

# Multiple Regression Analysis

Caio Vigo

**The University of Kansas**  
Department of Economics

Spring 2020

These slides were based on *Introductory Econometrics* by Jeffrey M. Wooldridge (2015)

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

## 1 Motivation for Multiple Regression

The Model with  $k$  Independent Variables

## 2 Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

## 3 The Expected Value of the OLS Estimators

## 4 The Variance of the OLS Estimators

Estimating the Error Variance

## 5 Efficiency of OLS: The Gauss-Markov Theorem

## Motivation:

- With a simple linear regression model we learned a model in which a **single** independent variable  $x$  explains (or affect) a dependent variable  $y$ .
- If we add more factors to our model that are useful for explaining  $y$ , then more of the variation in  $y$  can be explained.

We can build better models for predicting the dependent variable.

- Recall the  $\log(wage)$  example.

## Example: $\log(wage)$

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

- Might be the case that there are factors in  $u$  affecting  $y$ .
- For instance intelligence could help to explain  $wage$ .

- Let's use a **proxy** for it:  $IQ$ .
- By explicitly including  $IQ$  in the equation, we can take it out of the error term.
- Consider the following extension of the  $\log(wage)$  example:

**Example:**  $\log(wage)$  (extension)

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + u$$

Generally, we can write a model with two independent variables as:

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

where

$\beta_0$  is the intercept,

$\beta_1$  measures the change in  $y$  with respect to  $x_1$ , *holding other factors fixed*,

$\beta_2$  measures the change in  $y$  with respect to  $x_2$ , *holding other factors fixed*

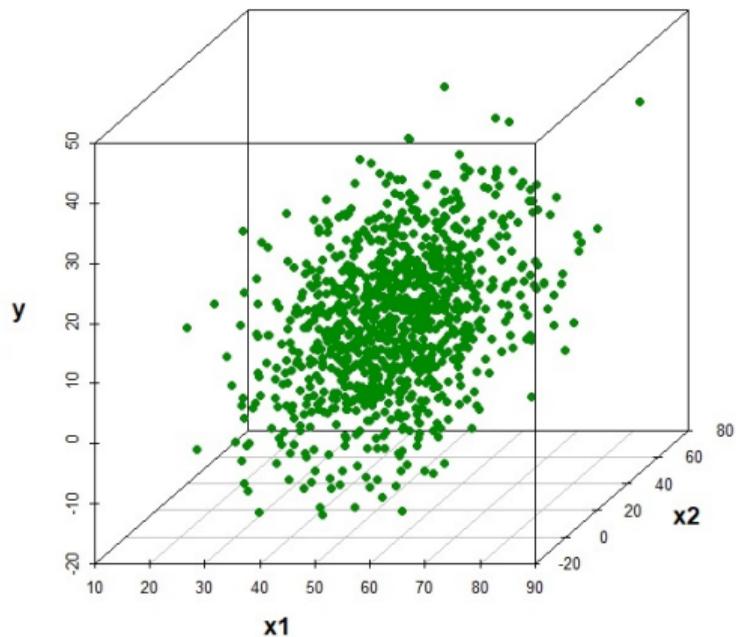
- In the model with two explanatory variables, the key assumption about how  $u$  is related to  $x_1$  and  $x_2$  is:

$$E(u|x_1, x_2) = 0.$$

- For any values of  $x_1$  and  $x_2$  in the population, the average unobservable is equal to zero.
- The value zero is not important because we have an intercept,  $\beta_0$  in the equation.

- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

3D Scatterplot



# The Model with 2 Independent Variable

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

- In the wage equation, the assumption is  $E(u|educ, IQ) = 0$ .
- Now  $u$  no longer contains intelligence, and so this condition has a better chance of being true.
- Recall that in the simple regression, we had to assume  $IQ$  and  $educ$  are unrelated to justify leaving  $IQ$  in the error term.
- Other factors, such as workforce experience and “motivation,” are part of  $u$ . Motivation is very difficult to measure. Experience is easier:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + u.$$

# The Model with $k$ Independent Variables

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

- The **multiple linear regression model** can be written in the population as

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

where,

$\beta_0$  is the **intercept**,

$\beta_1$  is the parameter associated with  $x_1$ ,

$\beta_2$  is the parameter associated with  $x_2$ , and so on.

- Contains  $k + 1$  **(unknown) population parameters**.
- We call  $\beta_1, \dots, \beta_k$  the **slope parameters**.

- Now we have **multiple explanatory** or **independent variables**  $x'$ s.
- We still have **one explained** or **dependent variable**  $y$ .
- We still have **an error term**,  $u$ .

- **Advantage of multiple regression:** it can incorporate fairly general functional form relationships.

- Let  $lwage = \log(wage)$ :

$$lwage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 exper^2 + u,$$

so that  $exper$  is allowed to have a quadratic effect on  $lwage$ .

- Thus,  $x_1 = educ$ ,  $x_2 = IQ$ ,  $x_3 = exper$ , and  $x_4 = exper^2$ . Note that  $x_4$  is a *nonlinear* function of  $x_3$ .

- The key assumption for the general multiple regression model is:

$$E(u|x_1, \dots, x_k) = 0$$

- We can make this condition closer to being true by “controlling for” more variables.

