HANDOUT 18
Multiple Regression III – Various Topics

1. **Introduction**
2. **Goodness of Fit**
3. **The Standard Error of OLS Estimators**

Source: Wooldridge (Ch 3), Hughes-Hallett (Math camp handouts)

1. INTRODUCTION

- Today we study 2 broad topics related to estimation in the context of multiple regression:
  - Goodness of fit (the famous R^2)
  - Variance of OLS estimators

2. GOODNESS OF FIT

- Consider the following terms:

  Total Sum of Squares = $TSS = \sum(Y_i - \bar{Y})^2$

  Explained sum of squares= $ESS = \sum(\hat{Y}_i - \bar{Y})^2$

  Residual sum of squares= $RSS = \sum \hat{u}_i{}^2$

- It turns out that TSS=ESS+RSS. (See Wooldridge for proof)
- The R-squared is defined to be

$$R^2 = \frac{ESS}{TSS}$$

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i{}^2}{\sum(Y_i - \bar{Y})^2}$$

- By definition R^2 is a number between zero and one (because TSS = ESS + RSS, ESS $\geq$ 0 and RSS $\geq$ 0).
- Interpretation of R^2: proportion of the sample variation in y that is explained by the OLS regression line.
- R^2 can also be shown to equal the ***squared correlation*** coefficient between the actual Y_i and the fitted values $\hat{Y}_i$. This is where the term "R-squared" comes from.

Example – Smoking and Lung Cancer

```
. regress lcd cigs, robust
Regression with robust standard errors                 Number of obs =       5
                                                       F(  1,      3) =   22.59
                                                       Prob > F      =  0.0177
                                                       R-squared     =  0.8658
                                                       Root MSE      =  63.921

------------------------------------------------------------------------------
             |               Robust
         lcd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |   .3445158    .072487     4.75   0.018     .1138297    .5752019
       _cons |     20.217   52.79902     0.38   0.727     -147.813     188.247
------------------------------------------------------------------------------
```

QUESTION: How do we interpret the R^2 in this particular example?

QUESTION: What happens to R^2 when an explanatory variable is added to a regression?

A. It must increase
B. It increases or stays the same
C. It must decrease
D. It decreases or stays the same
E. Not enough information provided

- Adjusted R^2: Penalizes you for using irrelevant explanatory variables
- R^2 provides a measure of how well the OLS line fits the data
  - An R^2=1 means all the points lie on the same line, i.e. OLS provides a perfect fit to the data
  - An R^2 close to zero means a poor fit of the OLS line

QUESTION: The larger the R^2, the lower the likelihood that our regression suffers from omitted variable bias (OVB)

A. True
B. False
C. I don't know

3. THE STANDARD ERROR OF OLS ESTIMATORS

Idea: The discussion of unbiasedness gives us an assessment of the central tendencies of $\hat{\beta}_j$. Now we would like to have a measure of the spread in the sampling distribution of $\hat{\beta}_j$.

Key idea: All else equal, we would like an estimator of $\hat{\beta}_j$ that has a low standard error. Why?

We first add an assumption to our model called **homoskedasticity**. We do so for two reasons:

(1) The formulas for the standard error of $\hat{\beta}_j$ are simplified, which allows us to develop more easily the intuition behind the determinants of the standard error
(2) OLS has important efficiency properties under the homoskedasticity assumption (see below)

ASSUMPTION MLR.5 [HOMOSKEDASTICITY]

$$Var[u|X_1, X_2, \dots, X_k] = \sigma^2$$

If this assumption fails, then the model exhibits heteroskedasticity. See Appendix #3 for details.

Assumptions MLR.1 through MLR.5 are collectively known as the Gauss-Markov assumptions (for cross-sectional regression)

Efficiency of OLS: The Gauss-Markov Theorem

Under assumptions MLR.1 through MLR.5, $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are the Best Linear Unbiased Estimators (BLUEs) of $\beta_0, \beta_1, \dots, \beta_k$ respectively.

Best: lowest variance
Linear: Can be expressed as a linear function of the data on the dependent variable
Unbiased: $E(\hat{\beta}_j) = \beta_j$
Estimator: Rule/Method/Formula that can be applied to any sample to produce an estimate

Key idea: The importance of the Gauss-Markov Theorem is that, when the standard set of assumptions holds, we need not look for alternative linear unbiased estimators: none will be better than OLS.

