ECOM20001
Econometrics 1

Lecture Note 6
Multiple Linear Regression - Estimation

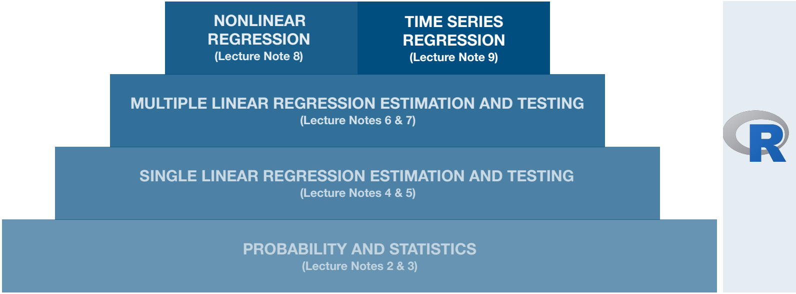
A/Prof David Byrne
Department of Economics
University of Melbourne

Stock and Watson: Chapter 6

Summary of Key Concepts

- Omitted Variable Bias
- Population Multiple Linear Regression Model
- Control Variables
- Heteroskedasticity
- OLS Estimator with Multiple Linear Regression
- Measures of Model Fit
- Least Squares Assumption in Multiple Linear Regression
- Perfect Multicollinearity
- Imperfect Multicollinearity

Building our Econometric Toolkit



NONLINEAR REGRESSION (Lecture Note 8)

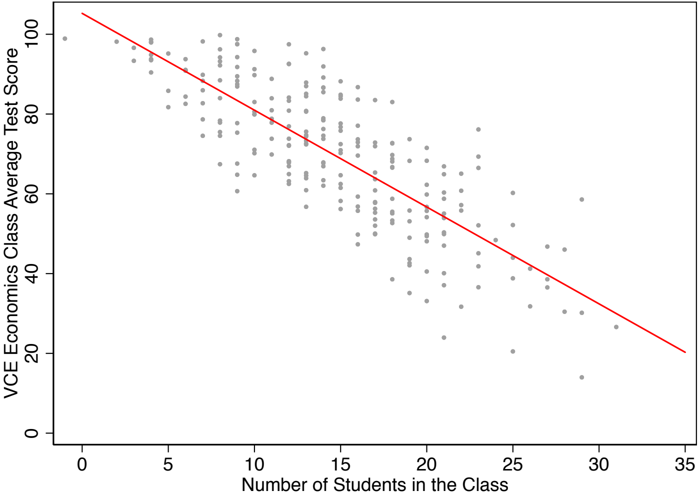
TIME SERIES REGRESSION (Lecture Note 9)

MULTIPLE LINEAR REGRESSION ESTIMATION AND TESTING (Lecture Notes 6 & 7)

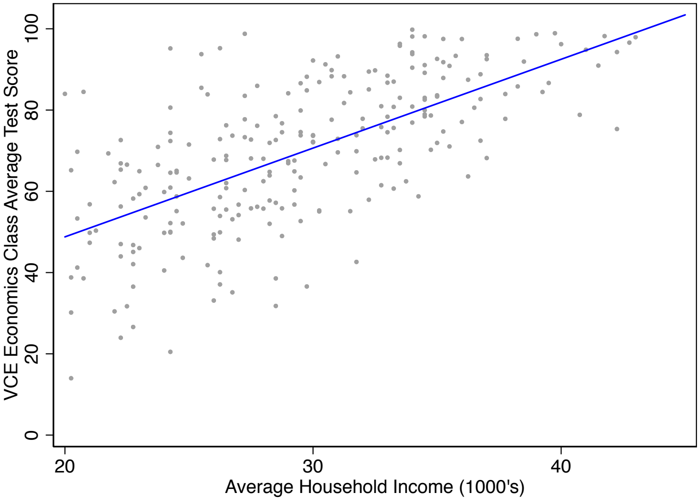
SINGLE LINEAR REGRESSION ESTIMATION AND TESTING (Lecture Notes 4 & 5)

PROBABILITY AND STATISTICS (Lecture Notes 2 & 3)

