Linear Regression

Hang Zheng

June 2025

Contents

Part I

Simple Linear Regression

Chapter 1

Linear Regression with One Predictor Variable

1.1 Simple Linear Regression Model with Distribution of Error Term Unspecified

Formal Statement of Model

The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.1}$$

where:

- Y_i is the value of the response variable in the ith trial
- β_0 and β_1 are parameters
- X_i is a known constant, namely, the value of the predictor variable in the ith trial
- ϵ_i is a random error term with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = \sigma^2$; ϵ_i and ϵ_j are **uncorrelated** so that their covariance is zero (i.e., $\sigma\{\epsilon_i, \epsilon_j\} = 0$ for all $i, j; i \neq j$)

Important Features of Model

1. Y_i is a random variable with sum of two components. The constant term $\beta_0 + \beta_1 X_i$, and the second term ϵ_i.
2. Since $E(\epsilon_i) = 0$,then we have

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_1$$

Which means:

$$E(Y_i) = \beta_0 + \beta_1 X_i \tag{1.2}$$

Therefore, the regression function for the model in is

$$E(Y) = \beta_0 + \beta_1 X \tag{1.3}$$

3. The variance of Y is

$$\sigma^2(Y) = \sigma^2(\epsilon_i) = \sigma^2 \tag{1.4}$$

Alternative Versions of Regression Model

Let X_0 be a constant identically equal to 1. Then, we can rewrite (1,1) as follows:

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \epsilon_i, \text{where } X_0 = 1 \tag{1.5}$$

To leave model (1.1) unchanged, we write :

$$Y_i = \beta_0 + \beta_1(X_i - \bar{X}) + \beta_1 \bar{X} + \epsilon_i$$

Thus the alternative model is:

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i \tag{1.6}$$

where:

$$\beta_0^* = \beta_0 + \beta_1 \bar{X} \tag{1.6a}$$

1.2 Estimation of Regression Function

Method of Least Squares

The criterion Q is defined as:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

given the observations $(X_1, Y_1),\ (X_2, Y_2), \cdots (X_n, Y_n)$. We set

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$
$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Then we have

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

1.3 Estimation of Error Term Variance σ^2

The sum of squares SSE has $n-2$ degree of freedom. Hence, the appropriate mean square, denoted by MSE or s^2, is

$$s = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-2}$$

It can be shown that:

$$E(s^2) = \sigma^2$$

Which means that the MSE is an unbiased estimator for regression in (1.1)

1.4 Definition of Linear Models

Definition 1.4.1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2) + \beta_3 X_1 X_2 + \epsilon$$

is a linear model.

Definition 1.4.2

$$e_i = y_i - \hat{y}_i$$

*is the residual of the response variable y_i.(Be careful to the **order**).*

We want to minimize the sum of squared error.which is:

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

The quantity called *coefficient of determination* is denoted by R^2:

$$R^2 = \frac{SSE(\bar{y}) - SSE(\hat{y})}{SSE(\bar{y})}$$

where $SSE(\hat{y})$ is the unexplained variation of y.
$SSE(\bar{y})$ is the total variation of y.
Remarks:

1. $0 \leq R^2 \leq 1$ and $R^2 = 1$ if perfect.
2. R^2 is unit-less and can be sensitive to extreme x values.

A numerical measurement of the strength of the linear association is *Pearson correlation coefficient*:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where n is the sample size and $s_x,\ s_y$ are sample variance.

1.5 The standard assumption

Given model $Y_i = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon_i$, we assume that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

Definition 1.5.1 *A regression model is a formal means for expressing two essential ingredients of a statistical relation:*

1. $Y \sim u(X_1, X_2, \cdots X_n)$
2. *The means of these distribution vary in some systematic ways as a function of* $(x_1, x_2, \cdots, x_n)$

Estimating σ^2:An unbiased estimator of σ^2 is MSE:

$$MSE = s^2 = \frac{SSE}{n-p}$$

where n is the sample size and p is the number of parameters.
Root Mean Square Error:

$$RMSE = s = \sqrt{MSE}$$

R calls s the **'residual standard error'**.