Terminology

For the purposes of the next section, it will be helpful to think about various R^2s, which we define here. Consider the following regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

The following R^2s can be defined:

Name	R^2 computed from the following regression:
R^2	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$
R_1^2	$X_1 = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_3 + v$
R_2^2	$X_2 = \delta_0 + \delta_1 X_1 + \delta_2 X_3 + \varepsilon$
R_3^2	$X_3 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \eta$

More generally, R_j^2 is the R-squared from regressing X_j on all other explanatory variables (and including an intercept).

QUESTION: When would you expect R_j^2 to be large?

THEOREM 3.2 [Sampling variances of the OLS slope estimators]

Under assumptions MLR.1 through MLR.5, conditional on the sample values of the explanatory variables,

$$Std.Error(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{TSS_j(1 - R_j^2)}} \qquad \mathbf{(3.51)}$$

for *j=1,2,…,k*, where $TSS_j = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$ is the total sample variation in X_j, and R_j^2 is the R-squared from regressing X_j on all other explanatory variables (and including an intercept).

Note: The proof of theorem 3.2 can be found in Wooldridge.

FORMULA FOR STANDARD ERROR	EXAMPLE
$$Std.Error(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{TSS_j(1-R_j^2)}}$$	

Determinant of Standard Error	Analysis	Sign of Relationship with Standard Error
(1) The variance of the error term (σ^2)		
(2) The Total Sample Variation in X_j (TSS_j): $$TSS_j = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$$		
(3) The Linear Relationships Among the Explanatory Variables (R_j^2)		

THE COMPONENTS OF THE STANDARD ERROR OF OLS ESTIMATORS

Eq. (3.51) shows that the standard error of $\hat{\beta}_j$ depends on three factors: σ^2, TSS_j, and R_j^2. We now consider each of these factors separately.

(1) The variance of the error term (σ^2)

Key: σ^2 is a feature of the population; it has nothing to do with sample size.

(2) The Total Sample Variation in X_j (TSS_j):

$$TSS_j = \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right)^2$$

Everything else equal, for estimating β_j, we prefer to have as much variation in X_j as possible. When sampling randomly from the population, TSS_j increases with sample size.

(3) The Linear Relationships Among the Explanatory Variables (R_j^2)

It is important to see that this R-squared is distinct from the R-squared in the regression of Y on X_1, X_2,…X_k.

Extreme cases:

- $R_j^2 = 0$ [smallest Var ($\hat{\beta}_j$) for a given σ^2 and TSS_j]
- $R_j^2 = 1$ [violates assumption MLR.3]

Key case: When R_j^2 is "close" to 1, $Var(\hat{\beta}_j)$ might become too large. High (but not perfect) correlation between two or more of the independent variables is called ***multicollinearity***.

Key idea #1: Worrying about high degrees of correlation among the independent variables in the sample is really no different from worrying about a small sample size: both work to increase $Var(\hat{\beta}_j)$.

Example: Estimating the effect of school expenditure categories on student performance.

Key idea #2: A high degree of correlation between certain explanatory variables can be irrelevant as to how well we can estimate other parameters in the model. For example, consider:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Say X_2 and X_3 are highly correlated. Then $Var(\hat{\beta}_2)$ and $Var(\hat{\beta}_3)$ may be large. But the amount of correlation between X_2 and X_3 has no direct effect on $Var(\hat{\beta}_1)$.

Suppose we estimate the following regression:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u}$$

Adding an explanatory variable $\boldsymbol{X_4}$ that is correlated with $\boldsymbol{X_1}$ will:

A. Increase the standard error of $\boldsymbol{\widehat{\beta}_1}$
B. Have no effect on the standard error of $\boldsymbol{\widehat{\beta}_1}$
C. Decrease the standard error of $\boldsymbol{\widehat{\beta}_1}$
D. Not enough information given
E. I don't know

Standard Errors in Misspecified Models

Key idea: The choice of whether or not to include a particular variable in a model can sometimes be made by analyzing the tradeoff between bias and variance.

Estimating the Standard Errors of the OLS Estimators

Problem: The formula for $Std\ Error(\hat{\beta}_j)$ (and hence the formula for the standard error) depends on σ^2, which we don't observe since it's a population parameter.

Solution: Obtain an unbiased estimator of σ^2, which will then allow us to obtained unbiased estimators of $Std\ Error(\hat{\beta}_j)$. See Appendix #4 for details.

