Econometrics

- Econometrics uses economics principles to build an empirical model. You can think of an empirical model as an observation or a proposition that one variable, labeled y is related to one or more variables labeled $x_1, x_2, \cdots, x_k$.

- Regression analysis is one of the most widely used statistical methodologies in economics, business, engineering, and the social sciences. It 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 vari*ables while the dependent variable is often referred to as the *response variable*, *target variable*, or the *output variable.*

Data Wrangling and Econometrics

- Before developing and applying econometric models, it's crucial to prepare the data.
- Data wrangling, also known as data munging or data cleaning, refers to the process of transforming and mapping raw data into a format suitable for analysis.
- In this lecture, we will a few perform important data wrangling tasks in Excel. Excel offers a user-friendly interface, making it accessible to individuals with varying levels of technical expertise.
- While Excel is not be as powerful as specialized programming languages like R or Python for large-scale data wrangling tasks, it can still be a useful tool for smaller datasets and basic data preparation.

Data Wrangling using Excel

Relative, Absolute, and Mixed References

By default, cell addresses in a formula, such as = B1+C1, are relative references and will change when a formula is copied to another cell. For example, if we enter the formula =B1+C1, in cell A1 and copy it to cell D4, the formula in cell D4 will appear as =E4+F4.

Absolute references allow us to maintain the original cell references when a formula is copied elsewhere. We specify absolute references by adding a dollar sign ($) in front of the column name and row number (e.g., $B$1).

We use mixed references by adding a dollar sign ($) in front of either the column name or the row number (e.g., $B1 or B$1), but not both. This will keep the reference to the specific column name or the row number constant.

Reference	Formula in cell A1	Formula in cell A1 copied to cell D4
Relative	=B1+C1	=E4+F4
Absolute	=$B$1+$C$1	=$B$1+$C$1
Mixed	=$B1+$C1	=$B4+$C4
Mixed	=B$1+C$1	=E$1+F$1

Data Wrangling using Excel

Example: Refer to the ***Tax*** worksheet. In Examples 1 and 2, insert a formula for the tax amount for January that can be copied and pasted to other months.

Example 1

		Tax Rate
		15%
Month	Earnings	Tax
January	$70,000	
February	$68,000	
March	$72,000	
April	$74,000	
May	$73,000	
June	$84,000	
July	$96,000	
August	$90,000	
September	$76,000	
October	$64,000	
November	$84,000	
December	$106,000	
Total		

Example 2

		Tax Rate		
		15%	20%	25%
Month	Earnings	Tax 1	Tax 2	Tax 3
January	$70,000			
February	$68,000			
March	$72,000			
April	$74,000			
May	$73,000			
June	$84,000			
July	$96,000			
August	$90,000			
September	$76,000			
October	$64,000			
November	$84,000			
December	$106,000			
Total				

Data Wrangling using Excel

In Excel, we use formulas to perform basic calculations. When we enter a formula in a cell, Excel performs the specified calculation and returns the result in the same cell.

Functions in Excel are predefined formulas. Like a formula, a function always begins with an equal sign (=) and must be written with the correct syntax enclosed within parentheses. The following are some important predefined formulas.

Function and syntax	Description	Example
=COUNT(array)	Returns the number of cells in the array with numerical values.	=COUNT(A1:A10)
=COUNTA(array)	Returns the number of cells in the array that are not blank.	=COUNTA(A1:A10)
=COUNTBLANK(array)	Returns the number of cells in the array that are blank.	=COUNTBLANK(A1:A10)
=COUNTIF(array, criteria)	Returns the number of cells in the array that meet a specific selection criteria.	=COUNTIF(A1:A10, “>10”)
=IF(logical statement, result if the statement is true, result if the statement is false)	Returns a result based on the outcome of the logical statement.	=IF(A1=“Yes”, 1, 0). If A1 = “Yes”, returns a 1. If not, returns a 0.
=SUM(array)	Adds and returns the sum of the numbers in the array.	=SUM(A1:A10)
=VLOOKUP(lookup value, reference table, column number in the reference table containing results)	Searches and retrieves information from a specified column in a reference table.	See the example below.

