Group 19 project 1st draft **Group memebers**
Stats 506 Mimi Tran, Zhenbang Jiao, Yijia Zhang
Date: Nov/26 2018

1 Topic

Topic: Linear regression

1.1 Introduction

Formal definition from *Linear Models with R*: Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the response, output or dependent variable; and one or more predictor, input, independent or explanatory variables, $X_1...X_p$. When $p = 1$, it is called simple regression but when $p > 1$ it is called multiple regression or sometimes multivariate regression. (Faraway, J 2009)
An informal interpretation is that linear regression establish a linear relationship between the response variable and independent variable(s) in a data set. It is widely applied in many areas such as sciences, engineering and machine learning.
Mathematically:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{1}$$

1.2 Solving linear regression problems with least square

One well defined way to solve for β is called the least square method:
Using above definition, we can write:

$$y = X\beta + \epsilon$$

with $y = [y_1, ..., y_n]^T$, $\epsilon = [\epsilon_1, ..., \epsilon_n]^T$, $\beta = [\beta_0, ..., \beta_m]^T$ and:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{d2} & \dots & x_{nm} \end{bmatrix} \tag{2}$$

In above expression, ϵ is the residual, and one way to obtain best estimation of β is to minimize:

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon \tag{3}$$

The above equation can be rewritten as:

$$(y - X\beta)^T (y - X\beta) \tag{4}$$

Now, we take the differential of (4) with respect to β, notice that to minimize (4), the graident should be zero; thus:

$$X^T X \hat{\beta} = X^T y \tag{5}$$

in which $\hat{\beta}$ is our estimation of β. Rearranging (5) we get:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{6}$$

This is the norm equation.

1.3 example with data

Need to fill

2 Diagnostic and Model Selection

Normally, there are numerous variables in a data set, and some of variables are actually correlated; then one problem is how to establish a rather simple model which could be easily interpret. One popular technique is the *criterion-based Model Selection*. The idea is to choose model with respect to a specific criterion that measures the behavior of fit. In our project, we will introduce three common criterion: the *Akaike information criterion (AIC)*, the *Bayes information criterion (BIC)*, and the *adjusted R square*.

2.1 AIC

Before defining AIC, we first define the following values:
Number of independent predictors p: $p = \#X_i$
Residual sum of square - RSS:

$$RSS = \hat{\epsilon}^T \hat{\epsilon} = (y - X\hat{\beta})^T (y - X\hat{\beta}) \tag{7}$$

And

$$AIC = n \ln (RSS/n) + 2(p + 1) \tag{8}$$

Pick the model minimizes AIC.

2.2 BIC

$$BIC = n \ln (RSS/n) + (p + 1) \ln n \tag{9}$$

Pick the model minimizes BIC.

2.3 Adjusted R square

The R square value is defined as:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{10}$$

The adjusted R square value is:

$$1 - \frac{n-1}{n-p-1}(1 - R^2) \tag{11}$$

Select a model that maximize the adjusted R^2 value

2.4 Examples

Fill

References

To be fill in