Key Ideas

- **Goodness of fit (R2)**: What it is and what it is not.
- **Standard Errors**:
  - We care about *magnitude* of coefficient but also about *standard error*
  - Important to understand *determinants* of standard errors to be able to better design and consume empirical studies
  - *Tradeoff* between *bias* and *variance*

APPENDIX #1– OLS IN MATRIX NOTATION

(Adapted from Johnston and Hughes Hallett)

- In this course, we have expressed the linear PRF for a regression with k explanatory variables in the following form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \qquad (4)$$

- We can write (4) using matrix algebra. This may be useful to you for two reasons:
    - Both in API-210 and in many academic papers you will see the PRFs written in matrix algebra form, so it is important for you to be familiar with this notation
    - Matrix algebra allows us to specify how to compute the OLS estimators when we have more than one explanatory variable in our PRF
- There are several matrix algebra notations used. We will focus on two that are commonly used:
    - Notation #1: Will be used in API-210 and has some computational advantages. This notation will be covered by Deb Hughes Hallett in Math Camp.
    - Notation #2: Used in classic textbooks such as Johnston and Greene.

Notation #1

- You can write the PRF: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$ in the following way: $y_i = x_i \beta + \varepsilon_i$, where:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad x_i = \begin{bmatrix} 1 \\ x_{1i} \\ x_{2i} \\ \vdots \\ \vdots \\ x_{ki} \end{bmatrix}$$

i denotes the observation, and ‘ denotes the transpose of the matrix.

- The OLS estimators from the linear PRF $y_i = x_i' \beta + \varepsilon_i$ can be computed as follows:

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i X_i' \right)^{-1} \sum_{i=1}^{N} X_i Y_i$$

Notation #2

The hypothesized model is:

$$y = X\beta + u$$

Where

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_n \end{bmatrix}$$

The OLS estimator of the population parameters represented in the vector β is given by:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

and under certain conditions the variance of this estimator is given by:

$$Var(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$$

APPENDIX #2 - STUDIES ABOUT CLASS SIZE AND TEST SCORES

Study #1 - Randomized Experiment in Tennessee (STAR)

```
. reg tscorek sck, robust;
Regression with robust standard errors                Number of obs =    5786
                                                       F(  1,  5784) =   40.67
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0073
                                                       Root MSE      =  73.483

------------------------------------------------------------------------------
             |               Robust
     tscorek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sck |   13.74055   2.154628     6.38   0.000     9.516677    17.96443
       _cons |   918.2013   1.135073   808.94   0.000     915.9762    920.4265
------------------------------------------------------------------------------
```

sck: dummy for small class size

Study #2 - Observational Study in California

```
. reg testscr str, robust;

Regression with robust standard errors                Number of obs =     420
                                                       F(  1,   418) =   19.26
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0512
                                                       Root MSE      =  18.581

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

str: student-teacher ratio

APPENDIX #3 – HETEROSKEDASTICITY

- Note that the standard error formula in (3.58) is not a valid estimator of sd($\hat{\beta}_j$) if the errors exhibit heteroskedasticity. Thus, while the presence of heteroskedasticity does not lead to bias in $\hat{\beta}_j$, it does lead to bias in the usual formula for the variance of $\hat{\beta}_j$, which then invalidates the standard errors.
- There are statistical tests to assess the presence of heteroskedasticity (see chapter 8 of Wooldridge for details).
- However, for the purposes of this course, we will adopt Stock and Watson's guideline of always calculating standard errors assuming the presence of heteroskedasticity. These are called heteroskedasticity-robust standard errors.
- The heteroskedasticity-robust standard error formula is:

$$se(\hat{\beta}_j) = \sqrt{\frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{RSS_j^2}}$$

  Where $\hat{r}_{ij}^2$ denotes the square of the residual from regressing X_j on all other explanatory variables, and RSS_j^2 is the sum of squared residuals from this regression.
- In Stata you get this standard error by using the **"robust" option** when you run a regression. For example, "regress lcd cigs, robust"

APPENDIX #4 - ESTIMATING THE STANDARD ERRORS OF THE OLS ESTIMATORS

Problem: The formula for $Std\ Error(\hat{\beta}_j)$ (and hence the formula for the standard error) depends on σ^2, which we don't observe since it's a population parameter.

Solution: Obtain an unbiased estimator of σ^2, which will then allow us to obtained unbiased estimators of $Std\ Error(\hat{\beta}_j)$.

The unbiased estimator of σ^2 in the general multiple regression case is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}^2}{(n-k-1)}$$

where n = number of observations and k = number of explanatory variables

The term n-k-1 is the ***degrees of freedom (df)*** for the general OLS model with n observations and k explanatory variables.

Standard error of $\hat{\boldsymbol{\beta}}_j$:

$$Std\ Error(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{TSS_j(1-R_j^2)}} \qquad (3.58)$$