Simple Linear Models

Model and Model Assumptions

Given the data

y	x
y_1	x_1
y_2	x_2
$\dots$	$\dots$
y_n	x_n

We would like to approximate y_i with a function of x_i, $f(x_i)$. Assume that this function is linear

$$y_i \approx \hat{y_i} = f_1(x_i) = \beta_0 + \beta_1 x_i \tag{1}$$

Let the error/residual of each approximation be $e_i = y_i - \hat{y_i}$

To do inferential statistics, estimating the error of $\hat{\beta_1}$ or finding confidence interval for β_0, we need to bring in a probability distribution. Assume that x_i is a known constant (non-random) and that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2}$$

where

$\epsilon_1, \epsilon_2, ..., \epsilon_n \sim^{iid} N(0, \sigma)$. From this equation, we see that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. S_{xy} and S_{yy} are random, S_{xx} is non-random.

Coefficient Estimators

There are multiple ways to define the quality of the approximation. For example, in the least square method, one wants to minimize the total sum of the errors or residual sum square (RSS) or sum square errors (SSE)

$$RSS = SSE = \sum e_i^2 = e_1^2 + e_2^2 + ... + e_n^2 = \sum e_i^2 \tag{3}$$

Notice that we use the summation notation ($\sum e_i$) to present a sum of all e_i^2 for i from 1 to n. A full version of $\sum$ is $\sum_{i=1}^{n}$. We will use $\sum$ instead of $\sum_{i=1}^{n}$ for simplicity.

In the least absolute error or least absolute deviations, one wants to minimize

$$\sum e_i = |e_1| + |e_2| + ... + |e_n| \tag{4}$$

The values of β_0 and $beta_1$ that minimizes RSS is denoted by $\hat{\beta_0}$ and $\hat{\beta_1}$, respectively.

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} \tag{5}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x_i}^2 \tag{6}$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y_i}^2 \tag{7}$$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{8}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{9}$$

$$RSS = SSE = \sum(y_i - \hat{y_i})^2 = \sum(y_i - \hat{\beta_0} - \hat{\beta_1}x_i)^2 \tag{10}$$

We have some properties for the β_0 and β_1 as follows.

$$E\hat{\beta_0} = \beta_0 \tag{11}$$

$$Var(\hat{\beta_0}) = \sigma^2\frac{\sum x_i^2}{nS_{xx}} \tag{12}$$

$$E\hat{\beta_1} = \beta_1 \tag{13}$$

$$Var(\hat{\beta_1}) = \frac{\sigma^2}{S_{xx}} \tag{14}$$

$S^2 = \frac{SSE}{n-2} = \frac{S_{yy}}{\hat{\beta_1}\S_{xy}}$ is an unbiased estimator for σ^2

Goodness of Fit

We have the following formula

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

The quantity $TSS = \sum_{i=1}^{n}(y_i - \bar{y})$ can be though as the total amount of `information` (variance) in the response variable y. This quantity is decomposed into $RSS = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})$ which can be thought as the `information` explained by the resgression models and $RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})$, which is the amount of `information` that can not be explained by the model (the sum of squared erros of the models).

This formula leads to the definition of R^2

$$R^2 = 1 - \frac{RSS}{TSS}$$

Another way to view this formula is that TSS is actually the sum squared errors of the naive model $y = \bar{y}$ (predicting y value by its average, disregarding x). Thus,

$$R^2 = 1 - \frac{\text{Sum squared errors of Linear Model}}{\text{Sum squared errors of the naive Model}}$$

Thus, R^2 also measure how good the linear model is when comparing with the naive model. This definition of R^2 can be applied to measure the goodness of fit of other models, not just linear model.

To measure the goodness of fit of this linear approximation, we compare the approximation with a naive approximation, which is to approximate all y_i with their average, $\bar{y}$. The naive model is a special case of the linear model where $\beta_1 = 0$. In the other words, the naive models is

$$y = f_2(x) = \bar{y} \tag{15}$$

The sum square error of the naive models is $\sum(y_i - \bar{y})^2$. This quantity is also called the total sum square (SST or TSS). We use the following R^2 to measure the goodness of fit of the linear model

$$R^2 = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} \tag{16}$$

It's noticed that if the linear approximation fit the data perfectly ($SSE = 0$), $R^2 = 1$.

Inferences on Coeffient Estimators

F- Test We are interested in testing the following hypotheses for β_1

$$H_0 : \beta_1 = 0 H_\alpha : \beta_1 \neq 0$$

$$F = \frac{RegSS}{RSS/(n-2)} \sim F_{1,n-2}$$

t-test

$$H_0 : \beta_1 = d H_\alpha : \beta_1 \neq d$$

We have that under H_0

$$t = \frac{\hat{\beta_1} - d}{\sqrt{s^2/S_{xx}}} \sim t_{n-2}$$

It's noticed that the F-test and the t-test have the same rejection region when $d = 0$.

Practice Problems