Motivation for  
Multiple  
Regression

The Model with  $k$   
Independent  
Variables

Mechanics and  
Interpretation  
of OLS

Interpreting the OLS  
Regression Line

The Expected  
Value of the  
OLS  
Estimators

The Variance  
of the OLS  
Estimators

Estimating the Error  
Variance

Efficiency of  
OLS: The  
Gauss-Markov  
Theorem

- ① Motivation for Multiple Regression  
The Model with  $k$  Independent Variables
- ② Mechanics and Interpretation of OLS  
Interpreting the OLS Regression Line
- ③ The Expected Value of the OLS Estimators
- ④ The Variance of the OLS Estimators  
Estimating the Error Variance
- ⑤ Efficiency of OLS: The Gauss-Markov Theorem

- Suppose we have  $x_1$  and  $x_2$  ( $k = 2$ ) along with  $y$ .
- We want to fit an equation of the form:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

given data  $\{(x_{i1}, x_{i2}, y_i) : i = 1, \dots, n\}$ .

- Sample size =  $n$ .

## Labels and indexing

Now the explanatory variables have two subscripts:

- $i$  = observation number
- $j$  = labels for particular variables (it is the second subscript - 1 and 2 in this case)

For example:

$$x_{i1} = educ_i, \quad i = 1, 2, \dots, n$$

$$x_{i2} = IQ_i, \quad i = 1, 2, \dots, n$$

- Motivation for Multiple Regression
- The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
- Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
- Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

## Least Squares Method

- As in the simple regression case, different ways to motivate OLS. We choose  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$  (so three unknowns) to minimize the sum of squared residuals,

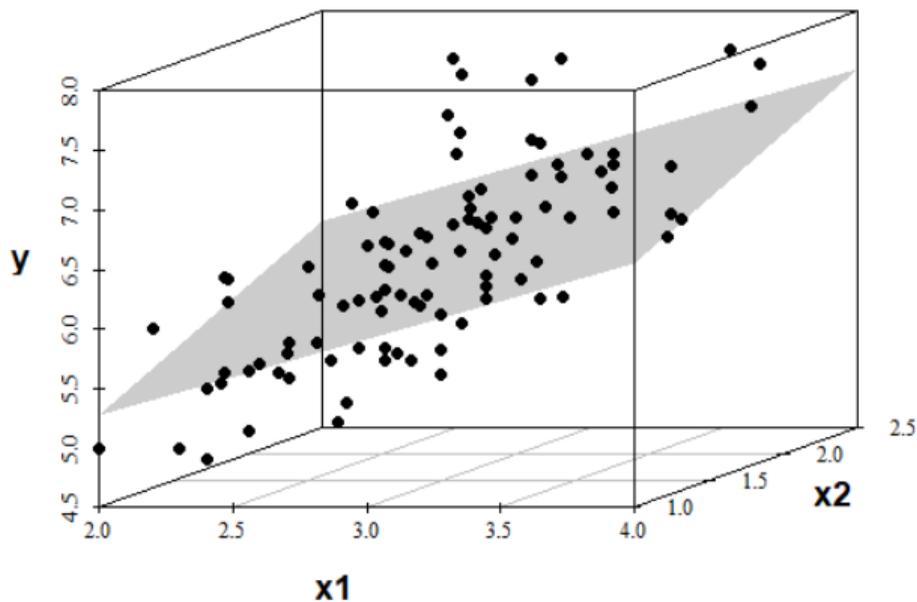
$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2$$

- The case with  $k$  independent variables is easy to state: choose the  $k + 1$  values  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ ,  $\hat{\beta}_2, \dots, \hat{\beta}_k$  to minimize

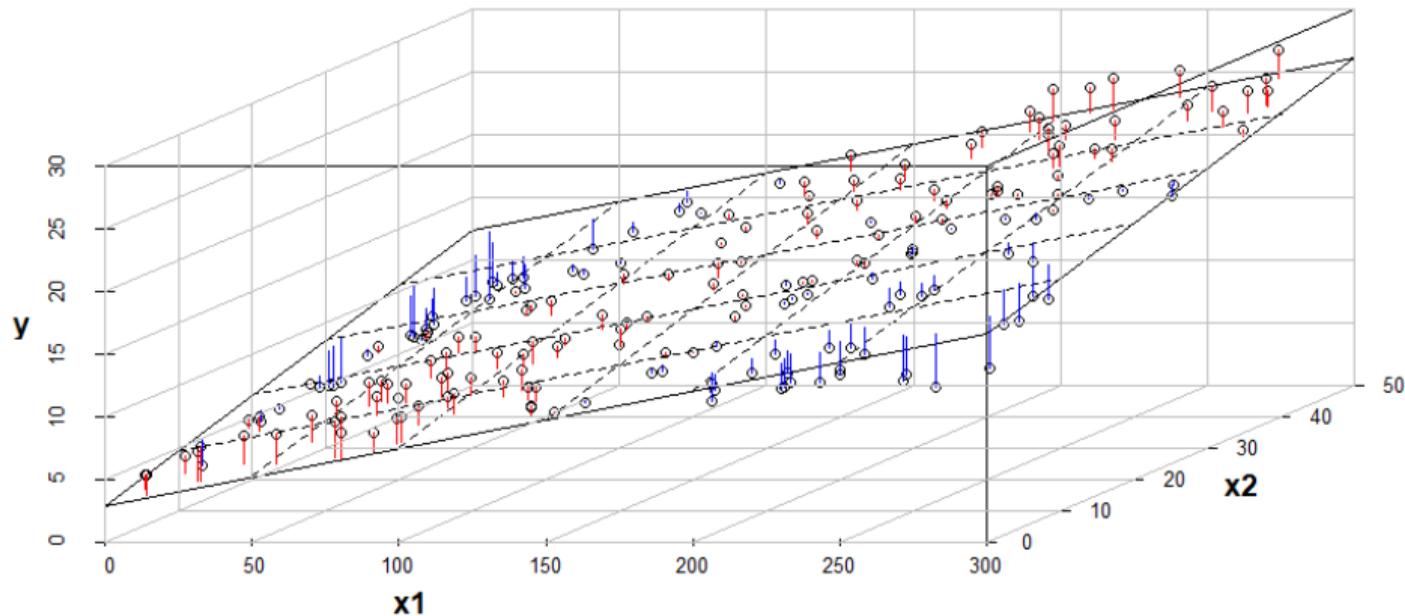
$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik})^2$$

- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

**Regression Plane**



- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem



- Motivation for Multiple Regression  
The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS  
Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators  
Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

- The **OLS first order conditions** (solved with multivariable calculus) are the  $k + 1$  linear equations in the  $k + 1$  unknowns  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ :

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^n x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

$$\sum_{i=1}^n x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

$$\vdots = \vdots$$

$$\sum_{i=1}^n x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

- As long as we add an assumption (**MLR.3** - we will see in the next topic), we can guarantee this system to have a unique solution.
- We will not find a closed solution to each  $\beta_j$  , for  $j = 0, 1, 2, \dots, k$ .
- We can use matrix algebra to easily find the solution.

The OLS regression line is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

- The slope coefficients now explicitly have ceteris paribus interpretations.
- Consider  $k = 2$ :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Then

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

allows us to compute how predicted  $y$  changes when  $x_1$  and  $x_2$  change by any amount.

- What if we “hold  $x_2$  fixed,” that is, its change is zero,  $\Delta x_2 = 0$ ?

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 \text{ if } \Delta x_2 = 0$$

In particular,

$$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x_1} \text{ if } \Delta x_2 = 0$$

In other words,  $\hat{\beta}_1$  is the slope of  $\hat{y}$  with respect to  $x_1$  when  $x_2$  is held fixed.

- Similarly,

$$\Delta \hat{y} = \hat{\beta}_2 \Delta x_2 \text{ if } \Delta x_1 = 0$$

and

$$\hat{\beta}_2 = \frac{\Delta \hat{y}}{\Delta x_2} \text{ if } \Delta x_1 = 0$$

- We call  $\hat{\beta}_1$  and  $\hat{\beta}_2$  **partial effects** or **ceteris paribus effects**.

## Terminology

We say that  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  are the OLS estimates from the regression

$$y \text{ on } x_1, x_2, \dots, x_k$$

or

$$y_i \text{ on } x_{i1}, x_{i2}, \dots, x_{ik}, \quad i = 1, \dots, n$$

when we want to emphasize the sample being used.

- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

- Recall the **wage** example:

## Example (Wage)

$$\widehat{wage}_n = -0.90 + 0.54 educ$$

$$n = 526, \quad R^2 = .16$$

- Then we did:

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

$$\widehat{lwage} = 0.58 + .08 educ$$

$$n = 526, R^2 = .19$$

- Let's write a multiple regression model:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

<i>Dependent variable:</i>	
lwage	
educ	0.092*** (0.007)
exper	0.004** (0.002)
tenure	0.022*** (0.003)
Constant	0.284*** (0.104)
Observations	526
R <sup>2</sup>	0.316
Adjusted R <sup>2</sup>	0.312
Residual Std. Error	0.441 (df = 522)
F Statistic	80.391*** (df = 3; 522)

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

$$\widehat{\log(wage)}_n = .284 + .092 \text{ educ} + .004 \text{ exper} + .022 \text{ tenure}$$

$$n = 526, R^2 = .32$$

## Interpretation:

- $.092$  means that, holding *exper* and *tenure* fixed, another year of education is predicted to increase  $\log(wage)$  by  $.092$ , i.e., **9.2% increase in wage**.
- Alternatively, we can take two people,  $A$  and  $B$ , with the *same exper* and *tenure*. Suppose person  $B$  has one more year of schooling than person  $A$ . Then we predict  $B$  to have a wage that is **9.2% higher**.

## What Does it Mean to “Hold Other Factors Fixed”?

- The power of multiple regression analysis is that it provides the *ceteris paribus* interpretation, even though the data have **not** been collected in a *ceteris paribus* fashion.

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + u$$

- Using the multiple regression model for wage as an example, it may seem that we actually went out and sampled people with the same *exper* and *tenure*.
- It's not the case. It's a random sample.

## Fitted Values and Residuals

- For each  $i$ ,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik}$$

$$\hat{u}_i = y_i - \hat{y}_i$$

**(1)** The sample average of the residuals is zero, i.e.,  $\sum_{i=1}^n \hat{u}_i = 0$ . This implies  $\bar{y} = \bar{\hat{y}}$ .

**(2)** Each explanatory variable is uncorrelated with the residuals in the sample. This follows from the first order conditions. It implies that  $\hat{y}_i$  and  $\hat{u}_i$  are also uncorrelated.

**(3)** The sample averages always fall on the OLS regression line:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \dots + \hat{\beta}_k \bar{x}_k$$

That is, if we plug in the sample average for each explanatory variable, the predicted value is the sample average of the  $y_i$ .

