Linear Regression

This chapter is about linear regression, a straightforward approach for supervised learning. Linear regression is useful for predicting quantitative response from a set of predictor variables. Moreover, it determines which variables in particular are significant predictors of the outcome variables and in what way do they *indicated by the beta estimates* impact the outcome variable. These regression estimates are used to explain the relationship between predictor variable X and response variable Y.

Simple Linear Regression

Simple linear regression is an approach for predicting a quantitative response Y on the basis of a single predictor variable X. Assumed there is approximately a linear relationship between X and Y. The linear relationship will be mathematically written as

$$Y \approx \beta_0 + \beta_1 X$$

β_0 and β_1 are unknown constants that designate the *intercept* and *slope* in the linear model and together they are known as the model *parameters* . Then, after we have estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model parameters by using the training data, we can predict the response variable Y by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

with $\hat{y}$ designates the prediction of Y on basis $X = x$.

Estimating the Parameters

As said before, β_0 and β_1 are unknown. Before we make predictions, we have to use the data to estimate the parameters. Let

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

designate n observation pairs, each of which consists of a measurement of X and Y. The main objective here is to get paramater estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that fit the linear model, so that

$$\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$i = 1, \ldots, n.$$

Here we want to acquire parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ that will make the regression line as close as possible to the n data points. We will take the *least squares* approach to acquire these parameters. Figure 1 represents the simple linear regression model.

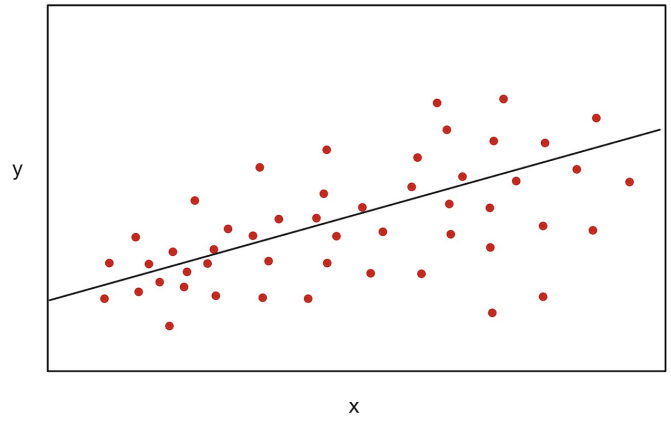

Figure 1: Simple linear regression model, the red dots represent data points

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on x_i (ith value of X). Then $e_i = y_i - \hat{y}_i$ designates the ith *residual*. *Residual* is the difference between ith actual response value and the ith predicted response value from our linear model. Next we define the *residual sum of squares* (RSS) as

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

and furthermore

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

To obtain optimum $\hat{\beta}_0$ and $\hat{\beta}_1$, we have to minimize the RSS. The *least squares parameter estimates* will be written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2},$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

with $\overline{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$ and $\overline{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$ as the sample means.

R^2 Statistics

The R^2 statistic provides the measure of fit. It is the proportion of variability in Y that can be explained by using X, has the value between 0 and 1, and also independent of the scale of Y. To calculate R^2, we use the formula

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

where TSS $= \sum(y_i - \overline{y})^2$ is the *total sum of squares*. TSS measures the total variance in Y. R^2 value near 0 indicates that the model did not explain much of the variability in the response, meaning that the regression model could be wrong.

Multiple Linear Regression

As explained before, simple linear regression is a practical approach to predict a response on the basis of a single predictor variable. Unfortunately, in reality we often come up with more than one predictor variable. How do we integrate these extra predictors to make our analysis? One solution would be by making separate simple linear regressions, each of them using different predictor. This solution, however, is not that effective, because we will have different regression equation for each predictor so that a single prediction would be hard to conclude. Another thing is that by using simple linear regression, each of the equation will exclude the other predictors in forming estimates for the parameters. Instead, we will extend the simple linear regression by giving each predictor its own separate slope in a single model. Assumed that we have i predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \epsilon,$$

where X_i designates the ith predictor and β_i is the slope for each ith predictor and is interpreted as the average effect on Y of a one unit increase in X_i, while other predictors fixed. ϵ designates the residual of the regression.

Estimating the Parameters

As we have shown in the simple linear regression, the regression parameters $\beta_0, \beta_1, \ldots, \beta_i$ are also unknown and therefore must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_i$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_i x_i.$$

The parameters are going to be estimated by using the least square approach, like it was the case in simple linear regression. We choose $\beta_0, \beta_1, \ldots, \beta_i$ to minimize the sum of squared residuals

$$RSS = \sum_{j=1}^{n} (y_j - \hat{y_j})^2$$

$$= \sum_{j=1}^{n} (y_j - \hat{\beta}_0 - \hat{\beta}_1 x_{j1} - \hat{\beta}_2 x_{j2} - \cdots - \hat{\beta}_i x_{ji})^2$$

Selecting Significant Variables

On many cases, it is possible that only some of the predictors are related with the response. To determine which predictors are related to the response, so that we can make a single model only for those predictors, is called *variable selection*. In our case, we will be using *Akaike Information Criterion*(AIC) to determine which variables are significant.

Logistic Regression

In linear regression model, the response variable Y is assumed quantitative. But in other situations, the response variable is instead qualitative or also referred as categorical. The task is now we predict the probability of each categories of a qualitative variable , as the base for making our final prediction. For example, if someone took a loan, then it's either they *have* or *have not* paid it back. Then we could code these qualitative response as:

$$Y = \begin{cases} 0 & \text{if not default} \\ 1 & \text{if default} \end{cases}$$

After that we could make a linear regression to this binary response, and predict **paid** if $Y > 0.5$ and **unpaid** otherwise. But rather than making a linear model to this response, logistic regression models the probability that Y belongs to a certain category. For example, the probability of default can be written as

$$p(X) = Pr(Y = 1|X)$$

Logistic Model

As mentioned before, the response variable in logistic regression is qualitative. Therefor we cannot model the probability by using linear regression model. One of the reasons is that the predicted probability value would not fall between 0 and 1 if we use linear model. We must then model $p(X)$ using a function that gives results between 0 and 1 for all values of X. In logistic regression, we use the *logistic function*,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

After a bit of manipulation, we got

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

The left-hand side of the equation is called the *odds* and have value between 0 and ∞. Then by giving both sides the logarithm, we will have

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

The left-hand side is now called the *log-odds* or *logit*.

Estimating the Parameters

The parameters β_0 and β_1 are also unknown like the case in linear regression and they need to be estimated based on training data. The preferred approach is the *maximum likelihood*. The idea is that we look for $\hat{\beta}_0$ and $\hat{\beta}_1$ by plugging these estimates into the model for $p(X)$ yields a number close to one for all $Y = 1|X$, and a number close to zero for all $Y = 0|X$. This can be formalized using *likelihood function*:

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize the likelihood function. For our topic we will not go deep into the mathematical details of maximum likelihood as it can be easily fit using **R** function that we will discuss more later on. Once the parameters have been estimated, we can put them in our model equation:

$$\hat{p}(X) = \frac{e^{\hat{\beta_0}+\hat{\beta_1}X}}{1 + e^{\hat{\beta_0}+\hat{\beta_1}X}}.$$

Multiple Logistic Regression

Assumed now that we have multiple predictors. Just like on linear regression, we can extend the formula that we have as follows:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X + \cdots + \beta_p X_p,$$

where $X = (X_1, \ldots, X_p)$ are p predictors. The equation can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X+\cdots+\beta_p X_p}}.$$

We also use the maximum likelihood to estimate $\beta_0, \beta_1, \ldots, \beta_p$.