Student Test Scores and Class Size

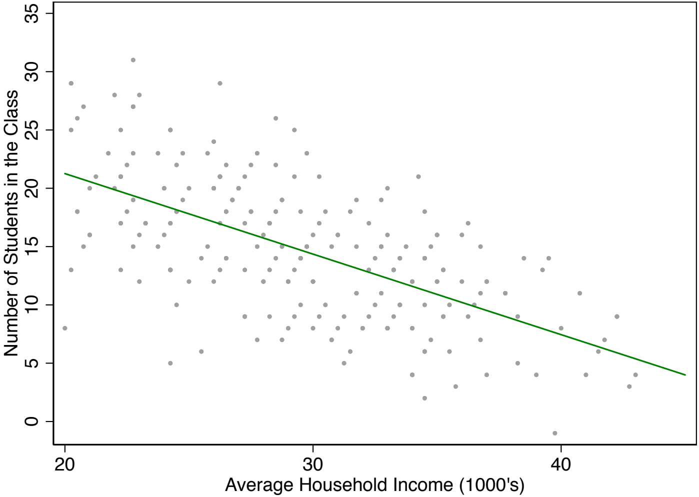


Student Test Scores and Household Income



]

Class Size and Household Income



Econometrically Modelling Test Scores

- We've seen that:
  - $(\uparrow ClassSize_i, \downarrow TestScore_i)$
- But we also now have that
  - $\downarrow Income_i, \uparrow ClassSize_i$ AND $\downarrow Income_i, \downarrow TestScore_i$

  So putting it together, we have:
  - $\downarrow Income_i \rightarrow (\uparrow ClassSize_i, \downarrow TestScore_i)$
- In words, if some other variable like $Income_i$ varies across classes, then this automatically creates a negative relationship between $ClassSize_i$ and $TestScore_i$

Econometrically Modelling Test Scores

- This has serious implications for interpreting the OLS coefficient estimate from the following single linear regression:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + u_i$$

- Remember, β_1 is meant to capture the individual relationship (or direct link) between *TestScore*$_i$ and *ClassSize*$_i$ <u>alone</u>

Econometrically Modelling Test Scores

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + u_i$$

- The OLS coefficient will fail to isolate the direct link between $ClassSize_i$ and $TestScore_i$
- Why?
  - Because $\downarrow Income_i \rightarrow (\uparrow ClassSize_i, \downarrow TestScore_i)$
- So the OLS estimate $\hat{\beta}_1$ from the single linear regression will be driven by two forces:
  1. $(\uparrow ClassSize_i, \downarrow TestScore_i)$, the negative direct relationship we want to determine empirically
  2. $\downarrow Income_i \rightarrow (\uparrow ClassSize_i, \downarrow TestScore_i)$, a separate negative indirect correlation between $ClassSize_i$ and $TestScore_i$ due to differences in $Income_i$ across classes

Econometrically Modelling Test Scores

- In words, what does all of this mean for an econometrician?
- Suppose you obtained a statistically significant (from 0) $\hat{\beta}_1$
- Suppose you further concluded from this that it means we can increase test scores by reducing class sizes.
  - This interpretation is based on the negative direct relationship between test scores and class size we have in mind

Econometrically Modelling Test Scores

- Once you make this claim, however, someone else who is watching your presentation says the following:
  "you may have a statistically significant (from 0) $\hat{\beta}_1$ estimate empirically, but couldn't that estimate just be driven by the fact that higher income schools tend to have smaller class sizes, and higher income kids tend to do better on tests?"
- This criticism of the above interpretation is based on the negative indirect relationship between test scores and class size that arises because of differences in income across schools
- It could be that there is no underlying direct relationship between test scores and class sizes driving your $\hat{\beta}_1$ estimate; it could be completely driven by an indirect relationship driven by differences income across schools

Econometrically Modelling Test Scores

- Should we expect the $\hat{\beta}_1$ estimate to be bigger or smaller than the population value of β_1?
- Conceptually, we interpretting what our OLS estimate $\hat{\beta}_1$ means, we can think of it containing two parts:

$$\hat{\beta}_1 = \underbrace{\beta_1}_{\textit{direct}} + \underbrace{\gamma}_{\textit{indirect}}$$

  where:

  - β_1: ($\uparrow$ *ClassSize$_i$*, $\downarrow$ *TestScore$_i$*), the true negative <u>direct</u> class size – test score relationship we want to determine empirically
  - γ: $\downarrow$ *Income$_i$* $\rightarrow$ ($\uparrow$ *ClassSize$_i$*, $\downarrow$ *TestScore$_i$*), the negative <u>indirect</u> class size – test score relationship being driven by differences in income across classes

- Given we expect $\gamma < 0$, this means that we can expect our single linear regression estimate to yield $\hat{\beta}_1 < \beta_1$, which means that it gives a biased estimate of the direct class size – test score relationship

Econometrically Modelling Test Scores

- So the magnitude of the OLS estimate of the relationship between $TestScore_i$ and $ClassSize_i$ from a single linear regression will be too large relative to the true value of β_1
- Intuition: $\hat{\beta}_1$ captures the true class size – test score relationship, but is ALSO confounded by the fact that richer kids are in smaller classes and tend to do better in school for reasons unrelated to class size

Omitted Variable Bias

- The example we have just discussed is an example of omitted variable bias in econometrics
- If the regressor ($ClassSize_i$) is correlated with a variable that has been omitted from the analysis ($Income_i$) AND that determines, in part, the dependent variable ($TestScore_i$), then the OLS estimator of the effect of interest (test size – class size relationship) will suffer from omitted variable bias (causing $\hat{\beta}_1 < \beta$ in our example)

Omitted Variable Bias

- Omitted variable bias occurs when the omitted variable (e.g,. income) satisfies two conditions:
  1. correlated with the included regressor (e.g, class size)
  2. helps determine the dependent variable (e.g., text scores)
- If omitted variable bias exists, then $E[\hat{\beta}_1] \neq \beta_1$, the OLS estimate of β_1 from a single linear regression is now biased, and all of our machinery for estimating and testing regression models fails

Omitted Variable Bias and OLS Assumption #1

- Omitted variable bias means that the first least squares assumption of $E[u_i|X_i] = 0$, fails
- Why? Recall u_i contains all factors other than X_i that are determinants of Y_i
- If one of these factors are correlated with X_i then this means u_i is correlated with X_i
- Example: if $Income_i$ is a determinant of $TestScore_i$ (e.g., Y_i) and we omit it then it is in u_i, AND if $Income_i$ is correlated with $ClassSize_i$ (e.g., X_i), then u_i will be correlated with X_i
- Because u_i and X_i are correlated in the presence of an omitted variable, the conditional mean of u_i given X_i is not zero → violates the first OLS assumption!
  - Recall that if $corr(u_i, X_i) \neq 0 \implies E[u_i|X_i] \neq 0$

Formula for Omitted Variable Bias

- Suppose least squares assumptions 2 (IID) and 3 (no outliers) hold, but assumption 1 (independence) does not hold
- Let $corr(u_i, X_i) = \rho_{Xu}$ is the correlation between X_i and u_i in the single linear regression
- If omitted variable bias is present, then as $n \rightarrow \infty$

$$\hat{\beta}_1 \rightarrow \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

- If there's no omitted variable bias, $\rho_{Xu} = 0$
- However, if there is omitted variable bias $\rho_{Xu} \neq 0$

Implications of Omitted Variable Bias

$$\hat{\beta}_1 \to \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}$$

- With omitted variable bias, as n gets large, $\hat{\beta}_1$ does not get close to β_1 with high probability
- The bias term $\rho_{Xu}\frac{\sigma_u}{\sigma_X}$ exists even if n is very large
- The size of the bias depends on the magnitude of ρ_{Xu}
- The direction of the bias in $\hat{\beta}_1$ depends on the sign of ρ_{Xu} (whether it's positive or negative)