Data Wrangling using Excel

Example: Refer to the ***Grades*** worksheet. This example uses the VLOOKUP feature of Excel.

Consider the following averages in ECON 339 using **Grades** data.

1. Convert averages to letter grades given the following scale:
2. Find the average for Students 1, 6, and 12

Student	Average
1	75.75
2	77.77
3	90.99
4	82.38
5	90.94
6	83.26
7	67.83
8	64.95
9	72.06
10	84.97

92% and above	A	72% - 76%	C+
88% - 92%	A-	68% - 72%	C
84% - 88%	B+	64% - 68%	C-
80% - 84%	B	60% - 64%	D
76% - 80%	B-	< 60%	F

Data Wrangling using Excel

Example: BalanceGig is a company that matches independent workers for short term engagements with businesses in the construction, automotive, and high-tech industries. The 'gig' employees work only for a short period of time, often on a particular project or a specific task. A manager at BalanceGig extracts the employee data from their most recent work engagement, including the hourly wage, the client's industry, and the employee's job classification. A portion of the ***Gig*** data set is shown in the following table.

EmployeeID	HourlyWage	Industry	Job
1	32.81	Construction	Analyst
2	46	Automotive	Engineer
⋮	⋮	⋮	⋮
604	26.09	Construction	Other

1. Find the number of missing observations for the HourlyWage, Industry, and Job variables.
2. Find the number of employees who
    - Worked in the automotive industry
    - Earned more than $30 per hour
    - Both
3. Find the lowest- and highest-paid accountants who worked in the automotive and the tech industries.

Data Wrangling using Excel

Summary

- There is a total of 604 records in the data set. There are no missing values in the hourly wage. The Industry and Job variables have 10 and 16 missing values, respectively.

- 190 employees worked in the automotive industry, 536 employees earned more than $30 per hour, and 181 employees worked in the automotive industry and earned more than $30 per hour.

- The lowest and the highest hourly wages in the data set are $24.28 and $51.00, respectively. The three employees who had the lowest hourly wage of $24.28 all worked in the Construction industry and were hired as Engineer, Sales Rep, and Accountant, respectively. Interestingly, the employee with the highest hourly wage of $51.00 also worked in the Construction industry in a job type classified as Other.

- The lowest and the highest paid accountants who worked in the automotive industry made $28.74 and $49.32 per hour, respectively. In the Tech industry, the lowest and the highest paid accountant made $36.13 and $49.49 per hour, respectively. Note that the lowest hourly wage for an accountant is considerably higher in the Tech industry compared to the automotive industry ($36.13 > $28.74).

Data Wrangling using Excel

Dealing with Missing Values

There are two common strategies for dealing with missing values.

- The **omission** strategy recommends that observations with missing values be excluded from subsequent analysis.
- The **imputation** strategy recommends that the missing values be replaced with some reasonable imputed values.
- For numerical variables, it is common to replace the missing values with the average (typical) values across relevant observations. For categorical variables, it is common to impute the most predominant category.
- For small data sets, you can simply use Excel functions like sorting or filtering to identify missing information and using omission or imputation strategies. Other packages, like R and Python, allow built-in algorithms for performing such tasks.

Using ***Gig*** data, show that with the omission strategy, you are left with 578 records. Here 26 (10+16) records are removed; fewer than 26 would have been removed if both Industry and Job variables had missing information for the same employee.

Data Wrangling using Excel

Subsetting

The process of extracting portions of a data set that are relevant to the analysis is called subsetting. It is commonly used to pre-process the data prior to analysis.

Example: Catherine Hill is a marketing manager at Organic Food Superstore. She has been assigned to market the company's new line of Asian inspired meals. Research has shown that the most likely customers for healthy ethnic cuisines are *college-educated millennials* (born between 1982 and 2000). In order to spend the company's marketing dollars efficiently, Catherine wants to focus on this target demographic when designing the marketing campaign. With the help of the Information Technology (IT) group, Catherine has acquired a representative sample of Organic Food Superstore's customers.

