Some Problems in Multiple Linear Regression Models

ISYE 6740, Fall 2020, Final Report

Shasha Liao

Georgia Tech

Contents

1 Problem Statement

Multiple linear regression model, one of the most basic statistic model, is very easily to be ignored after we learn fancier models likes Neural Networks. However, it is still a very common topic at data science interviews. In this project, we are going to explore some potential problems we may encounter when we fit a linear regression model to a particular dataset. More specifically, we are going to discuss how to identify and overcoming the following problems:

1. Nonlinearity of the response-predictor relationships.
2. Correlation of the error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

A simple linear regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{1}$$

where β_0 is the intercept term, X_i represents the ith feature, β_i quantifies the relation between X_i and the response Y, and ϵ is a mean-zero random error term.

2 Data Source

After searching a lot for a proper dataset, we found that it is not easy to find one with all the problems we want to explore. So we are going to simulate my own dataset.

- x_1: $x_1 \sim N(\mu_1, \sigma_1^2)$ with some outliers added, designed for high leverage points.
- x_2: $x_2 \sim N(\mu_2, \sigma_2^2)$ with some outliers added, designed for high leverage points.
- x_3: $x_3 \sim N(\mu_3, \sigma_3^2)$ with some outliers added, designed for high leverage points.
- x_4: $x_4 \sim N(\mu_4, \sigma_4^2)$, designed for nonlinearity after taking log.
- x_5: $x_1 x_2$, designed for nonlinearity.
- x_6: $x_1 + 2x_2 + 10x_3$, designed for multi-colinearity.
- ϵ: $x_1 \tilde{\epsilon}$ with $\tilde{\epsilon} \sim N(0, \sigma^2)$, designed for non-constant variance of error terms or dependence of the predictors and error terms.
- Y: $x_1 + x_2 + x_3 + \log x_4 + x_5 + x_6 + \epsilon$, with some outliers added. This will be our predictor.

The design of our data set is subject to change if necessary as our experiments goes on.

3 Methodology

We will try different methods to overcome each problem and evaluate the resulting model in different ways for comparison. Firstly, we will do some EDA to study the distribution of each predictor and do some transformations if necessary to change the distribution into a good shape.

1. Nonlinearity of the response-predictor relationships.

   We will fit our model and make a plot of fitted values VS residuals to see if there is a clear nonlinear pattern in residuals. If so, we should consider creating nonlinear transformation of predictors, such as $\log x_i$, $x_i x_j$ and x_i^2 for our model.

2. Non-constant variance of error terms.

   Again, we will fit our model and make a plot of fitted values VS residuals to check if the variance of the residuals changes. If so, we can consider either make a nonlinear transformation of the response Y or fit our model by weighted least squares, with weights proportional to inverse variance.

3. Outliers.

   Outliers are unusual response y_i given the predictor x_i. We will plot the studentized residuals, which is computed by dividing each residual by its standard error. Observations whose studentized residuals are greater than 3 in absolute value are likely to be outliers. We will remove the outliers, fit the model again, and compare the R^2 statistics.

4. High-leverage points.

   Observations with high-leverage have unusual value for predictor x_i. We will compute the leverage statistic of an observation and check if it greatly exceeds $(p+1)/n$. If so, this observation is considered as a high-leverage point. We will remove all the high-leverage points from our dataset, fit the model again, and compare the R^2 statistics.

5. Collinearity.

   We will compute the Variance Inflation Factor (VIF) of each predictor to check collinearity. A VIF value that is greater than 5 or 10 indicates a problematic amount of collinearity. We can either remove the problematic predictor or take an average of standardized versions of the collinear predictors in order to create a new variable.

6. Correlation of the error terms.

   If two errors are correlated, then knowing one error will help us to guess the other error. We can analyze the plot of the residuals and check if there is tracking to see if there is correlation between the residuals. We won't go deep to seek for solutions to solve this problem. In practice, the experiments should be carefully designed to ensure uncorrelated errors.

4 Evaluation

We will evaluate our linear regression models using statistics like RSS, RSE and R^2 given in the following formulas:

- Residual Sum of Squares: $RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$
- Residual Standard Error: $RSE = \sqrt{\frac{RSS}{n-p-1}}$ where $n-p-1$ is the degree of freedom since we have n data points and $p+1$ parameters. RSE measures the lack of fit of the models to the data. It is an absolute measure, which means it depends on the scale of y_i. We can use it together with the R^2 statistic.
- R^2 statistic: $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$ where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})$ is the total sum of squares. The R^2 statistic measures the proportion of variance explained by the fitted model.

5 Final Results