Econometrics and Regression

- As mentioned earlier, econometrics uses economics principles to build an empirical model.

- Regression analysis presumes that the *dependent variable y* is influenced by the *explanatory variables* $x_1, x_2, \cdots, x_k$. Consequently, we use information on the explanatory variables to predict and/or describe changes in the dependent variable. Alternative names for the explanatory variables are predictor variables, independent variables, or the input variables while the dependent variable is often referred to as the response variable, target variable, or the output variable.

- To develop a regression model, we also include the error term, ε, that captures the stochastic nature of the relationship (non-deterministic). It is impossible to include all *x*'s that can explain the variations in *y*. In analyzing salary, variables such as motivation are not even measurable (quantifiable).

- Tests of significance tests help determine which explanatory variables matter (are statistically significant) and which don't.

- For credible estimates and the significance tests, we resort to making certain assumptions. Later, we will examine the importance of these assumptions on the statistical properties of the estimator, as well as the validity of the testing procedures. We address common violations to the model assumptions, the consequences when these assumptions are violated, and offer some remedial measures.

Econometrics and Regression

There are two important perspectives for econometrics:

- Predictive modeling perspective that captures conditional expectation, $E(y|x_1, x_2, x_3, \dots)$
- Causal estimation perspective that captures the partial effect, $\partial y/\partial x_j$

Both of these perspectives are tremendously important.

- If you want to make a prediction the former is relevant
- If you want to make a decision, the latter is relevant
- More later!

The Simple Linear Regression Model

- A simple linear regression model uses only one explanatory variable.

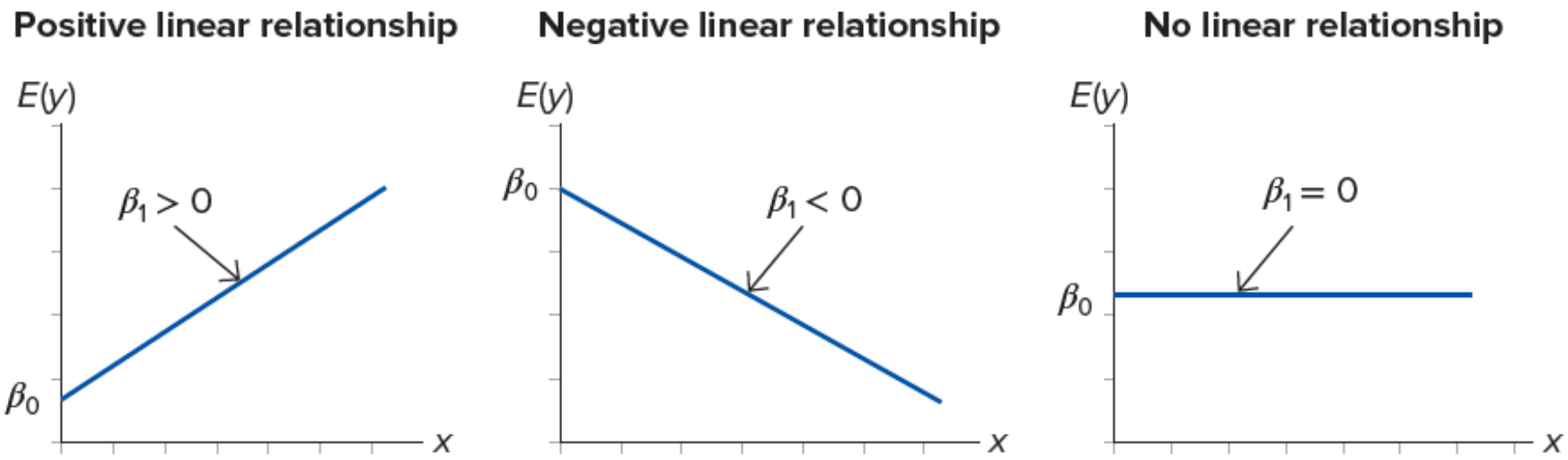
- Consider $y = \beta_0 + \beta_1 x + \varepsilon$

  - β_0 is the unknown intercept and β_1 the unknown slope
  - $\beta_0 + \beta_1 x$ is the deterministic component
  - ε is the stochastic component or random error term

- Conditional on x, $E(\varepsilon) = 0$. Therefore, $E(y) = \beta_0 + \beta_1 x$.

Positive, Negative, or no Relationship

The slope coefficient β_1 of the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, determines the direction of the relationship.