- Subset the ***Customers*** data to first identify college-educated millennial customers and then compare the profiles of female and male college-educated millennial customers.
- Show that the *Female* worksheet contains 21 observations of female college-educated millennials, and the *Male* worksheet contains 38 observations of male college-educated millennials. Refer to the appropriate tables (next slide) for a portion of the data of the *Female* and *Male* subsets.

Data Wrangling using Excel

Sex	HouseholdSize	Income	Spending2021	NumOfOrders	Channel
Female	5	53000	241	3	SM
Female	3	84000	153	2	Web
⋮	⋮	⋮	⋮	⋮	⋮
Female	1	52000	586	13	Referral

Sex	HouseholdSize	Income	Spending2021	NumOfOrders	Channel
Male	5	94000	843	12	TV
Male	1	97000	1028	17	Web
⋮	⋮	⋮	⋮	⋮	⋮
Male	5	102000	926	10	SM

Summary Table

	HouseholdSize	Income	Spending2021	NumOfOrders	SM
Females	3.10	61285.71	646.14	9.48	0.14
Males	3.13	83157.89	642.21	11.08	0.42
	Average	Average	Average	Average	Proportion

- Female and male customers are similar in terms of household sizes and total spending in 2018.
- More high-income earners are found among the male customers than female customers.
- Male customers seem to order more than female customers do even though their total spending is about the same.
- A large portion of the male customers were acquired through social media ads whereas female customers were often acquired through Web ads and referral programs.

Data Wrangling using Excel

Transforming Categorical Variables

- We use labels or names to identify the distinguishing characteristics of a categorical variable. Examples include gender (male or non-male), marital status (single, married, widowed, divorced, separated) and the performance of a manager (excellent, good, fair, poor).

- Most quantitative techniques are limited in their abilities to handle categorical data directly. A common transformation of categorical variables is to convert them into numerical values.

- A **dummy variable**, also referred to as an indicator or a binary variable, is commonly used to describe two categories of a variable. It assumes a value of 1 for one of the categories and 0 for the other category, referred to as the reference or the benchmark category. For example, we can define a dummy variable to categorize a person's gender using 1 for male and 0 for non-male where non-males is the reference category. Alternatively, without any loss of generality, we can define 1 for non-male and 0 for male, using males as the reference category. All interpretation of the results is made in relation to the reference category.

- Oftentimes, a categorical variable is defined by more than two categories. For example, the mode of transportation used to commute may be described by three categories: Public Transportation, Driving Alone, and Car Pooling. Later we will learn that given k categories of a variable, the general rule is to use create $k - 1$ dummy variables in the (regression) analysis, using the last category as reference.

Data Wrangling using Excel

Example: For the new Asian-inspired meal kits, Catherine feels that understanding the channels through which customers were acquired is important to predict customers' future behaviors. In order to include Channel in her predictive model, Catherine needs to convert the channel categories into dummy variables. Use the Channel variable in the ***Customers*** data to create the relevant dummy variables. Find the number and percentage of customers for each channel.

Open the ***Customers*** data file.

Enter "Referral" in a blank column and enter the formula =IF(N2="Referral",1,0). Copy and paste this formula to the rest of the column. Repeat the analysis for other channels.

Channel	Referral	SM	TV	Web
Number	38	39	57	66
Percentage	19.00%	19.50%	28.50%	33.00%

Summary Measures using Excel

Summary Measures

- We use numerical descriptive measures to extract meaningful information from data. These measures provide precise, objectively determined values that are easy to calculate, interpret, and compare with one another.

- The *mean* is the most used measure of **central location**. A known weakness of the mean is that it is unduly influenced by outliers. The *median* is the middle observation of a variable; that is, it divides the variable in half. The median is especially useful when outliers are present. The *mode* is the most frequently occurring observation of a variable. A variable may have no mode or more than one mode. The mode is the only meaningful measure of central location for a categorical variable.

- The *variance* and the *standard deviation* are the most used measures of *dispersion*. The standard deviation is the positive square root of the variance.