Signing Omitted Variable Bias

- In our example, we had a positive relationship between our omitted variable *Income$_i$* and our outcome variable Y which was *TestScore$_i$*.
  - This means *Income$_i$* enters u_i in the single linear regression with a positive sign **(+)**.
- Further, there was a negative **(-)** relationship between our omitted variable *Income$_i$* and our independent variable X which was *ClassSize$_i$*
- Therefore, the sign of the correlation between X and u is given by sign[ρ_{Xu}]=sign[**(+)** $\times$ **(-)**]=**(-)**
- Given that

$$\hat{\beta}_1 \rightarrow \beta_1 + \underbrace{\rho_{Xu}}_{\text{(-)}} \frac{\sigma_u}{\sigma_X}$$

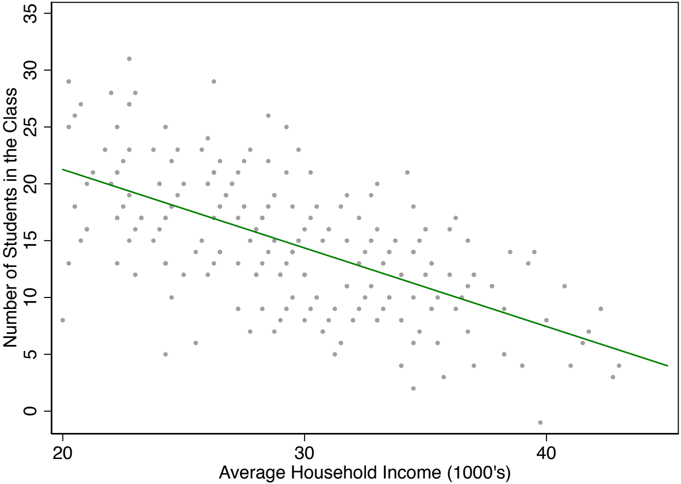
  there will be a negative bias in $\hat{\beta}_1$ relative to the true value β_1, that is: $\hat{\beta}_1 < \beta_1$

Fixing Omitted Variable Bias

- Conceptually, how might we try to fix the omitted variable bias problem in our example?
- Let's start with the source of the problem: income levels vary across schools, which is what creates the bias in $\hat{\beta}_1$
- What if instead of using all the schools in our sample, we focused on a group of schools that had similar income levels?
  - for example, only look at schools with average household income between $29,000 and $31,000