Motivation for  
Multiple  
Regression

The Model with  $k$   
Independent  
Variables

Mechanics and  
Interpretation  
of OLS

Interpreting the OLS  
Regression Line

The Expected  
Value of the  
OLS  
Estimators

The Variance  
of the OLS  
Estimators

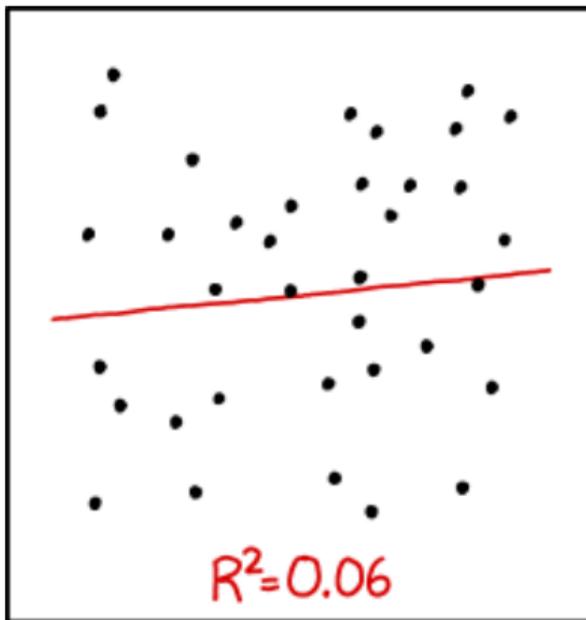
Estimating the Error  
Variance

Efficiency of  
OLS: The  
Gauss-Markov  
Theorem

$$R^2$$

**... again**

- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

## Goodness-of-Fit

- As with simple regression, it can be shown that

$$SST = SSE + SSR$$

where  $SST$ ,  $SSE$ , and  $SSR$  are the total, explained, and residual sum of squares.

- We define the  $R$ -squared as before:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- Recall,  $0 \leq R^2 \leq 1$
- Using the same set of data and the same dependent variable, the  $R^2$  **can never fall when another independent variable is added to the regression.** And, it almost always goes up, at least a little.
- This means that, if we focus on  $R^2$ , we might include silly variables among the  $x_j$ .
- Adding another  $x$  cannot make  $SSR$  increase. The  $SSR$  falls unless the coefficient on the new variable is identically zero.

Motivation for  
Multiple  
Regression

The Model with  $k$   
Independent  
Variables

Mechanics and  
Interpretation  
of OLS

Interpreting the OLS  
Regression Line

The Expected  
Value of the  
OLS  
Estimators

The Variance  
of the OLS  
Estimators

Estimating the Error  
Variance

Efficiency of  
OLS: The  
Gauss-Markov  
Theorem

- ① Motivation for Multiple Regression  
The Model with  $k$  Independent Variables
- ② Mechanics and Interpretation of OLS  
Interpreting the OLS Regression Line
- ③ The Expected Value of the OLS Estimators
- ④ The Variance of the OLS Estimators  
Estimating the Error Variance
- ⑤ Efficiency of OLS: The Gauss-Markov Theorem

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

## Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

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

where the  $\beta_j$  are the population parameters and  $u$  is the unobserved error.

## Assumption MLR.2 (Random Sampling)

We have a random sample of size  $n$  from the population,

$$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, \dots, n\}$$

- The data should be a representative sample from the population.

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

## Assumption MLR.3 (No Perfect Collinearity)

In the sample (and, therefore, in the population), none of the explanatory variables is constant, and there are **no exact linear** relationships among them.

If an independent variable in a *Multiple Regression model* is an **exact linear combination** of the other independent variables, we say the model suffers from **perfect collinearity**, and it cannot be estimated by OLS.

- Under perfect collinearity, there are no unique OLS estimators. **R**, **Stata** and other regression packages will indicate a problem.

- We must rule out the (extreme) case that one (or more) of the explanatory variables is an exact *linear* function of the others.

Usually perfect collinearity arises from a **bad specification** of the population model.

- Assumption MLR.3 can only hold if  $n \geq k + 1$ , that is, we must have at least as many observations as we have parameters to estimate.

# The 4 Assumptions for Unbiasedness

- Motivation for Multiple Regression
- The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
- Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
- Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

- Suppose that  $k = 2$  and  $x_1 = educ$ ,  $x_2 = exper$ . If we draw our sample so that

$$educ_i = 2exper_i$$

for every  $i$ , then Assumption MLR.3 is violated.

- This is very unlikely unless the sample is small.
- In any realistic population there are plenty of people whose education level is not twice their years of workforce experience.

# The 4 Assumptions for Unbiasedness

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

Do not include the same variable in an equation that is measured in different units.

## Example: CEO Salary

In a CEO salary equation, it would make no sense to include firm sales measured in dollars along with sales measured in millions of dollars. There is no new information once we include one of these.

Be careful with functional forms! Suppose we start with a constant elasticity model of family consumption:

$$\log(\text{cons}) = \beta_0 + \beta_1 \log(\text{inc}) + u$$

- How might we allow the elasticity to be nonconstant, but include the above as a special case? The following does *not* work:

$$\log(\text{cons}) = \beta_0 + \beta_1 \log(\text{inc}) + \beta_2 \log(\text{inc}^2) + u$$

because  $\log(\text{inc}^2) = 2 \log(\text{inc})$ , that is,  $x_2 = 2x_1$ , where  $x_1 = \log(\text{inc})$ .

- Instead, we probably mean something like

$$\log(\text{cons}) = \beta_0 + \beta_1 \log(\text{inc}) + \beta_2 [\log(\text{inc})]^2 + u$$

which means  $x_2 = x_1^2$ . With this choice,  $x_2$  is an exact *nonlinear* function of  $x_1$ , but this (fortunately) is allowed in MLR.3.