Summary Measures using Excel

Example: Growth funds invest in companies whose stock prices are expected to grow at a faster rate, relative to the overall stock market, and value funds invest in companies whose stock prices are below their true worth. Consider the **Growth_Value** for the annual return data for Fidelity's Growth Index mutual fund and Fidelity's Value index mutual from 1984 to 2019.

1. Calculate and interpret the typical return for these two mutual funds.
2. Calculate and interpret the investment risk for these two mutual funds.
3. Determine which mutual fund provides the greater return relative to risk.

Over the 36-year period:

- The mean return for Growth was greater than the mean return for Value, or, equivalently, 15.76% > 12.01%. Therefore, Growth has higher return.
- The sample standard deviation (or variance) for Growth is greater than that of Value (23.799 > 17.979). Therefore, Value is less risky.
- This shows the importance of both measures (in this case the mean and the standard deviation) in the analysis of data.

Summary Measures using Excel

Example: The marketing analyst of an online retail company is trying to understand spending behavior of customers during the holiday season. She has compiled information on 130 existing customers that includes the customer's sex (Sex = Female or Male) and spending (in $) in the following categories: clothing (Clothing), health and beauty (Health), technology (Tech), and miscellaneous items (Misc). Use the AVERAGEIF function in Excel to find the average spending for each of the product categories for female customers and for male customers. Then, help the manager determine whether it seems appropriate to target females or males for the different product categories. Data: **Online.**

Sex	Clothing	Health	Tech	Misc
Female	225.67	100.25	47.10	159.88
Male	97.93	100.64	310.97	85.84

Given the means for the two groups, the manager should target females for clothing and miscellaneous products and males for technology products. Because females and males spend approximately the same on health products, the manager need not differentiate this market.

Practice Problems

1. The ***Match*** Data file has two worksheets. The first worksheet contains CEO compensation for 100 firms whereas the second worksheet contains total assets of only 86 firms. In Column C of the Assets worksheet, include the compensation of the corresponding 86 CEOs, matched by the firm name. Use the VLOOKUP function for matching. Find the mean of the matched compensation.

2. A social science study conducts a survey of 418 individuals about how often they exercise, marital status, and annual income. The ***Fitness*** file contains relevant data.
   - How many of the individuals who are married, and exercise sometimes earn more than $110,000 per year?
   - How many missing values are there in each variable?
   - How many individuals are married and unmarried?
   - How many married individuals always exercise? How many unmarried individuals never exercise?

3. **CarBuyers and CarPurchases.** A luxury car dealership wants to start an outreach campaign to its loyal customers. A sales manager downloads two relevant data files. The first file contains contact information of the customers who had purchased multiple sports cars in recent years including their customer IDs, names, cities where they live, age, and sex. The second data file includes the make and model of the sports cars that they had purchased, the date of purchase, and purchase price. The common variable between the two data sets is the customer ID. Merge the two data sets by matching the customer IDs. The consolidated data should include all purchased cars and variables from the two original data sets.
   - In the consolidated data, what is the average price of the vehicles purchased by male customers?
   - In the consolidated data, how many Porsche cars were purchased by the customers who live in Boston?

4. The accompanying file contains a portion of data from the National Longitudinal Survey (NLS), which follows over 12,000 individuals in the United States over time. Variables in this analysis include the following information on individuals: Urban (1 if lives in urban area, 0 otherwise), Siblings (number of siblings), White (1 if white, 0 otherwise), FamilySize, Height, Weight (in pounds), and Income (in $). The ***Longitudenal*** file contains relevant data. Are there any missing values in the data set? If there are, which variables have missing values? Which observations have missing values? Omit all observations (rows) that have missing values. How many observations are removed due to missing values?

5. Consider the ***IceCream*** file for an ice cream truck driver's daily income (Income in $), number of hours on the road (Hours), whether it was a hot day (Hot = 1 if the high temperature was above 85° F, 0 otherwise), and whether it was a holiday (Holiday = 1 if holiday, 0 otherwise).
   - Calculate the mean Income and the mean Hours.
   - Find the mean Income for a hot day and the mean Income for a non–hot day. Which subgroup has a higher Income?
   - Find the mean Income for a holiday and the mean Income for a nonholiday. Which subgroup has a higher Income?