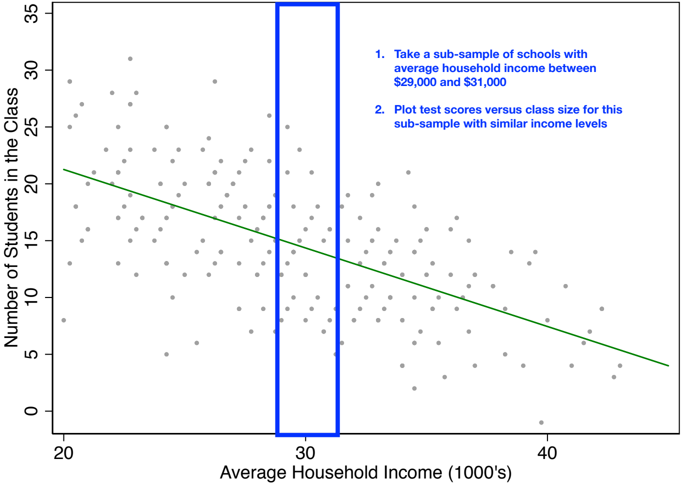
Fixing Omitted Variable Bias

Source of the bias: variation in income



Fixing Omitted Variable Bias

Fixing the problem: taking a sub-sample with similar income

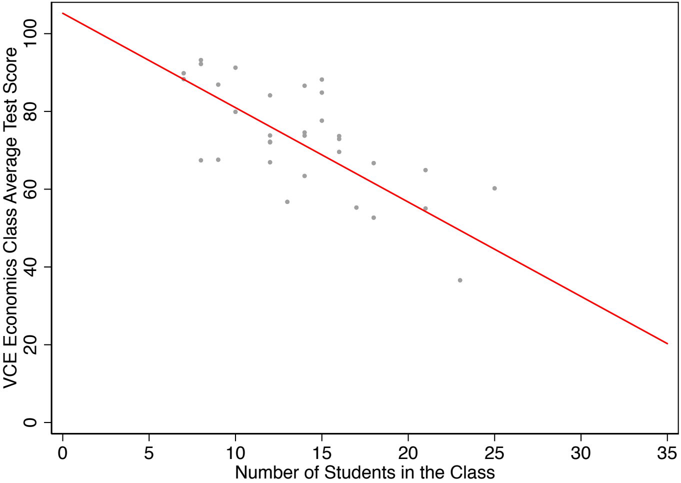


1. Take a sub-sample of schools with average household income between $29,000 and $31,000

2. Plot test scores versus class size for this sub-sample with similar income levels

Fixing Omitted Variable Bias

Test score – class size relationship based on the sub-sample



Fixing Omitted Variable Bias

- In the sub-sample we focused on, income is now similar across classes, so if we find a negative relationship between $ClassSize_i$ and $TestScore_i$ in the sub-sample, then it is more likely to be driven by a direct relationship rather than an indirect relationship because of income differences
- This highlights the idea of holding income fixed in estimating the direct class size – test score relationship

Multiple Linear Regression

- The multiple linear regression model extends the single linear regression model to include additional variables as regressors
- The model allows us to estimate the effect on Y_i of changing one variable (X_{1i}) while holding other regressors ($X_{2i}, X_{3i}, X_{4i}, \ldots$) constant (or fixed)
- In our example, we can use multiple linear regression to isolate the effect on test scores (e.g., Y_i) of class size (e.g., X_{1i}) while holding household income (e.g., X_{2i}) fixed
- In this way, multiple linear regression is a tool for eliminating omitted variable bias

Population Regression Model

- Population regression model with k regressors is defined as:

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

- For expositional purposes, in what follows we will work with a regression including two regressors, X_{1i}, X_{2i}.
  - However, everything I discuss immediately extends to the general case with k regressors $X_{1i}, X_{2i}, \ldots X_{ki}$
  - Also, X_{1i} and X_{2i} may be any combination of continuous or dummy variables

- Population regression model with $k = 2$ regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \ldots, n$$

- From our test scores example with $i = 1, \ldots, n$ classes:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + u_i, \quad i = 1, \ldots, n$$

Control Variables

- Taking conditional expectations at a point where $X_{1i} = x_1$ and $X_{2i} = x_2$, the population regression function is:

  $$E(Y_i|X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

  where β_1 and β_2 are regression coefficients on X_{1i} and X_{2i}
- We often refer to some of the regressors in a multiple linear regression as control variables
- Interpreting β_1, we say it is the relationship between X_{1i} on Y_i holding X_{2i} fixed (or, equivalently, controlling for X_{2i})
  - From our example, if we used multiple linear regression, β_1 would be the relationship between class size and test score, controlling for income

Coefficient Interpretation

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

- When we interpret coefficients, we imagine changing only one regressor at a time, leaving the others fixed
- For example, we increase X_{1i} by ΔX_1 and leave X_{2i} fixed
- Changing X_{1i} we lead to an expected change in Y_i, which we label ΔY, and is defined from the population regression line:

$$Y_i + \Delta Y = \beta_0 + \beta_1 (X_{1i} + \Delta X_1) + \beta_2 X_{2i} \tag{2}$$

- Subtracting equations (1) and (2), we have:

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \text{ holding } X_2 \text{ constant}$$

- β_1 is the expected change in Y_i from a one-unit change in X_{1i}.
- It is often called the partial effect of X_{1i} on Y_i, which emphasises our focus on changing just one regressor while holding all other regressors fixed

The Constant

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

- The intercept β_0 is called the constant term and it is interpreted as the average value of Y_i when $X_{1i} = 0$ and $X_{2i} = 0$
- We can equivalently write the regression including a third regressor X_{0i} which is a dummy variable that equals one for all observations:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where X_{0i} is often called the constant regressor

Heteroskedasticity

- The error term in the regression is homoskedastic if the variance conditional on all of the regressors, $var(u_i|X_{1i}, X_{2i}, \ldots, X_{ki})$ is constant for $i = 1, \ldots, n$
- Otherwise, the error term is heteroskedastic
- We will continue to work under the more general assumption of heteroskedasticity in computing standard errors and conducting statistical inference (next lecture note)

