

# FINC 305 LN2 - Linear Regression: Theory

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,
- $\beta_0$  is the intercept,
- $\beta_1, \beta_2, \dots, \beta_p$  are the coefficients,

- $x_1, x_2, \dots, x_p$  are the independent variables,
- $\epsilon$  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,
- $\beta$  is a  $k \times 1$  vector of coefficients,
- $\epsilon$  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,
- $\beta$  is a  $k \times 1$  vector of coefficients,
- $\epsilon_i$  is a scalar error term.

## 2 Assumption

- **Linearity:** the model specifies a linear relationship between  $y$  and  $x_1, \dots, x_k$ . 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,  $\text{rank}(X) = k$ , 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 Independent 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  $\epsilon_i$ .

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

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

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

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

### 3 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 sum of squared residuals (SSR) ( $\epsilon$ ), which is making

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

as small as possible. The estimator is

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

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  $\beta$ ,  $X$  and  $y$  are matrices containing all the observations of  $X$ s and  $y$ .

## 4 Goodness of Fit

The most frequent measure of goodness of fit (e.g., how much of the variation in  $y$  is explained by variation of  $x$ 's in linear model) is the R-Square ( $R^2$ ). To calculate the  $R^2$ , we need to determine:

- **Total Sum of Squares(SST)**,  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$
- **Explained Sum of Squares (SSE)**,  $SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
- **Sum of Square residuals (SSR)**,  $SSR = \sum_{i=1}^n (y_i - x_i\hat{\beta})^2$
- **SST = SSR+SSE**

For some notations, the predicted  $y_i$  is  $\hat{y}_i = x_i\hat{\beta}$ ,  $y_i = x_i\hat{\beta} + e_i$ , where  $e_i = \hat{\epsilon}_i$  is the predicted residual, then  $y_i = \hat{y}_i + e_i$ .

The percentage of variation of  $y$  explained by regression is:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

While the  $R^2$  is the standard measure of fit, it has the potentially unattractive feature that it always increase with more x's (independent variables). Recall the equation to calculate  $R^2$ ,

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\hat{\sigma}_\epsilon^2}{\sigma_y^2}$$

But it doesn't correct because these are biased estimate of these variances. To generate unbiased estimation, we should have unbiased  $\hat{\sigma}_\epsilon^2 = \frac{SSR}{n-k-1}$ , and unbiased  $\sigma_y^2 = \frac{SST}{n-1}$ , where k is the number of regressors on the right hand side of the equation, n is the number of observations. Then substitute back to previous equation,

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-k-1} \times (1 - R^2)$$

- The first thing that we can notice is that  $\frac{n-1}{n-k-1}$  is always greater than 1. Therefore,  $\bar{R}^2$  is always less than  $R^2$ .
- This "penalizes" are measured by decreasing in the number of parameters (k).
- By adding a new regressor in the equation,  $SSR$  falls, which will increase  $\bar{R}^2$ . On the other hand,  $\frac{n-1}{n-k-1}$  increases as we add one more regressor, which will decrease  $\bar{R}^2$ . Whether  $\bar{R}^2$  increases or decreases depends on which of these two effects is stronger.
- $\bar{R}^2$  can be negative. Whenever  $\frac{SSR}{TSS}$  is large enough that it may not be able to offset  $\frac{n-1}{n-k-1}$ ,  $\frac{n-1}{n-k-1} \cdot \frac{SSR}{TSS}$  may be greater than 1, which will result in  $\bar{R}^2 < 0$ .

## 5 Potential Issues in Linear Regression (Tentative)

### 5.1 Multicollinearity

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated. This makes it difficult for the model to isolate the effect of each independent variable on the dependent variable. For example, you are predicting a company's stock price based on variables net income, total cost and total revenue. Then

$$R_i = \beta_1 NI_i + \beta_2 TC_i + \beta_3 TR_i$$

However, Net income also can be determined by:

$$NI_i = TR_i - TC_i$$

Since net income is determined by total revenue and total cost, which are highly correlated, multicollinearity will exist.

The existence of multicollinearity may inflate the **standard errors** of the estimated coefficients, making them statistically insignificant even when they may be meaningful. Also, it reduces the interpretability of the coefficients because it becomes unclear how much each variable contributes uniquely to the dependent variable.

To detect multicollinearity, we apply **Variance Inflation Factor (VIF)**, which is calculated by:

$$VIF = \frac{1}{1 - R^2}$$

A  $VIF > 10$  is a common threshold for severe multicollinearity. To solve multicollinearity, you can either remove redundant variables, or combine two high-correlated variables as a new features to regress.

## 5.2 Heteroscedasticity

## 5.3 Overfitting