- It's fairly easy to quantify the bias

- Suppose our true model satisfied Assumptions 1-4:

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

  but we can't see x_2 so instead we estimate

$$y = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_1 + \widetilde{u},$$

  where $\widetilde{u} = \beta_2 x_2 + u$

- $\widetilde{\beta}_1$ will be biased, but by how much?

Quantifying omitted variable bias

- We already have a relationship between $\widetilde{\beta}_1$ and the unbiased estimator, $\hat{\beta}_1$:

$$\widetilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \widetilde{\delta}_1$$

  where $\widetilde{\delta}_1$ is the slope coefficient of a regression of x_2 on x_1

- $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased (or would be if we could compute them):

$$\begin{aligned} E(\widetilde{\beta}_1) =& E(\hat{\beta}_1) + E(\hat{\beta}_2)\widetilde{\delta}_1 \\ =& \beta_1 + \beta_2 \tilde{\delta}_1 \\ Bias(\widetilde{\beta}_1) =& \beta_2 \widetilde{\delta}_1 \end{aligned}$$

When there's no bias

- $\tilde{\delta}_1$ has the same sign as the sample correlation $Corr(x_1, x_2)$
- No bias in two (esoteric) cases:
  1. $\beta_2 = 0$, i.e. x_2 uncorrelated with y
  2. $Corr(x_1, x_2) = 0$, i.e. $\widetilde{\delta}_1 = 0$
- In general: biased
- Mathematical support for "confounder: u is correlated with both y and x"

Direction of bias

- Important to tell if $\widetilde{\beta}_1$ is "too high" or "too low" compared to $\hat{\beta}_1$

- Direction depends on both $\hat{\beta}_2$ and $Corr(x_1, x_2)$ (i.e. $\widetilde{\delta}_1$)

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive Bias	Negative Bias
$\beta_2 < 0$	Negative Bias	Positive Bias

- Examples?

Bias with more than two variables

- Much more difficult to determine direction of bias with 3+ x's

- Remember: correlation of any x_j with $\widetilde{u}$ generally causes bias in **all** of the OLS estimators, not just in $\widetilde{\beta}_j$

Variance of OLS Estimators

An additional assumption

- We want to say something about $Var(\hat{\beta}_j)$ [useful for hypothesis testing]

- As in the simple regression case, add assumption of homoskedasticity:

$$Var(u|x_1, x_2, \ldots, x_k) = Var(u) = \sigma^2$$

- Homoskedasticity is usually violated

- Go along with it for now (It makes things simpler)

- Later in the course we'll talk about how to test its validity

The Gauss-Markov Assumptions

- All five assumptions combined are called the **Gauss-Markov Assumptions**:

1. Linear in Parameters
2. Random Sampling
3. No Perfect Collinearity
4. $E(u|x_1, \ldots, x_k) = 0$
5. $Var(u|x_1, x_2, \ldots, x_k) = \sigma^2$ for all $x_1, x_2, \ldots, x_k$

The variance of the β_j's

- The formula for the variance is

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, j = 1, 2, ..., k$$

where

$$\begin{aligned} \sigma^2 &= Var(u) \\ SST_j &= (N-1)Var(x_j) = \sum_{i=1}^{N}(x_{ij} - \bar{x}_j) \\ R_j^2 &= R^2 \text{ from a regression of } x_j \text{ on all other } x\text{'s} \end{aligned}$$

What factors affect $Var(\beta_j)$?

1. **Error variance:** As $\downarrow \sigma^2$, $\downarrow Var(\beta_j)$

2. **Total variation in x_j:** As $\uparrow SST_j$, $\downarrow Var(\beta_j)$

3. **Correlation w/other x's:** $R_j^2 \to 1$, $Var(\hat{\beta}_j) \to \infty$

- Last one is called **multicollinearity**

- If x_j is unrelated to all other x's, it is easier to estimate its ceteris paribus effect on y

- $R_j^2 = 0$ is very rare—even small values are not common

Multicollinearity: How high is too high?

- R_j^2 "close" to 1 is called the "problem" of **multicollinearity**
- We can't generally define what we mean by "close"; just can't have $R_j^2 = 1$
- Multicollinearity is **not** a violation of Gauss-Markov!
- Thus, impossible to state hard rules about when it is a "problem"
- Some people like to use Variance Inflation Factor (VIF)
- But VIF can be offset by large sample size (i.e. large SST_j)

Estimating σ^2

- Just like with simple regression, our estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N-K-1}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{1}{df}SSR$$

  where df is the degrees of freedom

- Here, $df = N - K - 1$

- Why? $K + 1$ parameters to estimate (the β's)

- So subtract those from the total sample size

Standard error formulas

- Regression packages automatically report $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

- Regression packages typically report the **standard error** of β_j (SE_{β_j})

$$SE_{\hat{\beta}_j} = \sqrt{\widehat{Var}(\beta_j)} = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - R_j^2)}}$$

- If you know linear algebra:

$$SE_{\hat{\beta}_j} = \sqrt{\hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{[j,j]}}$$