FINC 305 LN2 - Linear Regression

Linxian Huang

December 27, 2024

In both Economics and Finance studies, we always begins with the following premise: y and x are two variables, representing some data we retrieved from real world, and we are interested in "how y varies with changes in x", such as "how return of Apple's stock change when the market portfolio return changes." Based on this consideration, y is called the **dependent variable**, the **independent variable**, **response variable**, or **predicted variable**. x is called **independent variable**, the **explanatory variable**, the **control variable**, the **predictor variable** or the **regressor**. Also, we can easiliy imagine that it has more than one factor that can significantly influence Apple's stock return. In this lecture, we will go further to review and study simple and multiple regresssions, discussing potential applications, and how to inprove the explanatory and predictive power of linear models you build.

1 Simple and Multiple Linear Regression

Linear regression model tries to explain a *dependent variable* (y) and one or more independent variables x_i. Just like how I interpreted previously, it has lots of terminology get used. A generic form of linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where:

- y is the dependent variable,
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients,

- $x_1, x_2, \ldots, x_p$ are the independent variables,
- ϵ is the error term.

By converting this regression model to a matrix form, it can be interpreted as:

$$Y = X\beta + \epsilon$$

where:

- Y is an $n \times 1$ vector of dependent variables,
- X is an $n \times k$ matrix of independent variables,
- β is a $k \times 1$ vector of coefficients,
- ϵ is an $n \times 1$ vector of errors.

For a single observation i, the model becomes:

$$y_i = x_i'\beta + \epsilon_i$$

where:

- y_i is a scalar dependent variable,
- x_i is a $k \times 1$ vector of independent variables,
- β is a $k \times 1$ vector of coefficients,
- ϵ_i is a scalar error term.

1.1 Assumption

- **Linearity**: the model specifies a linear relationship between y and $x_1, \ldots, x_k$. Variables can be transformed or non-linear, but the relationship is linear. For examples,

$$\ln y = \beta_1 + \beta_2 \ln x + \epsilon, \quad \text{or} \quad y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon$$

- **Full Rank**: According to the matrix form of linear regression, $rank(X) = k$, which means each variable (i.e., a column in X is linearly independent. Columns of X are not linear functions of other columns.
  Notes: x and x^2 are non-linear, so that's ok to present in a same regression

Example: $y = \alpha + I\beta_1 + \frac{I}{2}\beta_2 + \epsilon$, converting to $y = \alpha + I(\beta_1 + \frac{\beta_2}{2}) + \epsilon$, in such relationship, we can only identify $\tilde{\beta}$, which is $y = \alpha + I\tilde{\beta} + \epsilon$. Notes: It's ok for two variables (i.e., two columns of X) to be correlated as long as they are not linearly related.

- **Exogeneity of the Indenpendent Variables**: The expected value of the disturbance at observation i is not a function of the independent variables. In other words, the independent variables do not carry useful information for predicting $epsilon_i$.

$$E[\epsilon_i | x_{j1}, x_{j2}, \cdots, x_{jk})] = 0, \quad or \quad E[\epsilon | X] = 0$$

- **Homoscedasticity and Non-correlation**: Each disturbance ϵ_i has the same finite variance σ^2 and is uncorrelated with every other disturbance ϵ_j. (Note: 'Constant variance' = 'Homoscedasticity'. 'Heteroscedasticity' is opposite.) For Example:

$$Var[\epsilon_i | X] = \sigma^2, \quad \forall i = 1, \ldots, n, \quad \text{and} \quad \text{Cov}\left(\epsilon_i, \epsilon_j \mid x\right) = 0$$

- **Normal Distribution**: The distrubances are normally distributed.

1.2 Ordinary Least Squares Estimation (OLS)

The **Ordinary least square** (OLS) methods chooses the estimates to minimize the sum of squared residuals. In the linear regression with one independent variable only, the estimated OLS equation is written in a form:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

OLS tries to minimize the sume of squared residuals (SSR) (ϵ), which is making

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}\right)^2$$

as small as possible. The estimator is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sum_{i=1}^{n} \left((x_i - \bar{x})^2\right)}.$$

If we have multiple independent variables in the linear model, our target is still minimizing the SSR by using a squared residual to penalizes large

residuals, which is

$$\begin{aligned} SSR(\beta) &= (y - X\beta)'(y - X\beta) \\ &= (y' - \beta' X')(y - X\beta) \\ &= y'y - \beta' X'y - y'X\beta + \beta' X'X\beta \\ &= y'y - 2y'X\beta + \beta' X'X\beta \\ &= y'y - 2a'\beta + \beta' A\beta, \quad \text{where } a = X'y \text{ and } A = X'X. \end{aligned}$$

$$\frac{\partial SSR(\beta)}{\partial \beta} = -2a + 2A\beta = 0 \quad \Rightarrow \quad -2X'y + 2X'X\beta = 0,$$

$$(X'X)\beta = X'y \quad \text{thus} \quad \hat{\beta} = (X'X)^{-1}X'y.$$

In the above formula, $\hat{\beta}$ is a vector containing all the estimators of β, X and y are matrices containing all the observations of Xs and y.