Topic 6: Multiple Regression Estimation

- We covered

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where we only have one regressor (explanatory variable, independent variable) X_i

- But what if true specification is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

e.g. Y is wage; X_1 is year of education; X_2 is age;

X_3 is female binary variable;…etc.

- If all those variables determine Y, but we leave them out, they are in effect captured by u.

1) Omitted Variable Bias

- The true model is

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

- But we instead run

$$Y_i = \beta_0 + \beta_1 X_{1i} + \tilde{u}_i, \text{ where } \tilde{u}_i = \beta_2 X_{2i} + u_i$$

- If X_2 and X_1 are correlated (corr(X_2,X_1) is not equal to 0) and X_2 is a truly determinant of Y (beta2 is not equal to 0), then our OLS Assumption 1 breaks down: $E(\tilde{u}_i|X_{1i}) \neq 0$
- Then $\hat{\beta}_1$ will not be unbiased, nor will it be consistent.

1) Omitted Variable Bias

- OVB (omitted variable bias)

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\frac{cov(X_{1i}, \tilde{u}_i)}{var(X_{1i})}}_{bias}$$

$$= \beta_1 + \underbrace{\frac{cov(X_{1i}, \beta_2 X_{2i} + u_i)}{var(X_{1i})}}_{bias} = \beta_1 + \underbrace{\beta_2 \frac{cov(X_{1i}, X_{2i})}{var(X_{1i})}}_{bias}$$

- In this context, we can determine the sign of bias.

  The sign is determined by β_2 and $cov(X_{1i}, X_{2i})$.

$\underbrace{\beta_2}_{+}$ and $\underbrace{cov(X_{1i}, X_{2i})}_{+} \Rightarrow$ positive bias $\quad$ $\underbrace{\beta_2}_{-}$ and $\underbrace{cov(X_{1i}, X_{2i})}_{-} \Rightarrow$ positive bias

$\underbrace{\beta_2}_{+}$ and $\underbrace{cov(X_{1i}, X_{2i})}_{-} \Rightarrow$ negative bias $\quad$ $\underbrace{\beta_2}_{-}$ and $\underbrace{cov(X_{1i}, X_{2i})}_{+} \Rightarrow$ negative bias

1) Omitted Variable Bias

- OVB Example
- Suppose that the correct model is

$$Wage_i = \beta_0 + \beta_1 Education_i + \beta_2 WorkingExperience_i + u_i$$

- But we use the model

$$Wage_i = \beta_0 + \beta_1 Education_i + \tilde{u}_i$$

- Then we will have an OVB.
- What is direction of OVB?

$\beta_2(+)$ and $cov(Education_i, WorkingExperience_i)(-)$

Negative bias: the estimated beta1_hat is smaller than the true beta1. We underestimate the effect of education on wage.

1) Omitted Variable Bias

- How do we solve the OVB problem?
- If we have data of the omitted variable, we just include them in the regression, i.e., using multiple regression
- If we do not have data of the omitted variable, this is a much more complicated problem. One potential solution is to use instrumental variable (covered in ECON 4284)

2) Multiple Regression Model

- Clearly, if we leave out relevant variables, we have got problems. In most cases, a multivariate model specification is appropriate.
- $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$
- $X_1, X_2, X_3 \ldots, X_k$ are the k different regressors
- beta0 is constant
- beta1: the effect of X_1 on Y holding all other variables constant.
- beta2: the effect of X_2 on Y holding all other variables constant.

  ….
- u: the error terms, all other factors that affect Y.
- Everything from univariate regression carries over. Now we have to estimate not just beta0 and beta1, but also beta2, beta3,…betak.

3) OLS in multiple Regression Model

- We still want estimators of the beta's so sum squared errors are minimized:

$$u_i = Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} ... - \beta_k X_{ki}$$

$$\min \sum u_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} ... - \beta_k X_{ki})^2$$

- FOC: $$\frac{\partial \sum (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} ... - \beta_k X_{ki})^2}{\partial \beta_0} = 0$$

