Linear Regression

Simple Linear Model

$y_i = \beta_o + \beta_1 x_i + \epsilon_i$

where, β_o : y-intercept, value of x when y is zero.

β_1 : slope, predicted increase in Y resulting from a one unit increase in x.

Estimated Linear Model

$\hat{y_i} = \hat{\beta_o} + \hat{\beta_1} x_i + \epsilon_i$

Error

$\epsilon_i = y_i - \hat{y_i}$

Sum of square due to error

$\sum_{i-1}^{n} \epsilon_i^2 = \sum_{i-1}^{n} (y_i - \beta_o - \beta_1 x_i)^2$

Differentiating above equation with respect to β_o and β_1 we get

$\hat{\beta_1} = \frac{\sum_{i-1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i-1}^{n} (x_i - \bar{x})^2}$

and

$\hat{\beta_o} = \bar{y} - \hat{\beta_1} \bar{x}$

Error Sum of Squares

$\epsilon_i = y_i - \hat{y_i}$ and sum of squares due to error is denoted by SSE and is given by: $SSE = (y_i - \hat{y_i})^2$

Sum of squares SSE has n-1 degrees of freedom associated with it. Two degrees of freedom is lost because both βo and β_1 had to be estimated in obtaining estimated means $\hat{y_i}$.

$s^2 = \frac{SS}{n-2}$

$s^2 = \frac{(y_i - \hat{y_i})^2}{n-2}$

Coefficient of Determination and Correlation

The total deviation $y_i - \bar{y}$ (used in the measure of the total variation of the observations y_i without taking the predictor variable x_i into account) can bedecomposed into two components:

$y_i - \bar{y} = y_i - \hat{y_i} + \hat{y_i} - \bar{y}$

Total deviation = deviation around fitted regression line + deviation of fitted regression value around mean

After some mathematical calculation we can verify that :

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

Total sum of Squares = Sum of Squares due to Error + Sum of Squares due to regression

TSS = SSE + SSR

Corefficient of determination is defined as

$R^2 = \frac{SSR}{SST}$

Non-Linear Functionl forms

In some cases linear relationship between the variables may not be adequate to model their relationship. There are cases for which a non-linear functional form is more suitable.

Simply transforming variables y and/or x and then estimating a regression model using the transformed variables is the simplest way of obtaining a non-linear specification.

A log-log functional form is specified as;

$log(y_i) = \beta_o + \beta_1 log(x_i) + \epsilon_i$

Some of the Useful transformations

Regular Linear form

$y_i = \beta_o + \beta_1 x_i + \epsilon_i$

log-log form

$log(y_i) = \beta_o + \beta_1 log(x_i) + \epsilon_i$

Linear-log form

$y_i = \beta_o + \beta_1 log(x_i) + \epsilon_i$

Log-linear form

$log(y_i) = \beta_o + \beta_1 x_i + \epsilon_i$

Question : How to interpret the coefficients ?