- Tracking down perfect collinearity can be harder when it involves more than two variables.

- Motivation for Multiple Regression
- The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
- Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
- Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

## Example: Vote

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u$$

where  $expendA$  is campaign spending by candidate A,  $expendB$  is spending by candidate B, and  $totexpend$  is total spending. All are in thousands of dollars. Mechanically, the problem is that, by definition,

$$expendA + expendB = totexpend$$

which, of course, will also be true for any sample we collect.

# The 4 Assumptions for Unbiasedness

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

- One of the three variables has to be dropped.
- The model makes no sense from a ceteris paribus perspective. For example,  $\beta_1$  is suppose to measure the effect of changing *expendA* on *voteA*, holding fixed *expendB* and *totexpend*. But if *expendB* and *totexpend* are held fixed, *expendA* cannot change!
- We would probably drop *totexpend* and just use the two separate spending variables.

## Key Point

Assumption MLR.3 does *not* say the explanatory variables have to be uncorrelated in the population or sample.

Nor does it say they cannot be “highly” correlated.

**MLR.3** rules out *perfect correlation* in the sample, that is, correlations of  $\pm 1$ .

- Multiple regression would be useless if we had to insist  $x_1, \dots, x_k$  were uncorrelated in the sample (or population)!
- If the  $x_j$  were all pairwise uncorrelated, we could just use a bunch of simple regressions.

# The 4 Assumptions for Unbiasedness

- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

**MLR.1:**  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + u$   
**MLR.2:** random sampling from the population  
**MLR.3:** no perfect collinearity in the sample

- The last assumption ensures that the OLS estimators are unique and can be obtained from the first order conditions (minizing the sum of squared residuals).
- We need a final assumption for unbiasedness.

## Assumption MLR.4 (Zero Conditional Mean)

$$E(u|x_1, x_2, \dots, x_k) = 0 \text{ for all } (x_1, \dots, x_k)$$

- Remember, the real assumption is  $E(u|x_1, x_2, \dots, x_k) = E(u)$ : the average value of the error does not change across different slices of the population defined by  $x_1, \dots, x_k$ .
- Setting  $E(u) = 0$  essentially defines  $\beta_0$ .

If  $u$  is correlated with any of the  $x_j$ , **MLR.4** is violated.

- When Assumption MLR.4 holds, we say  $x_1, \dots, x_k$  are **exogenous explanatory variables**.
- If  $x_j$  is correlated with  $u$ , we often say  $x_j$  is an **endogenous explanatory variable**.

## Theorem: Unbiasedness of OLS

Under Assumptions MLR.1 through MLR.4,

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

for any values of the population parameters  $\beta_j$ . In other words, the OLS estimators are unbiased estimators of the population parameters.

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

- ① Motivation for Multiple Regression  
The Model with  $k$  Independent Variables
- ② Mechanics and Interpretation of OLS  
Interpreting the OLS Regression Line
- ③ The Expected Value of the OLS Estimators
- ④ The Variance of the OLS Estimators  
Estimating the Error Variance
- ⑤ Efficiency of OLS: The Gauss-Markov Theorem

- So far, we have assumed

**MLR.1:**  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$

**MLR.2:** random sampling from the population

**MLR.3:** no perfect collinearity in the sample

**MLR.4:**  $E(u|x_1, x_2, \dots, x_k) = 0$

- Under MLR.3 we can compute the OLS estimates in our sample.
- The other assumptions then ensure that OLS is unbiased (conditional on the outcomes of the explanatory variables).

- Now, our goal is to find  $Var(\hat{\beta}_j)$ .
- In order to do that we need to add another assumption: **homoskedasticity (constant variance)**.
- Why should we add another assumption?
  - ① Imposing this assumption, the OLS estimator has an important feature/property: **efficiency**.
  - ② We can obtain simple formulas with it too.

## Assumption MLR.5 (Homoskedasticity)

The variance of the error,  $u$ , does not change with any of  $x_1, x_2, \dots, x_k$ :

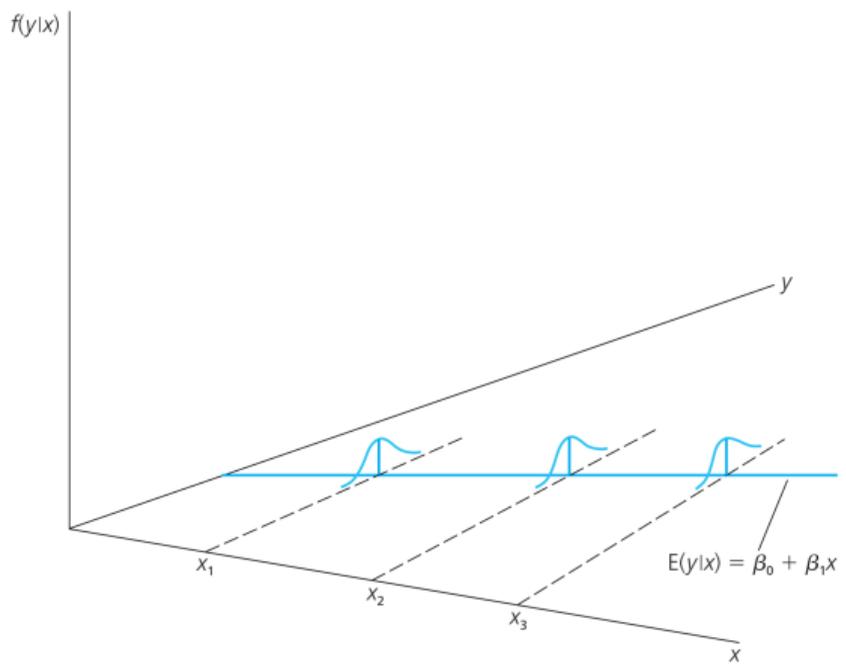
$$\text{Var}(u|x_1, x_2, \dots, x_k) = \text{Var}(u) = \sigma^2$$

- What it is saying is that the variance of the unobservable,  $u$ , conditional on  $x_1, x_2, \dots, x_k$  is constant.