The Multiple Regression Model

- Multiple Linear Regression
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$
  - $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are unknown parameters

- Sample data: n observations of $y, x_1, x_2, \ldots, x_k$

- Use the sample data to obtain $b_0, b_1, b_2, \ldots b_k$ which are estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$.

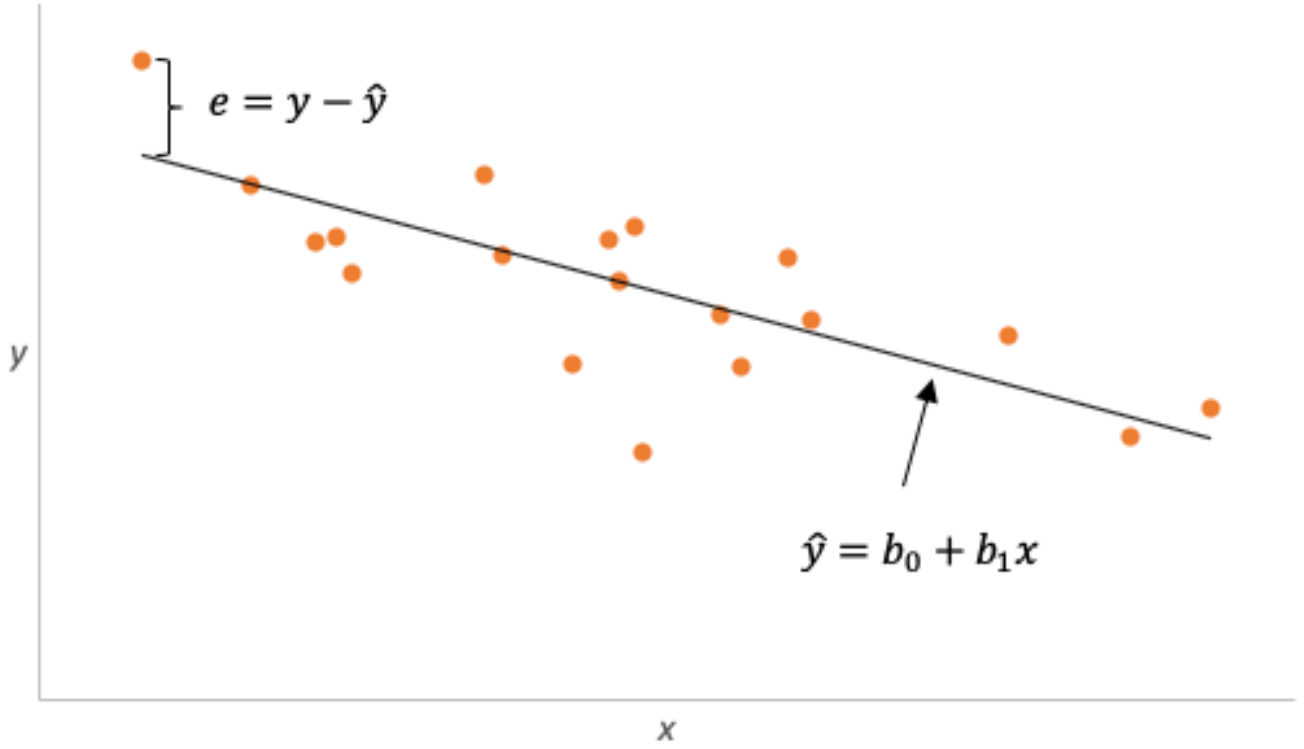
The Estimated Regression

- $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$
  - $b_0, b_1, b_2, \ldots, b_k$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$
  - $\hat{y}$ is the predicted value.

- A residual is defined as $e = y - \hat{y}$.

- Ordinary least squares (OLS) chooses the estimated regression equation by minimizing the error (residual) sum of squares, $SSE = \sum (y - \hat{y})^2 = \sum e^2$.
  - Desirable properties if certain assumptions hold (more later)
  - Gives an equation "closest" to the data

Scatterplot for a Simple Regression Model



As noted earlier, $\hat{y} = b_0 + b_1 x$ is obtained by minimizing $\text{SSE} = \sum e^2$.

Interpreting the Regression Coefficients

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

- b_0 is the estimate of β_0
  - Predicted value of $\hat{y}$ when each explanatory variable assumes a value of 0
  - Not always meaningful

- b_j is the estimate of β_j
  - Change in the predicted value of the dependent variable given a unit increase in x_j, holding all other explanatory variables constant
  - Partial influence of x_j on $\hat{y}$

Dummy Variables

- Explanatory variables can be numerical or categorical
- Examples of categorical variables include email (legitimate or spam) and loan default (yes or no). We will discuss multiple categories later.
- Categorical variable cannot be used in their original form—that is, in a non-numerical format. We convert a categorical variable into a dummy (indicator) variable.
- A dummy variable *d* assumes a value of 1 for one of the categories and 0 for the other.
- Assuming the absence of nonbinary cases, when categorizing a person's sex, we can define d as 1 for male and 0 for female. Alternatively, we can define d as 1 for female and 0 for male, with no change in inference.