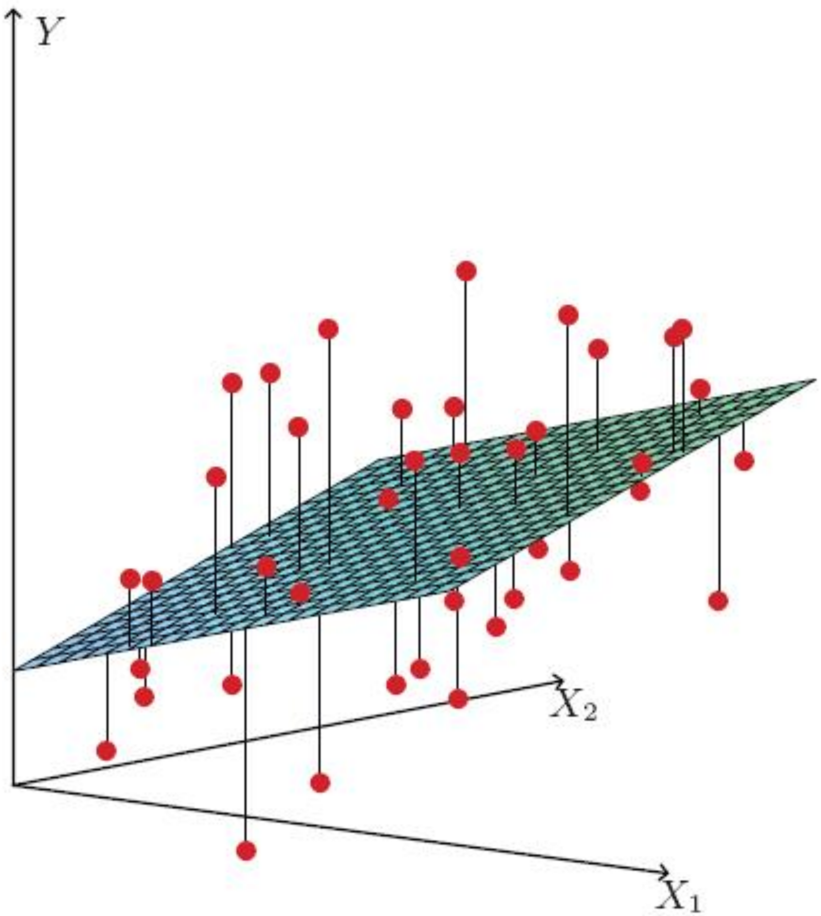
$$\frac{\partial \sum (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} ... - \beta_k X_{ki})^2}{\partial \beta_1} = 0$$

$$...$$

$$\frac{\partial \sum (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} ... - \beta_k X_{ki})^2}{\partial \beta_k} = 0$$

3) OLS in multiple Regression Model

- We want find estimators of the beta's (or the linear function of X's) so the sum of squared residuals is minimized.



3) OLS in multiple Regression Model

- There are k+1 unknowns (beat0, beta1, beta2…,betak) and k+1 equations.
- A lot of algebra. But the idea is the same as univariate case.
- OLS "line" will be:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$$

$\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_k$ are estimators from OLS procedure.

- Again, residual is $\hat{u}_i = Y_i - \hat{Y}_i$
- Again, $\sum \hat{u}_i = 0$

3) OLS in multiple Regression Model

- Regression of *TestScore* against *STR*:

$$\hat{TestScore} = 698.9 - 2.28 \times STR$$

- Now include percent English Learners in the district (*PctEL*):

$$\hat{TestScore} = 686 - 1.10 \times STR - 0.65 \times PctEL$$

- What happens to the coefficient on *STR*?
- Why? (*Note*: corr(*STR*, *PctEL*) = 0.19)

3) OLS in multiple Regression Model


```
. reg testscr str el_pct, r

Linear regression                               Number of obs =     420
                                                F(  2,    417) =  223.82
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4264
                                                Root MSE      =  14.464
```

testscr	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-1.101296	.4328472	-2.54	0.011	-1.95213	-.2504616
el_pct	-.6497768	.0310318	-20.94	0.000	-.710775	-.5887786
_cons	686.0322	8.728224	78.60	0.000	668.8754	703.189

4) Fit

- As with the univariate case, everything extends to multiple regression:

$$SER = \sqrt{\frac{1}{n-k-1}\sum \hat{u}_i^2} = \sqrt{\frac{SSR}{n-k-1}}$$

where n-k-1 is the degree of freedom (the number of parameters to estimate is k+1)

- R^2 is again the proportion of variation in Y explained by the regressors in the liner model

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

4) Fit

- R^2 can never decrease by adding additional variables, this implies you can keep adding variables without penalty, even if the variables are ridiculous.
- The adjusted R^2, $\bar{R}^2$ addresses this:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

- Let's relate $\bar{R}^2$ with R^2

$$R^2 = 1 - \frac{SSR}{TSS} \Rightarrow \frac{SSR}{TSS} = 1 - R^2$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

4) Fit

- The $\bar{R}^2$ (the "adjusted R^2") corrects this problem by "penalizing" you for including another regressor – the does not necessarily increase when you add another regressor.
- Note $0 \leq R^2 \leq 1$

  If $R^2 = 0$, then $\bar{R}^2 = 1 - \dfrac{n-1}{n-k-1} < 0$

  If $R^2 = 1$, then $\bar{R}^2 = 1$
- Note that $\bar{R}^2 \leq R^2$
- however, if n is large, the two will be very close.