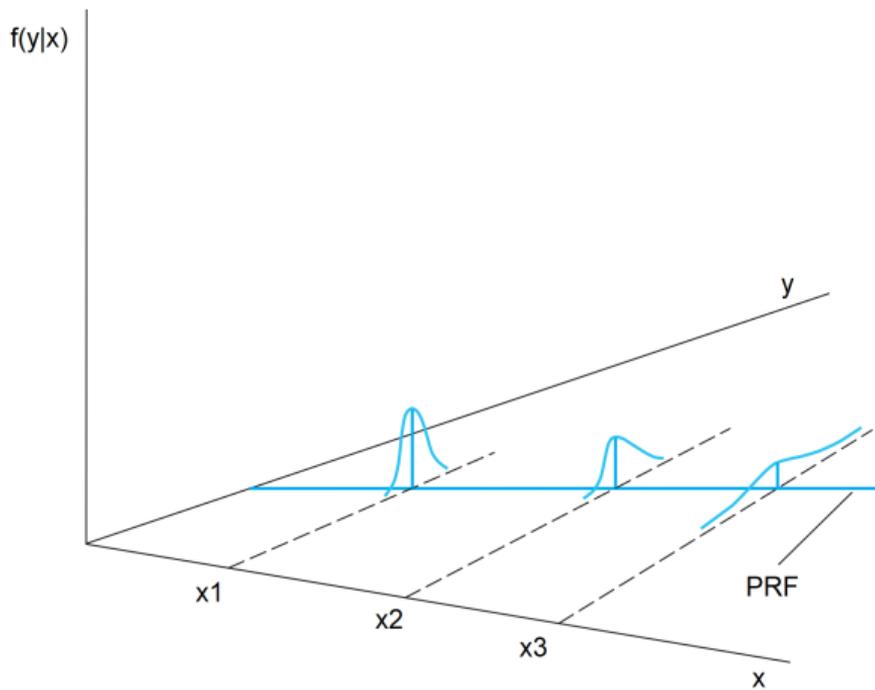
- The homoskedasticity assumption is common in cross-section analysis. However there are many problems where it does not hold.
- For time series **hardly (!)** you can make this assumption.
- When  $Var(u|x_1, x_2, \dots, x_k)$  depends on  $x_j$ , the error term exhibits **heteroskedasticity** (*nonconstant variance*)
- Since  $Var(u|x_1, x_2, \dots, x_k) = Var(y|x_1, x_2, \dots, x_k)$ , we have **heteroskedasticity** when  $Var(y|x_1, x_2, \dots, x_k)$  is a function of  $x$ .

- The homoskedasticity assumption plays no role in showing that  $\hat{\beta}_j$  are unbiased.
- $\sigma^2$  is the unconditional variance of  $u$ .
- $\sigma^2$  : **error variance** or **disturbance variance**.
- $\sqrt{\sigma^2} = \sigma$  : **standard deviation of the error**.

- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem



- Motivation for Multiple Regression
- The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
- Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
- Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem



- Assumptions **MLR.1** and **MLR.4** imply

$$E(y|x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

and when we add **MLR.5**,

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

- Assumptions **MLR.1** through **MLR.5** are called the **Gauss Markov assumptions**.

Motivation for  
Multiple  
Regression

The Model with  $k$   
Independent  
Variables

Mechanics and  
Interpretation  
of OLS

Interpreting the OLS  
Regression Line

The Expected  
Value of the  
OLS  
Estimators

The Variance  
of the OLS  
Estimators

Estimating the Error  
Variance

Efficiency of  
OLS: The  
Gauss-Markov  
Theorem

## Gauss Markov assumptions

**MLR.1:**  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + u$

**MLR.2:** random sampling from the population

**MLR.3:** no perfect collinearity in the sample

**MLR.4:**  $E(u|x_1, x_2, \dots, x_k) = 0$

**MLR.5:**  $Var(u|x_1, x_2, \dots, x_k) = Var(u) = \sigma^2$

Recall, our goal is to find  $Var(\hat{\beta}_j)$   
 (We will not find  $Var(\hat{\beta}_0)$  - which has different formula)

- Let's define the total variation in  $x_j$  in the sample:

$$SST_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

Notice that the  $R$ -squared can also be understood as the squared correlation between  $x_j$  and the other explanatory variables.

- Let's define  **$R$ -squared**  $R_j^2$ :  
a measure of correlation between  $x_j$  and the other explanatory variables (in the sample) is the  $R$ -squared from the regression:

$$x_{ij} \text{ on } x_{i1}, x_{i2}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{ik}$$

We are regressing  $x_j$  on all of the *other* explanatory variables.