Case Study: College Scorecard

- What explains variation in post-college earnings?
- Dependent Variable
  - Annual post-college earnings (Earnings in $)
- Explanatory Variables
  - The average annual cost (Cost in $)
  - The graduation rate (Grad in %)
  - The percentage of students paying down debt (Debt in %)
  - The location of the school (City or Noncity)
- Information on 116 schools (Data: ***College***)

A	B	C	D	E	F
School	**Earnings**	**Cost**	**Grad**	**Debt**	**Location**
St. Ambrose C (NC)	44800	22920	62	88	City
Albion College (Albion, MI)	45100	23429	73	92	Non-city
Alfred University (Alfred, NY)	42300	19567	63	87	Non-city
Allegheny College (Meadville, PA)	49200	25147	78	92	Non-city
Beloit College (Beloit, WI)	37900	21979	78	93	City
Bentley University (Waltham, MA)	74900	29886	86	98	City
Boston University (Boston, MA)	60600	34603	84	95	City

Dummy Variables

As discussed, a dummy variable d is used to describe two categories of a categorical variable.

- $d = 1$ for one of the categories
- $d = 0$ for the other(s)
  - ✓ Reference or benchmark category
  - ✓ All comparisons are made relative to this category

- In above case: we define City as 1 if city, 0 otherwise

- Treat dummy variables like any other variable for estimation and testing.

Case Study Continued

$$Earnings = \beta_0 + \beta_1 Cost + \beta_2 Grad + \beta_3 Debt + \beta_4 City + \varepsilon$$

a. What is the estimated regression equation?

b. Interpret the slope coefficients.

c. Predict the annual post-college earnings if a college's average annual cost is $25,000, its graduation rate is 60%, its percentage of students paying down debt is 80%, and it is located in a city.

Case Study Using Excel

Microsoft Excel is arguably the most widely used computer application among business professionals. Accountants, economists, financial analysts, marketers, HR managers, and many others use Excel spreadsheets for everyday business tasks. Oftentimes, these tasks involve entering, editing, and formatting data as well as performing data analysis.

Excel's Analysis ToolPak Add-In

Excel offers a number of add-ins that come preinstalled. The Analysis ToolPak add-in is used for statistical analysis. The following instructions activate the Analysis Toolpak add-in.

Activating the Analysis ToolPak Add-In

a. For Microsoft Windows, in Excel, go to **File > Options > Add-Ins**. For macOS, go to **Tools > Excel Add-ins** and continue to step C.

b. In the Manage Excel Add-ins section (toward the bottom of the screen), click **Go**. . . . .

c. On the *Add-ins* dialog box, check the Analysis ToolPak box and click **OK**.

d. In Excel, go to the **Data** tab and verify that the Data Analysis command button appears in the Analyze (or Analysis) group.

Case Study Using Excel

a. Make sure to add-in the Analysis ToolPak for statistical analysis.

b. Open the *College* data file. Choose Data > Data Analysis > Regression from the menu.

c. Insert a blank column between Debt and Location so that we can place the dummy variable together with other predictor variables. Enter the column heading City in cell F1. In cell F2, enter the formula =IF(G2="City", 1, 0). Fill the range F3:F117 with the formula in F2.

d. In the *Regression* dialog box, click on the box next to *Input Y Range*, and then select the data for Earnings. Click on the box next to *Input X Range*, and then *simultaneously* select the data for Cost, Grad, Debt, and City. Select *Labels*. Click OK.