As $n \to \infty$, $\dfrac{n-1}{n-k-1} \to 1$, so $\bar{R}^2 \to 1-(1-R^2) = R^2$

- Don't focus on $\bar{R}^2$ or R^2 to decide your model. Let theory dictate your variables. Use $\bar{R}^2$ or R^2 as an indication of whether you may have left stuff out.

5) Multiple Regression Assumption

- **Assumption #1: the conditional mean of u given the included X's is zero.**
- $E(u_i|X_{1i},\ldots, X_{ki}) = 0$
- This has the same interpretation as in regression with a single regressor.
- If an omitted variable (1) belongs in the equation (so is in *u*) and (2) is correlated with an included *X*, then this condition fails
- Failure of this condition leads to omitted variable bias
- The solution – if possible – is to include the omitted variable in the regression.

5) Multiple Regression Assumption

- **Assumption #2: $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.**
- This is satisfied automatically if the data are collected by simple random sampling.

- **Assumption #3: large outliers are rare (finite fourth moments)**
- This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).

5) Multiple Regression Assumption

- **Assumption #4: There is no perfect multicollinearity**
- Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors:
- Example: Suppose you accidentally include STR twice:
- Type: reg testscr str str in Stata:

```
. reg testscr str str
note: str omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   22.58
       Model |  7794.11004     1  7794.11004           Prob > F      =  0.0000
    Residual |  144315.484   418  345.252353           R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0490
       Total |  152109.594   419  363.030056           Root MSE      =  18.581

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
         str |          0  (omitted)
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

5) Multiple Regression Assumption

- **Assumption #4: There is no perfect multicollinearity**
- Example: $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$$

Suppose X_3 : age in years

X_2 : age in days

$$X_2 = 365 X_3; \quad corr(X_3, X_2) = 1$$

- This does not make sense. How do you interpret beta2?
- Literally, if we hold X_3 and all other variables constant, Y will change by beta2 if we change X_2 by 1 unit. How can you change X_2 (age in days) but also keep X_3 (age in year) constant?
- Stata will drop one automatically.

6) Distribution of OLS estimators

- The standard error of beta1_hat, beta2_hat,…,betak_hat become difficult to do without matrix algebra.
- Just like univariate, each beta_hat has a normal distribution in large samples (as n goes to infinite) by applying CLT.
- The usual hypothesis testing can be done, such as t-stat, p-values, and confidence interval. We will do this in Chapter 7.

7) Multicollinearity

- Perfect v.s. Imperfect multicollinearity
- Perfect multicollinearity: two major reasons that happens and OLS cannot run.

(i) The age example: age in years and age in days has the same exact information.

Interest rate in decimal (e.g. 0.08) or basis points (e.g. 8)

…

Solution: remove one of them (otherwise Stata will automatically drop one.)

7) Multicollinearity

- Perfect multicollinearity:

  (ii) dummy variable trap

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
&= \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3
\end{aligned}
\qquad where\ \ X_0 = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix}
$$

Suppose that X_2 is female dummy variable and X_3 is male dummy variable.

e.g.

$$
\begin{array}{ccccc}
X_1 & & X_2 & & X_0 \\
\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} & + & \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} & = & \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}
\end{array}
$$

So, don't include X2 (or X3).

7) Multicollinearity

- Imperfect multicollinearity:
- X_2 and X_3 might be highly correlated, but not perfectly.
- OLS can still run, but the standard error on beta2, beta3 or both may be very high, which will result in a wide confidence interval, low t-stat, high p-value (more likely NOT to reject the null beta=0.)
- Classic Example:

$$Consumption_i = \beta_0 + \beta_1 Income_i + \beta_2 Wealth + ... + u_i$$

- Income and wealth will usually have a correlation>0.90
- What should you do? You can collect more data, you can combine the variables, or just do nothing and point the potential problem.