Class 1: Regression Analysis: Introduction

Table of Contents / Agenda

1 Regression Analysis: Introduction

- How can we make predictions about real-world quantities, like sales or life expectancy?
- Most often in real world applications we need to understand how one variable is determined by **a number of others**

For example:

- How does sales volume change with changes in price. How is this affected by changes in the weather?
- How is the interest rate charged on a loan affected by credit history and by loan amount?
- We already used correlation coefficient to look at the relationship between *two* variables, but ...
- We cannot say that the correlation coefficient is a "pure" effect of one variable's change on another variable
  - e.g., What if x_1 (e.g., price) and x_2 (e.g., advertising) are also correlated?

ρ	Sales	Price	Advertising
Sales	1	-0.8	0.8
Price		1	-0.9
Advertising			1

1.1 Regression Analysis

- Let's you
  - Discover relationship between a dependent variable (y) and multiple independent variables (x's) jointly
  - Identify and measure each independent variable (x)'s impact on y separately
    - While *controlling for* (holding others constant) other variables

1.2 Relationship between x and y

- Essentially, we want to figure out the relationship between y (dependent variable) and x (independent, explanatory) variables:

$$y_i = f(x_{1i}, x_{2i}, \cdots)$$

  - Where
    - i: i'th observation, n: total number of observations
    - y_i: dependent variable
    - x_{ki}: i'th observation of k'th independent (explanatory) variable
    - $f(\cdot)$: the function specifying the relationship between y and x

- e.g.,

$$\underbrace{y_i}_{\text{Sales}_i} = f(\underbrace{x_{1i}}_{\text{Price}_i}, \underbrace{x_{2i}}_{\text{Promotion}_i})$$

- We basically want to know what $f()$ is. For example,

$$y_i = f(x_{1i}, x_{2i}) = 1 + 2 \times x_{1i} + 3 \times x_{2i}$$

1.3 Functional Form of f: Linear Regression

- In linear regression, we assume the dependent variable (y_i) to be a linear function of independent (or explanatory) variables (x_k's), coefficients (β_k's) and the error term (ε_i):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

- Where
  - β_k: coefficient for independent variable x_k, which represents the importance of x_k in y

- ε_i: the remaining part (error)
  - Unpredictable with x's
    - e.g., random-walk of stock prices

Note that β_0 is by itself since it corresponds to the constant term. That is, it represents the intercept, and you can think of it as x_{0i} being 1 everywhere ($\beta_0 \times 1 = \beta_0$).

2 Regression Analysis: Estimation

2.1 Estimation: Ordinary Least Squares (OLS)

- Again the regression model is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

- You can rearange terms and characterize the error by:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_{1i}$$

- Since y_i and x_{1i} are data so they do not vary. Then, as you change β_0 and β_1, ε_i will change.

Estimation Objective: **minimize** the sum of squared errors across all observations:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_{n-1}^2 + \varepsilon_n^2$$

- We want to find values of β_0 and β_1 that minimize the sum of squared errors

Fortunately, we have analytical solutions for the β_0 and β_1:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{1i} - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta}\,\bar{x}_1$$

- Where $\widehat{\beta}_k$: estimate (actual number) of coeffcient β_k

3 Interpretation of Regression Results: Fit (Model Level)

- Remember how we estimate coefficients (β_k's)?
- β_k which minimize the sum of squared errors are the estimates, $\widehat{\beta}_k$
- How do we measure how well our model performs?

3.1 Sum of Squares

Total sum of squares (SS_{total})

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- How much variation is in y (It's similar to variance)

Sum of Squared Errors (SS_{error})

$$SS_{err} = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_{n-1}^2 + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left\{ y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\text{predicted}} \right\}^2$$

3.2 Sum of Squared Errors (Residuals)

- SS_{error} is a measure of how wrong the regression estimates will be overall
- SS_{error} is a measure of variance
- y_i is sometimes higher, sometimes lower than the regression line
- Actual value of y_i varies because unobserved factors and randomness
- The regression can never be a perfect predictor

3.3 How well does regression fit?

- We can use these to construct a value which represents:
  - what % of total variance do we explain with our model?

$$\Rightarrow \frac{\text{explained variance}}{\text{total variance } (SS_{total})}$$

  - which can also be represented as

$$1 - \frac{\text{unexplained variance } (SS_{error})}{\text{total variance } (SS_{total})}$$

3.3.1 R^2

R^2 the percentage of variance in the dependent variable (y) explained by the independent variables (x's):

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}}$$

- R^2 is between 0 and 1 (0% to 100%)

4 Interpretation of Regression Results: Coefficients

- $\hat{\beta}_1$ (estimated coefficient for x_1): How much the **dependent variable (y)** is expected to change when the **independent variable (x_1)** increases by **one** unit
- Suppose we have x_1's value as 50, and $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 3$. Then, the predicted y value is:

$$\underbrace{\hat{\beta}_0}_{1} + \underbrace{\hat{\beta}_1}_{3} \times 50 = 151$$

- If we increase x_1 by 1:

$$\underbrace{\hat{\beta}_0}_{1} + \underbrace{\hat{\beta}_1}_{3} \times (50 + 1) = 154$$

- That is, y increases by $\hat{\beta}_1$ when we increase

'eee' is not recognized as an internal or external command, operable program or batch file.

