Financial Modeling and Analysis
Course Code: MATH 242
Module 4: Regression

Dr. Zahra Lakdawala

July 7, 2020

Outline

Day 1: Least Squares Approach

Regression Analysis

- Statistical technique to describe the relationship between variables
- Simplest case is to examine is one in which a variable Y is referred to as a dependent or target variable may be related to the independent variable X - also known as a simple regressor.
- If the relationship between X and Y is linear, then the equation of the line may take the form:

$$Y = \beta_1 + \beta_2 X$$

- The purpose of regression is to find the best fit line or the equation that expresses the relationship between Y and X

Exact or Deterministic Relationship

- Deterministic relationships are sometimes (although very rarely) encountered in business environments. For example, in accounting:

$$\begin{aligned} \text{assets} &= \text{liabilities} + \text{owner} \\ \text{equity total costs} &= \text{fixed costs} + \text{variable costs} \end{aligned}$$

- Deterministic relationships are the exception rather than the norm.

Fact is...

- Data encountered in finance are more likely to appear as if Y and X largely obey an approximately linear relationship, but it is not an exact relationship
- Still, it may be useful to describe the relationship in equation form, expressing Y as X alone - the equation can be used for forecasting and analysis, allowing for the existence of errors (since the relationship is not exact).
- So how to fit a line to describe the 'broadly linear' relationship between Y and X when the (x, y) pairs do not all lie on a straight line?

Approaches

- Consider the pairs (x_i, y_i). Let $\hat{y}_i$ be the 'predicted' value of y_i associated with x_i if the fitted line is used. Define $e_i = y_i - \hat{y}_i$ as the residual representing the 'error' involved
- If over- and under - predictions of the same magnitude arec onsidered to be equally undesirable, then the object would be to fit a line to make the absolute error as small as possible, but noting that the sample contains no bservations and given the relationship is inexact, it would not be possible to minimise all e_i's simultaneously.
- Thus, our criterion must be based on some aggregate measures.

Approaches to Line Fitting

- Eyeballing

Approaches to Line Fitting

- Eyeballing
- Minimise the sum of the errors:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Approaches to Line Fitting

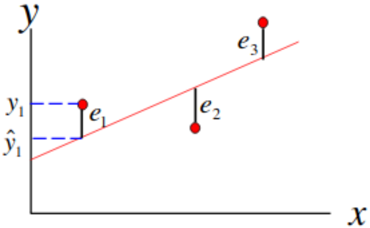
- Eyeballing
- Minimise the sum of the errors:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Welcome to the world of regression!

- The most common approach to estimating a regression equation is the least squares approach

Depiction of Error



Least Squares

This approach leads to a fitted line that minimises the sum of the squared errors, i.e.

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2\end{aligned}$$

Finding b_1, b_2

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i) = 0 \quad (1)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_2} = -2\sum_{i=1}^{n} x_i(y_i - b_1 - b_2 x_i) = 0 \quad (2)$$

- These are known as normal equations

Finding b_1, b_2 - contd.

- Solving the 2 equations leads to:

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_1 &= \bar{y} - b_2\bar{x} \end{aligned}$$

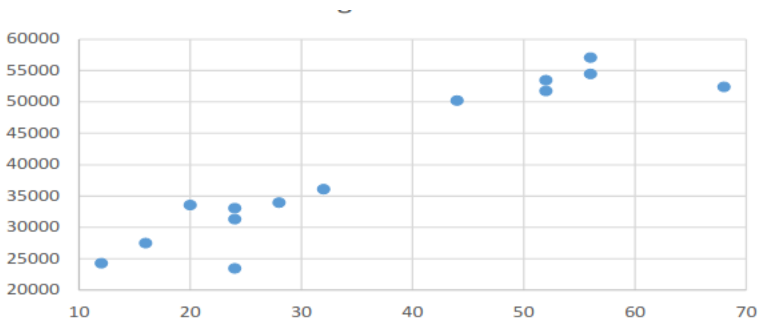
OR

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_1 &= \bar{y} - b_2\bar{x} \end{aligned}$$

Example

Example 1

To try to predict the price of Stock A, data were taken from the previous 12 months. The price of Stock A ($Y =$ Price) over time ($X =$ Time) was the only available information.



Steps in LS

- Find $x_i y_i$
- Find $\bar{x}$
- Find $\bar{y}$
- Find $\sum_{i=1}^{n} x_i^2$
- Find b_1, b_2 by substituting the values in the previously derived equations.

Advantages

The least squares approach has several advantages:

- The objective function $\sum_{i=1}^{n} e_i^2$ is strictly convex, meaning that the solution is always unique.
- The solution is easy to compute.
- The least squares approach leads to a set of well-established inference procedures
- The solution has an optimal mathematical property*

Disadvantages

There are also disadvantages associated with the least squares approach:

- The solution is sensitive to **outliers**
- The objective function $\sum_{i=1}^{n} e_i^2$ is symmetric, meaning that over and under-estimation of the same magnitude will be equally penalised. In some situations, one of the two errors may be considered as more serious than the other.

However, the advantages of least squares outweigh its disadvantages in most situations encountered in practice.

Least Squares and matrix formulation

- The basic matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \beta = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

- Notice that $\mathbf{x}\beta = \begin{bmatrix} b_1 + b_2 x_1 \\ \vdots \\ b_1 + b_2 x_n \end{bmatrix}$

Linear model in matrix form

- Define $\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$
- Note that

$$\mathbf{e} = \mathbf{y} - \mathbf{x}\beta$$

- Also

$$\sum_{i=1}^{n} \mathbf{e}_i^2 = \mathbf{e}^T \mathbf{e}$$

Linear Model in matrix form

This brings us to the following:

$$
\begin{aligned}
\mathbf{e}^T\mathbf{e} &= (\mathbf{y}-\mathbf{x}\beta)^T(\mathbf{y}-\mathbf{x}\beta) \\
&= \mathbf{y}^T\mathbf{y}-\beta^T\mathbf{x}^T\mathbf{y}-\mathbf{y}^T\mathbf{x}\beta+\beta^T\mathbf{x}^T\mathbf{x}\beta \\
&= \mathbf{y}^T\mathbf{y}-2\beta^T\mathbf{x}^T\mathbf{y}+\beta^T\mathbf{x}^T\mathbf{x}\beta
\end{aligned}
$$