Student Test Score Example

- We could imagine adding additional regressors to our test score regression:

$$\begin{aligned} TestScore_i = & \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 ParentEduc_i + \\ & + \beta_4 FamilySize_i + \beta_5 ShareFemale_i + \beta_6 ShareImmig \\ & + \beta_7 Urban_i + \beta_8 Private_i + u_i, \quad i = 1, \ldots, n \end{aligned}$$

- Richer and richer regressions allows us to control for (or hold fixed) many other variables that predict test scores in avoiding omitted variable bias to isolate the relationship between $ClassSize_i$ on $TestScore_i$, which is the main effect of interest

OLS Estimation with Multiple Linear Regression

- Just like with the singe linear regression, we use the Ordinary Least Squares (OLS) estimator to estimate the regression coefficients of a multiple linear regression model
- Recall that the OLS estimator aims to find the regression coefficients that together minimise the mistakes the model makes in predicting the dependent variable Y_i given the k regressors $X_{1i}, X_{2i}, \ldots, X_{ki}$
- For a given set of regression coefficients, $b_0, b_1, b_2, \ldots, b_k$, the model's mistake in predicting Y_i is:

$$Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \ldots - b_k X_{ki}$$

- The sum of squared prediction mistakes across all $i = 1, \ldots, n$ observations is:

$$\sum_{i=1}^{n} \left(\underbrace{Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \ldots - b_k X_{ki}}_{\text{prediction mistake}} \right)^2$$

OLS Estimation with Multiple Linear Regression

- The OLS estimators of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ correspond to the $b_0, b_1, b_2, \ldots, b_k$ values that together minimise the sum of squared prediction mistakes
- As usual, the OLS estimators are denoted by $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$
- The OLS regression function is the (k-dimensional) line constructed using the OLS estimators:

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

- The OLS predicted value of Y_i given $X_{1i}, X_{2i}, \ldots, X_{ki}$ is:

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

- The OLS residual for observation i is the difference between Y_i in the data and the model's predicted value for Y_i

$$\hat{u}_i = Y_i - \hat{Y}_i$$

Test Scores and Class Size Example

- Single linear regression of test scores on class size

$$\widehat{TestScore}_i = \underset{(1.84)}{105.24} - \underset{(0.12)}{2.43}\, ClassSize_i$$

Note: slightly different dataset than from previous lectures

- Multiple linear regression of test scores on class size and average income

$$\widehat{TestScore}_i = \underset{(6.15)}{73.99} - \underset{(0.14)}{1.96}\, ClassSize_i + \underset{(0.16)}{0.82}\, Income_i$$

- The inclusion of $Income_i$ as a regressor causes the magnitude of the coefficient on $ClassSize_i$ to fall as omitted variable bias related to $Income_i$ in the regression is accounted for

Measures of Fit in Multiple Linear Regression

- R^2, which recall is the fraction of the sample variance in Y_i that is explained or predicted by the regressors, is again a common measure of fit for multiple linear regression
- It is computed identically in multiple linear regression:

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

where the explained sum of squares $ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ and the total sum of squares $TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

Measures of Fit in Multiple Linear Regression

- The R^2 always rises when regressors are added to the regression unless the estimated coefficient on a regressor is exactly 0 (which very rarely happens in practice)
- Because of this, we often work the adjusted R^2, denoted $\bar{R}^2$, which is a modified version of R^2 that does not necessarily increase when a new regressor is added
- The adjusted R^2 is computed as:

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

- The difference with the standard $R^2 = 1 - \frac{SSR}{TSS}$ formula is the multiplication by $\frac{n-1}{n-k-1}$

Measures of Fit in Multiple Linear Regression

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

- $\bar{R}^2$ is always less than than R^2 and therefore always less than 1
- Adding a regressor to the regression has two effects on $\bar{R}^2$:
    - SSR falls, which causes $\bar{R}^2$ to rise
    - $\frac{n-1}{n-k-1}$ (because k goes up), which causes $\bar{R}^2$ to fall
- $\bar{R}^2$ can actually be negative if all the regressors together do not decrease enough SSR to offset the $\frac{n-1}{n-k-1}$ factor

