1 Maximum Likehood for Linear Regression

We consider the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ with n observation points $(x_1, y_1), \ldots, (x_n, y_n)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We estimate β_0, β_1 and σ^2 by b_0,b_1, and s^2 respectively which we have chosen. For a certain response y_i, we want to find the probability of observing this response given our chouse in parameters. We get

$$P(y_i \mid x_i; b_0, b_1, s^2) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(y_i - (b_0 + b_1 x_1))^2}{2s^2}}.$$

What we are calculating here is the probability of the error ϵ covering the difference between y_i and $b_0 + b_1 x_i$. Similarly, for the probability of observing the entire dataset, we obtain

$$\prod_{i=1}^{n} P(y_i \mid x_i; b_0, b_1, s^2).$$

We can take the log to get

$$\mathcal{L}(b_0, b_1, s^2) = \log \prod_{i=1}^{n} P(y_i \mid x_i; b_0, b_1, s^2) = \sum_{i=1}^{n} \log P(y_i \mid x_i; b_0, b_1, s^2).$$

Finally, we can maximize the above equation to get the estimators.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\sigma^2} &= \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta^1} x_i))^2
\end{aligned}$$

2 Optimizing the Likelihood Function

The scripts are provided in `gradientDescent.R` and `likelihoodFunction.R`.

3 Model Validation Methods

1. Through the script written in `P2_3.R`, we find the table of predictions.

	True	
Pred	0	1
0	434	11
1	10	228

From this table, we get a total misclassification percentage of approximately 3.07%.

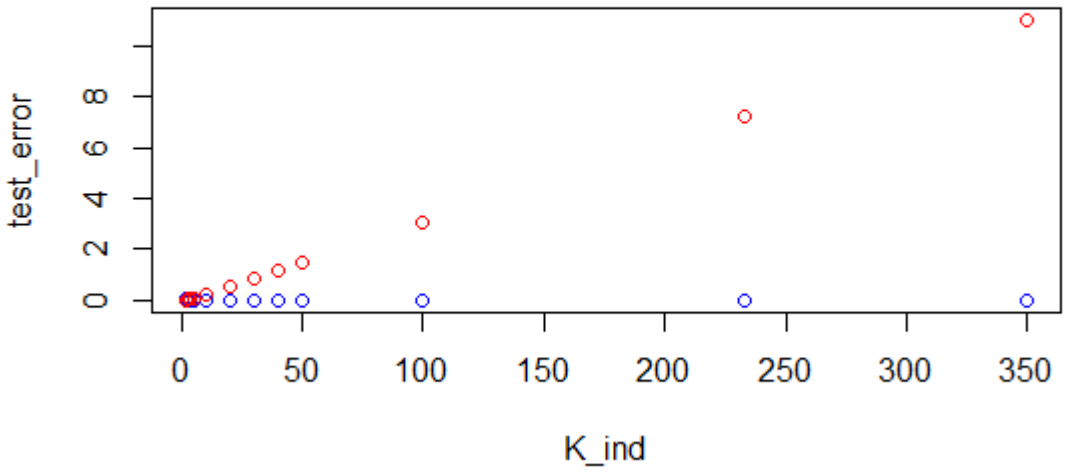

Figure 1: Red: Train Error, Blue: Test Error

2. Through the script written in `P2_3_2.R`, we find the table of predictions. For the training/testing split, I used a 70/30 split.

	True	
Pred	0	1
0	133	5
1	2	66

   From this table, we get a total misclassification percentage of approximately 3.40%.

3. From `P2_3_3.R`, we calculate 10-Fold cross-validation and get a misclassification percentage of 3.35%.

4. We use the script in `P2_3_4.R` to see the error image in Figure 1. The training error appears to be monotonically increasing, so in this case it is best to use a low value for K around $2, 3$ or 4.

4 Cross-Validation Intuition

In this case, I would choose the second method. The reason is that LOOCV has high variance, and if the dataset has 10,000 predictor variables, it is likely to be large in row size as well making LOOCV computationally expensive. In the second method, we drop the variables with the lowest variance, meaning the variables are relatively unchanging and have little effect on the response variable. Although only variables with relatively higher variance remain, 10-fold cross validation does not inherintly have high variance, so this variance is kept under control.