Case Study Using Excel

Regression Statistics						
Multiple R	0.6552					
R Square	0.4292					
Adjusted R Square	0.4087					
Standard Error	5645.831					
Observations	116					
ANOVA						
	df	**SS**	**MS**	**F**	**Significance F**	
Regression	4	2660691959	665172990	20.868	7.56E-13	
Residual	111	3538169765	31875403			
Total	115	6198861724				
	Coefficients	**Standard Error**	**t Stat**	**p-value**	**Lower 95%**	**Upper 95%**
Intercept	**10004.9665**	7634.3338	1.311	0.1927	−5122.98	25132.91
Cost	**0.4349**	0.1110	3.917	0.0002	0.21	0.65
Grad	**178.0989**	69.1940	2.574	0.0114	40.99	315.21
Debt	**141.4783**	117.2120	1.207	0.2300	−90.79	373.74
City	**2526.7888**	1103.4026	2.290	0.0239	340.32	4713.25

Case Study Using Excel

a. *What is the estimated regression equation?*

$$\widehat{\text{Earnings}} = 10004.97 + 0.4349\text{Cost} + 178.10\text{Grad} + 141.48\text{Debt} + +2526.79\text{City}$$

b. *Interpret the slope coefficients.* All coefficients are positive, suggesting a positive influence of each predictor variable on the response variable.

- Holding all other predictor variables constant, if the average annual costs increase by $1, then, on average, predicted earnings are expected to increase by $0.4349.
- All else constant, predicted earnings are $2,526.79 higher for graduates of colleges located in a city.

c. *Predict the annual post-college earnings if a college's average annual cost is $25,000, its graduation rate is 60%, its percentage of students paying down debt is 80%, and it is located in a city.*

$$\widehat{\text{Earnings}} = 10{,}004.97 + 0.434 \times 25{,}000 + 178.10 \times 60 + 141.48 \times 80 +$$
$$+ 2{,}526.79 \times 1 = 45{,}408.7991 \text{ or about } \$45{,}409$$

Case Study Using R

- Installation of both R and RStudio is straightforward and requires no special modifications to your system. Refer to Getting Started with R in Lecture Notes.

- RStudio does not come with R; *therefore, both pieces of software need to be installed separately.*

- R is case sensitive.

- Due to different fonts and type settings, copying and pasting R functions from the notes may cause errors. Try replacing special characters such as quotation marks and parentheses or delete extra spaces in the functions.

- An easy way to import data file: select **File > Import Dataset > From Excel.** You can also import a comma- or tab-delimited text files.

Case Study Using R

- Import the ***College*** data file into a data frame (table) and label it myData.
- By default, R will report the regression output using scientific notation. We can turn this option off by entering options(scipen=999) at the prompt. To turn scientific notation back on, we enter options(scipen=0).
- We can find the mean of, say cost, with the command in the console pane: mean(myData$Cost) and get 25251.69
- Or find the mean of all variables: colMeans(myData[,c(2:5)]). Note that Column 6 is non-numeric.
- Or, for more measures: summary(myData[,c(2:5)])
- You can access any previous command by pressing the up-arrow key, etc.

Case Study Using R

Question: Estimate the model and interpret the slope coefficients. Predict the annual post-college earnings if a college's average annual cost is $25,000, its graduation rate is 60%, its percentage of students paying down debt is 80%, and it is located in a city.

- # Import College data
- options(scipen=999)
- myData$City <- ifelse(myData$Location=="City",1,0)
- Model <- lm(Earnings ~ Cost + Grad + Debt + City, data = myData)
- summary(Model)

```
Call:
lm(formula = Earnings ~ Cost + Grad + Debt + City, data = myData)

Residuals:
     Min       1Q   Median       3Q      Max
-12375.3  -3065.2   -589.9   2946.5  20189.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10004.9665  7634.3338   1.311 0.192724
Cost            0.4349     0.1110   3.917 0.000155 ***
Grad          178.0989    69.1940   2.574 0.011373 *
Debt          141.4783   117.2120   1.207 0.229987
City         2526.7888  1103.4026   2.290 0.023912 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5646 on 111 degrees of freedom
Multiple R-squared:  0.4292,  Adjusted R-squared:  0.4087
F-statistic: 20.87 on 4 and 111 DF,  p-value: 0.0000000000007564
```

- predict(Model, data.frame(Cost=25000,Grad=60,Debt=80,City=1))

Practice Problem

The accompanying data file shows information that he has collected on salary (Salary in $ millions), pass completion rate (PC in %), total touchdowns scored (TD), and age for 32 quarterbacks during a recent season. (Data: **Quarterbacks**)

a. Estimate and interpret: Salary $= \beta_0 + \beta_1 PC + \beta_2 TD + \beta_3 Age + \varepsilon.$

b. Player 8 earned 12.9895 million dollars during the season. According to the model, what is his predicted salary if PC $= 70.6$, TD $= 34$, and Age $= 30$?

c. Player 16 earned 8.0073 million dollars during the season. According to the model, what is his predicted salary if PC $= 65.5$, TD $= 28$, and Age $= 32$?

d. Compute and interpret the residual salary for Player 8 and Player 16.