Measures of Fit in Multiple Linear Regression

- We can also measure fit of the regression using the standard error of the regression (SER) which estimates the standard deviation of the error term u_i:

  $$SER = \sqrt{s_{\hat{u}}^2}, \text{where } s_{\hat{u}}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{SSR}{n-k-1}$$

  where n is the number of observations, and k is the number of regression coefficients beyond the constant, and $SSR = \sum_{i=1}^{n} \hat{u}_i^2$ is the sum of squared residuals
- The use of $n-k-1$ in the denominator in computing SER is called a degrees of freedom adjustment (one for every one of the k variables in the model, and one more for the constant)

Test Scores and Class Size Example

- Regression results from single linear regression of test scores on class size

$$\widehat{TestScore}_i = \underset{(1.84)}{105.24} - \underset{(0.12)}{2.43}\, ClassSize_i, \quad R^2 = 0.65, \bar{R}^2 = 0.64$$

Note: slightly different dataset than from previous lectures

- Regression results from multiple linear regression of test scores on class size and average income

$$\widehat{TestScore}_i = \underset{(6.15)}{73.99} - \underset{(0.14)}{1.96}\, ClassSize_i + \underset{(0.16)}{0.82}\, Income_i \;\; R^2 = 0.69, \bar{R}^2 = 0.68$$

- The inclusion of $Income_i$ as a regressor causes the magnitude of the coefficient on $ClassSize_i$ to fall as omitted variable bias related to $Income_i$ in the regression is accounted for

Beware Interpretations of R^2 and $\bar{R}^2$

- $\bar{R}^2$ is useful because it summarises the extent to which the regressors explain the variation in the dependent variable
- However, "maximising" the $\bar{R}^2$ in practice is rarely the goal to economically or statistically addressing most questions
- In fact, if $\bar{R}^2$ is very close to 1 is often a sign that there is a logical problem with the regression model!
- We return to strategies for evaluating econometric models later in the subject, but having an extremely high $\bar{R}^2$ is typically <u>not</u> the goal of econometric analysis
- For the remainder of the subject, we will focus on reporting $\bar{R}^2$ for empirical analyses in tutorials and on assignments

The Least Squares Assumptions in Multiple Linear Regression

- As with single linear regression, multiple linear regression relies on 4 key assumptions that are critical for the sampling distributions of the OLS estimators, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$
- Assumption 1: Independence

$$E[u_i | X_{1i}, X_{2i}, \ldots, X_{ki}] = 0$$

- Assumption 2: $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ are IID
- Assumption 3: Large Outliers Are Unlikely
- Assumption 4: No Perfect Multicollinearity (new)

Perfect Multicollinearity

- Two regressors exhibit perfect multicollinearity if one of the regressors is a perfect linear combination of other regressors
- Assumption 4 requires that no regressors exhibit perfect multicollinearity
- Example: suppose you tried to run this regression by accident:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 ClassSize_i + u_i$$

  where the two regressors are perfectly collinear
- Conceptually, β_1 is the impact of $ClassSize_i$ (e.g,. the first regressor) on $TestScore_i$ holding $ClassSize_i$ (e.g, the second regressor) fixed, which does not make any sense
- Same problem if you attempt to estimate, for example:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 (Income_i/2) + u_i$$

Perfect Multicollinearity

- In general, if a group of regressors are perfectly collinear, then it is impossible to hold one regressor fixed to estimate the effect of one of the other collinear regressors on the dependent variable
- In practice, statistics software will either drop one of the perfectly multicollinear variables or will give an error message if you try to run a regression with perfect multicollinearity
- We fix perfect multicollinearity problems by modifying the set of regressors to eliminate the problem.