## Theorem: Sampling Variances of OLS Slope Estimators

Under Assumptions **MLR.1** to **MLR.5**, and condition on the values of the explanatory variables in the sample,

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

- Clearly, all five Gauss-Markov assumptions are needed to ensure this formula is correct.

- If,

$$\text{Var}(u|x_1, x_2, \dots, x_k) = f(x_j)$$

- **Example:** On the white board.
- This violates **MLR.5**, and the standard variance formula is *generally* incorrect for **all** OLS estimators, not just  $\text{Var}(\hat{\beta}_j)$ .

- Is  $R_j^2 = 1$  allowed? Answer: **No**.
- Any value  $0 \leq R_j^2 < 1$  is permitted.
- **Multicollinearity** As  $R_j^2$  gets closer to one,  $x_j$  is more linearly related to the other independent variables.

- The variance

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

has three components:

- $\sigma^2$
- $SST_j$
- $1 - R_j^2$

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

**Factors Affecting  $\text{Var}(\hat{\beta}_j)$ :**

**(1)** If the error variance  $\sigma^2 \downarrow$ ,  
 $\Rightarrow \text{Var}(\hat{\beta}_j) \downarrow \quad \Rightarrow \text{Var}(u|\mathbf{X}) \downarrow$  adding more explanatory variables

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

**Factors Affecting  $\text{Var}(\hat{\beta}_j)$ :**

**(2)** If the  $SST_j \uparrow$ ,

$\text{Var}(\hat{\beta}_j) \downarrow \Rightarrow$  the higher is the sample variation in  $x_j$  the better (increase the sample size  $n$ :  $SST_j$  is roughly a linear function of  $n$ ).

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

**Factors Affecting  $\text{Var}(\hat{\beta}_j)$ :**

**(3)** As  $R_j^2 \rightarrow 1$ ,

$\text{Var}(\hat{\beta}_j) \rightarrow \infty \Rightarrow R_j^2$  measures how linearly related  $x_j$  is to the other explanatory variables.

- We get the smallest variance for  $\hat{\beta}_j$  when  $R_j^2 = 0$ :

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j},$$

- If  $x_j$  is unrelated to all other independent variables  $\Rightarrow$  easier to estimate its ceteris paribus effect on  $y$ .
- $R_j^2 \approx 0$  (uncommon).
- $R_j^2 \approx 1$  (more common)  $\Rightarrow$  the estimate of  $\beta_j$  is not precise.

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

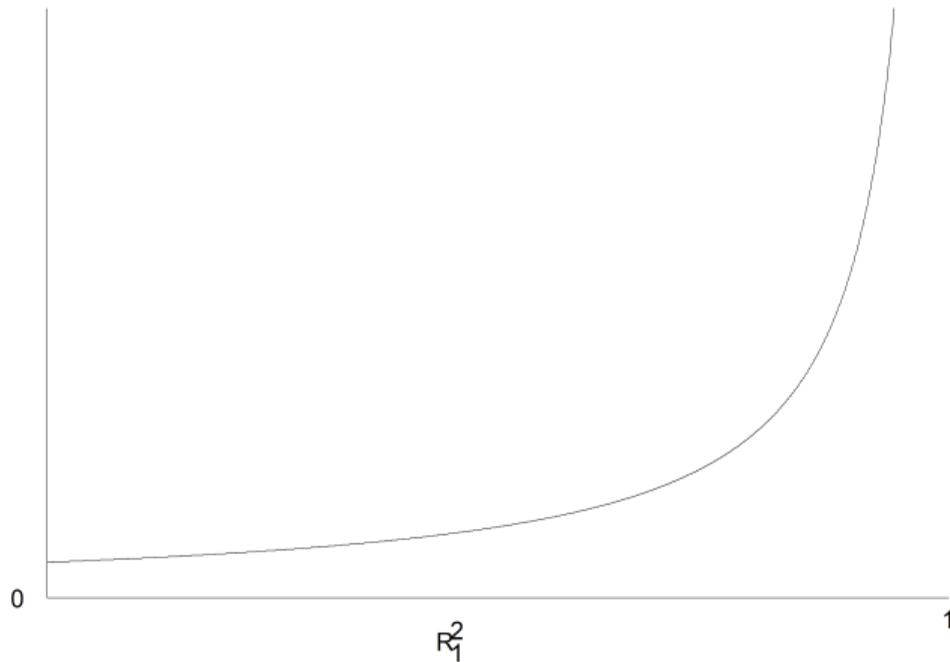
The Expected Value of the OLS Estimators

**The Variance of the OLS Estimators**

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

Figure: Graph of  $Var(\hat{\beta}_1)$  as a function of  $R_1^2$



Recall,

**Multicollinearity:**  $R_j^2$  close to one. (problem of ...)

**Perfect Collinearity:**  $R_j^2 = 1$  (not allowed under **MLR.1 - MLR.4**)

- Does multicollinearity (high correlation among two or more independent variables) violates any of the Gauss-Markov assumptions (including MLR.3.)?

Answer: **No.** Multicollinearity does not cause the OLS estimators to be biased. We still have  $E(\hat{\beta}_j) = \beta_j$ .

**Goal:** We need to estimate  $\sigma^2$ .

- **Problem:** we don't observe  $u_i$ .
- We could use our residuals  $\hat{u}_i$  (that we obtain when we run a regression) to find  $\sigma^2$ .