- Mathematically,

$$\frac{\partial y}{\partial x} = \frac{\partial(\beta_0 + \beta_1 x)}{\partial x} = \beta_1$$

5 Multiple Regression

5.1 Multiple Regression

- Sales vs. Promotion Discount is an example of simple linear regression
- But sales of a brand depend upon many things
  - TV Ads, In-store promotions, Coupons etc ...
- When many things vary at the same time, it is hard to visually see the impact of each factor
- Multiple regression lets you look at an isolated effect of one variable

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \cdots + \beta_K x_{i,K} + \varepsilon_i$$

- Interpretation of $\hat{\beta}_k$: holding other variables constant, the change in y if you increase x_k by 1 unit
- Just like the simple regression, mathematically,

$$\frac{\partial y}{\partial x_k} = \frac{\partial(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_K x_K)}{\partial x_k} = \beta_k.$$

5.2 R^2 and Adjusted R^2

- Recall

$$R^2 = 1 - \frac{\text{unexplained variance } (SS_{error})}{\text{total variance } (SS_{total})} = 1 - \frac{SS_{error}}{SS_{total}}$$

- R^2 is between 0 and 1 (0% to 100%)

5.2.1 R^2 in multiple regression

- R^2 **always** becomes larger when we add more independent variables
- So we CANNOT use R^2 to compare the fit of two different regressions with different numbers of independent variables

5.2.2 Adjusted R^2

- We use **adjusted** R^2 to compare regressions with different numbers of independent variables

$$R^2_{adj} = 1 - \left\{ \frac{SS_{error}}{SS_{total}} \times \frac{n-1}{n-K-1} \right\}$$

  - n: number of observations
  - K: number of independent (x) variables included in the model

- Basically, you give a little bit of penalty for higher K
- A variable needs to reduce SS_{error} significantly to overcome the penalty
- Occam's razor:

  "Among competing hypotheses, the one with the fewest assumptions should be selected"

- Albert Einstein:

  "Everything should be made as simple as possible, but no simpler"

6 Running Regression Analysis in Excel

Note for Mac users

- It is required that you have **Excel 2016 for Mac** installed in your device for this course, as `Data Analysis ToolPak` and `Solver` add-ins are not available in previous versions of Microsoft Excel for Mac. It still has some limitations comparing to the Windows version, but you will be

able to accomplish most of tasks.

6.1 Activating `Analysis ToolPak` and `Solver` in Excel

- You have to activate the `Analysis ToolPak` and `Solver`. (You only need the former for running Regression, but we may need the latter as well)

6.1.1 On Windows

1. Click `File` tab -> Click `Options` -> `Add-Ins` category.
2. Near the bottom of the `Excel Options` dialog box, make sure that `Excel Add-ins` is selected in the `Manage` box, and then click `Go...`.
3. In the `Add-Ins` dialog box, select the check boxes for `Analysis ToolPak` and `Solver Add-in`, and then click `OK`.
4. If Excel displays a message that states it can't run this add-in and prompts you to install it, click `Yes` to install the add-ins.
5. On the `Data` tab, note that an `Analyze` group has been added. This group contains command buttons for `Data Analysis` and for `Solver`.

6.1.2 On Mac

1. Go to the `Tools` menu, select `Add-ins`
2. Check `Analysis ToolPak` and `Solver Add-in` and then click `OK`

6.2 Running Regression Analysis in Excel

- Click `Data` -> `Data Analysis` -> Select `Regression` -> `OK`
- `Input Y Range`: Specify the range of cells that contain values for the dependent variable (y), including the variable label (the first row)
- `Input X Range`: Specify the range of cells that contain values for the independent (predictor) variable (x), including the variable label (the first row)
  - For multiple regression, all x variables must be adjacent to each other. (no gaps)
  - Tip: You can use `Shift + Down Arrow` and `Ctrl + Shift + Down Arrow` to select a range of data conveniently
- Make sure `Labels` checkbox is checked
- Click `OK` (it will put the results on a new sheet)