Class 2: 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 Python

- First, let's import basic modules for data analysis:

```python
import os
import numpy as np
import pandas as pd
```

6.1 statsmodels module

- `statsmodels` is the de-facto statistical analyses library in Python.
- There are two ways of using statsmodels: 1) passing data explicitly, and 2) passing data as a `pd.DataFrame` with a formula specifying the model. We will focus on the second method, which is more intuitive
- To start, import `statsmodels.formula.api`:

```
>>> import statsmodels.formula.api as smf
```

- It has many statistical models that you can use. Let's inspect elements of the imported module by `smf.<TAB>`.
- You will see many models. In general, lowercased model names indicate that they will accept R-like formula. (e.g., `ols`)

6.2 patsy formula

- The R-like formula in Python is provided by `patsy`. In general `patsy` formula has the form of `y ~ x1 + x2`, which corresponds to the following model (constant skipped):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Let's create a formula where we regress variable `Compensation` on `WinPercentage`:

```
>>> formula = 'Compensasion ~ WinPercentage'
```

6.3 Building a Model

- Now we are ready to build the regression model with the formula and data. First, you can build your model object like this:

```
>>> smf.ols(formula=formula, data=df)
```

- Let's assign this object to a variable `model` and inspect its elements.

```
>>> model = smf.ols(formula=formula, data=df)
```

6.4 Running Regression Analysis

- You can use `.fit()` method of the `model` object to actually run the regression. You can assign the resulting object to a variable:

```
>>> results = model.fit()
```

- The most frequently used method of the fitted object is `.summary()`. It will print out the regression output:

```
>>> results.summary()
```

You can also inspect individual statistics of the results as well.