- Motivation for Multiple Regression
  - The Model with  $k$  Independent Variables
- Mechanics and Interpretation of OLS
  - Interpreting the OLS Regression Line
- The Expected Value of the OLS Estimators
- The Variance of the OLS Estimators
  - Estimating the Error Variance
- Efficiency of OLS: The Gauss-Markov Theorem

- **Degrees of freedom:** With  $n$  observations and  $k + 1$  parameters, we only have

$$df = n - (k + 1)$$

degrees of freedom. Recall we lose the  $k + 1$   $df$  due to  $k + 1$  restrictions on the OLS residuals:

$$\sum_{i=1}^n \hat{u}_i = 0$$

$$\sum_{i=1}^n x_{ij} \hat{u}_i = 0, \quad j = 1, 2, \dots, k$$

## Estimator of $\sigma^2$

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

- Regression packages (e.g. **R**) reports:
  - $\sqrt{\hat{\sigma}^2} = \hat{\sigma}$
  - **Names:** *Residual std. error, std. error of the regression, root mean squared error, standard error of the estimate, root mean squared error*

Note that  $SSR$  falls when a new explanatory variable is added, but  $df$  falls, too. So  $\hat{\sigma}$  can increase or decrease when a new variable is added in multiple regression.

## Theorem: Unbiased Estimation of $\sigma^2$

Under the Gauss-Markov assumptions **MLR.1** through **MLR.5**

$$E(\hat{\sigma}^2) = \sigma^2$$

i.e.,  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ .

**Goal:** Now we want to find the **standard error** of each  $\hat{\beta}_j$ .

**Standard deviation of  $\hat{\beta}_j$**

$$sd(\hat{\beta}_j) = \frac{\sigma}{\sqrt{SST_j(1 - R_j^2)}}$$

**Standard error of  $\hat{\beta}_j$**

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - R_j^2)}}$$

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

	<i>Dependent variable:</i>
	lwage
educ	0.092*** (0.007)
exper	0.004** (0.002)
tenure	0.022*** (0.003)
Constant	0.284*** (0.104)
Observations	526
R <sup>2</sup>	0.316
Adjusted R <sup>2</sup>	0.312
Residual Std. Error	0.441 (df = 522)
F Statistic	80.391*** (df = 3; 522)

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Motivation for  
Multiple  
Regression

The Model with  $k$   
Independent  
Variables

Mechanics and  
Interpretation  
of OLS

Interpreting the OLS  
Regression Line

The Expected  
Value of the  
OLS  
Estimators

The Variance  
of the OLS  
Estimators

Estimating the Error  
Variance

Efficiency of  
OLS: The  
Gauss-Markov  
Theorem

- ① Motivation for Multiple Regression  
The Model with  $k$  Independent Variables
- ② Mechanics and Interpretation of OLS  
Interpreting the OLS Regression Line
- ③ The Expected Value of the OLS Estimators
- ④ The Variance of the OLS Estimators  
Estimating the Error Variance
- ⑤ Efficiency of OLS: The Gauss-Markov Theorem

## Theorem: Gauss-Markov

Under Assumptions MLR.1 through MLR.5 (Gauss-Markov assumptions), the OLS estimators  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  are the **best linear unbiased estimators (BLUEs)**

- To understand each component of the acronym “BLUE” let’s start from the end.

Motivation for Multiple Regression

The Model with  $k$  Independent Variables

Mechanics and Interpretation of OLS

Interpreting the OLS Regression Line

The Expected Value of the OLS Estimators

The Variance of the OLS Estimators

Estimating the Error Variance

Efficiency of OLS: The Gauss-Markov Theorem

**E (estimator):** It is a rule that can be applied to any sample of data to produce an estimate.

**U (unbiased):**  $\hat{\beta}_j^{OLS}$  is an unbiased estimator of the true parameter, i.e.,  $\beta_j$ .

$$\Rightarrow E(\hat{\beta}_j^{OLS}) = \beta_j \text{ for any } \beta_0, \beta_1, \beta_2, \dots, \beta_k$$

(conditional on  $\{(x_{i1}, \dots, x_{ik}) : i = 1, \dots, n\}$ ).

**L (linear):** The estimator is a linear function of  $\{y_i : i = 1, 2, \dots, n\}$ , but *it can be a nonlinear function of the explanatory variables.*, i.e.,

$$\tilde{\beta}_j = \sum_{i=1}^n w_{ij} y_i$$

where the  $\{w_{ij} : i = 1, \dots, n\}$  are any functions of  $\{(x_{i1}, \dots, x_{ik}) : i = 1, \dots, n\}$ .

- The OLS estimators can be written in this way.

**B (best):** This means smallest variance (which makes sense once we impose unbiasedness).

$$\text{Var}(\hat{\beta}_j) \leq \text{Var}(\tilde{\beta}_j) \text{ all } j$$

usually the inequality is strict. (conditional on the explanatory variables in the sample).

- If we do not impose unbiasedness, then we can use silly rules – such as  $\tilde{\beta}_j = 1$  always – to get estimators with zero variance.

- If the Gauss-Markov assumptions hold, and we insist on unbiased estimators that are also linear functions of  $\{y_i : i = 1, 2, \dots, n\}$ , then

OLS delivers the smallest possible variances.

- We are not looking nonlinear functions of  $\{y_i : i = 1, 2, \dots, n\}$ .

● **Remember:** Failure of MLR.5 does not cause bias in the  $\hat{\beta}_j$ , but it does have two consequences:

1. The usual formulas for  $Var(\hat{\beta}_j)$ , and therefore for  $se(\hat{\beta}_j)$ , are wrong.
2. The  $\hat{\beta}_j$  are no longer BLUE.