Perfect Multicollinearity Example: Huge Classes

- Supposed we created a dummy variable $HugeClass_i$ which equals one if a class has more than 35 students and is 0 otherwise. Here is the regression:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 HugeClass_i + u_i$$

- The regression exhibits perfect multicollinearity. Why?
  - In our sample, no class has more than 35 students, which means $HugeClass_i = 0$ for all observations
  - Recall the constant β_0 in the regression is equivalent to having $\beta_0 X_{0i}$ in the regression, where X_{0i} is the constant regressor that is always equal to 1
  - Therefore, $HugeClass_i = 1 - X_{0i}$ and we have perfect multicollinearity
- Two important aspects of this example:
  1. Perfect multicollinearity can arise because of the constant
  2. Perfect multicollinearity is specific to the dataset you have at hand; we could imagine classes with more than 35 students

Dummy Variable Trap

- A possible source of multicollinearity arises when multiple dummy variables are used as regressors
- For example, suppose the schools where split into either being urban or rural schools, and you created two dummy variables:
    - $Urban_i = 1$ if school i is in an urban location and is 0 otherwise
    - $Regional_i = 1$ if school i is in an regional location and is 0 otherwise
    - For each school, either:
        - $Urban_i = 1$ and $Regional_i = 0$ (urban school)
          OR
        - $Urban_i = 0$ and $Regional_i = 1$ (regional school)
- Therefore, the sum of the two dummy variables equals 1:
  $Urban_i + Regional_i = 1$ for all $i = 1, \ldots, n$

Dummy Variable Trap

- Suppose you tried to run the regression:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 Urban_i + \beta_4 Regional_i + u_i$$

- You again face multicollinearity because of the constant; $Urban_i + Regional_i = 1 = X_{0i}$ for $i = 1, \ldots, n$
- That is, the urban and regional dummy variables add up to equal the constant regressor for each observation $i = 1, \ldots, n$

Dummy Variable Trap

- The situation when a group of dummy variables add up to always equal another dummy variable (or the constant regressor) is called the dummy variable trap
- You can avoid the dummy variable trap by dropping one of the dummy variables (or dropping the constant):

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 Urban_i + u_i$$

  where $Regional_i$ is dropped, and β_3 is the difference in test scores in urban schools relative to regional schools, holding all other regressors constant

- In general, if:
  - there are G dummy variables
  - each observation falls into one and only category
  - there is an intercept in the regression
  - all G binary variables are included in the regression

  then the regression fails because of perfect multicollinearity as you have falled into the dummy variable trap

Avoiding the Dummy Variable Trap

- We typically avoid the dummy variable trap by only including $G-1$ of the G dummy variables in the regression
- The dummy that we do not include in the regression is the base category or base group or omitted category
- We interpet all the other dummies as the change in the outcome variable when a given dummy variable is equal to 1 relative to base group, holding all other regressors constant
- Alternatively, we can include all G dummies and drop the constant in the regression (very uncommon)
- In sum, when your software indicates that you have perfect multicollinearity, it is important to eliminate it by:
    1. determining the source of perfect multicollinearity
    2. creating a base group
    3. ensuring you properly interpret regression coefficients for the dummies in the regression relative to the omitted base group

Multicollinearity

- A related issue is imperfect multicollinearity, which arises when one of the regressors is highly correlated, but not perfectly correlated, with other regressors
- This does not prevent statistics programs from providing OLS estimates, but it does result in regression coefficients being estimated imprecisely and having large standard errors, and therefore statistically insignificant regression coefficients
- Intuitively, if two regressors are highly correlated and almost always co-moving together, it is hard to disentangle their individual impacts on the dependent variable in the regression
- Whereas perfectly multicolliearity typically arises because of a logical mistake in your regression set-up, imperfect multicollinearity is not necessarily an error but a feature of your data, OLS, and the question you are trying to address

Distribution of OLS Estimators in Multiple Linear Regression

- Because random samples vary from one sample to the next, different samples product different values for the OLS estimators, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$
- That is, these estimators are random variables with a distribution
- Under the 4 least squares assumptions, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are unbiased and consistent estimators of their population true values $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$
- In large samples, the sampling distribution of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ is well approximated by a multivariate normal distribution, with each $\hat{\beta}_j$ having a marginal distribution that is $N(\beta_j, \sigma^2_{\hat{\beta}_j})$ for $j = 0, 1, 2, \ldots, k$
- We can use these results to conduct hypothesis tests with multiple linear regression models using t-statistics and p-values similar to what we did with single linear regression models