

# Machine Learning: Block 1 Lab 1 (Group K6)

Yiran Wang, Tore Andersson and Shashi Nagarajan Iyer

11/11/2020

## Statement of Contribution:

All assignments were completed independently by each team member and approaches/results were discussed within the group. Reports for assignments 1, 2 and 3 were produced primarily by Yiran, Shashi and Tore respectively.

## 1. Assignment 1 - Handwritten digit recognition with K-NN

### 1.1 Import data and divide it into training, validation and test sets

The first task is to import digit data and divide it into three sets for training, testing and validation respectively with 50%, 25% and 25% of the data. We use `sample()` with seed set as 12345.

### 1.2 Train KNN model and estimate the result

This task consists of training a knn model with 30 nearest neighbors and estimate the result with confusion metrics and misclassification errors for both the training and test data.

The confusion matrix for train data is:

```
##      pred_train
##      0   1   2   3   4   5   6   7   8   9
## 0 202   0   0   0   0   0   0   0   0   0
## 1   0 179  11   0   0   0   0   1   1   3
## 2   0   1 190   0   0   0   0   1   0   0
## 3   0   0   0 185   0   1   0   1   0   1
## 4   1   3   0   0 159   0   0   7   1   4
## 5   0   0   0   1   0 171   0   1   0   8
## 6   0   2   0   0   0   0 190   0   0   0
## 7   0   3   0   0   0   0   0 178   1   0
## 8   0  10   0   2   0   0   2   0 188   2
## 9   1   3   0   5   2   0   0   3   3 183
```

The confusion matrix for test data is:

```
##      pred_test
##      0   1   2   3   4   5   6   7   8   9
## 0  77   0   0   0   1   0   0   0   0   0
## 1   0  81   2   0   0   0   0   0   0   3
## 2   0   0  98   0   0   0   0   0   3   0
## 3   0   0   0 107   0   2   0   0   1   1
## 4   0   0   0   0   0  94   0   2   6   5
## 5   0   1   1   0   0  93   2   1   0   5
## 6   0   0   0   0   0   0  90   0   0   0
## 7   0   0   0   1   0   0   0 111   0   0
```

```
## 8 0 7 0 1 0 0 0 0 70 0
## 9 0 1 1 1 0 0 0 1 0 85
```

With these two matrix, we can simply have the accuracy rate based on correct numbers over total numbers for each digits:

```
##      0      1      2      3      4      5      6      7
## 1.0000000 0.9179487 0.9895833 0.9840426 0.9085714 0.9447514 0.9895833 0.9780220
##      8      9
## 0.9215686 0.9150000

##      0      1      2      3      4      5      6      7
## 0.9871795 0.9418605 0.9702970 0.9639640 0.8623853 0.9029126 1.0000000 0.9910714
##      8      9
## 0.8974359 0.9550562
```

For train data, the quality of predicting different digits is with a large range between 91% and 100%. The prediction for digit 0 is 100%, and following with 6, 2, 3 and 7 with 97% to 99% rate. While the lower ones are for digits 4, 9 and 1 with about 91% correct rate. This could indicate that for some classes with 91% accuracy rate in the train data, the model does not fit quite well.

For test data, the accuracy rates for prediction on digits do not have a huge difference than for train data. Some of them are higher than the train data prediction rate such as for 1, 6, 7 and 9. The highest one is 100% for digit 6. And other ones are below the train ones with a lowest rate for digit 4 being 86%. As we can see here, the rates for digits comparing with them for train data do not have a big difference, which means the model generalizes well and not over-fitting.

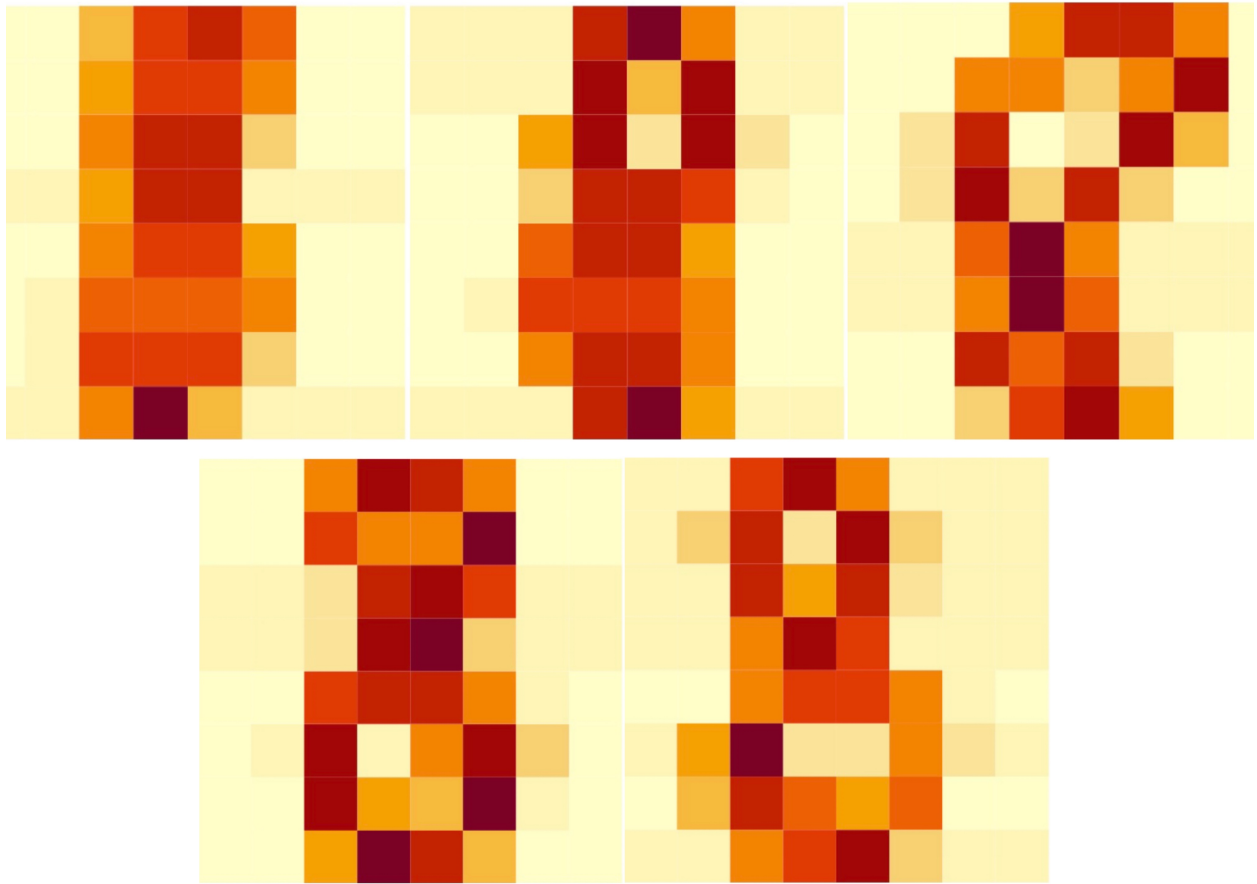
We can also have the mis-classification rate calculated as below:

```
## The mismatch rate for train data is: 0.04500262
## The mismatch rate for test data is: 0.05329154
```

For the overall mismatch rate 0.045 for train data and 0.053 for test data, we can see that the rates for both of them are fairly low and the difference between them is also not high. This can mean that the model is decent. But for a digit classification case with only 10 numbers, 5% error rate could mean that 1 in 20 observations would be classified wrong, which seems to be not good enough.

### 1.3 Identify and plot cases with low and high correct rates

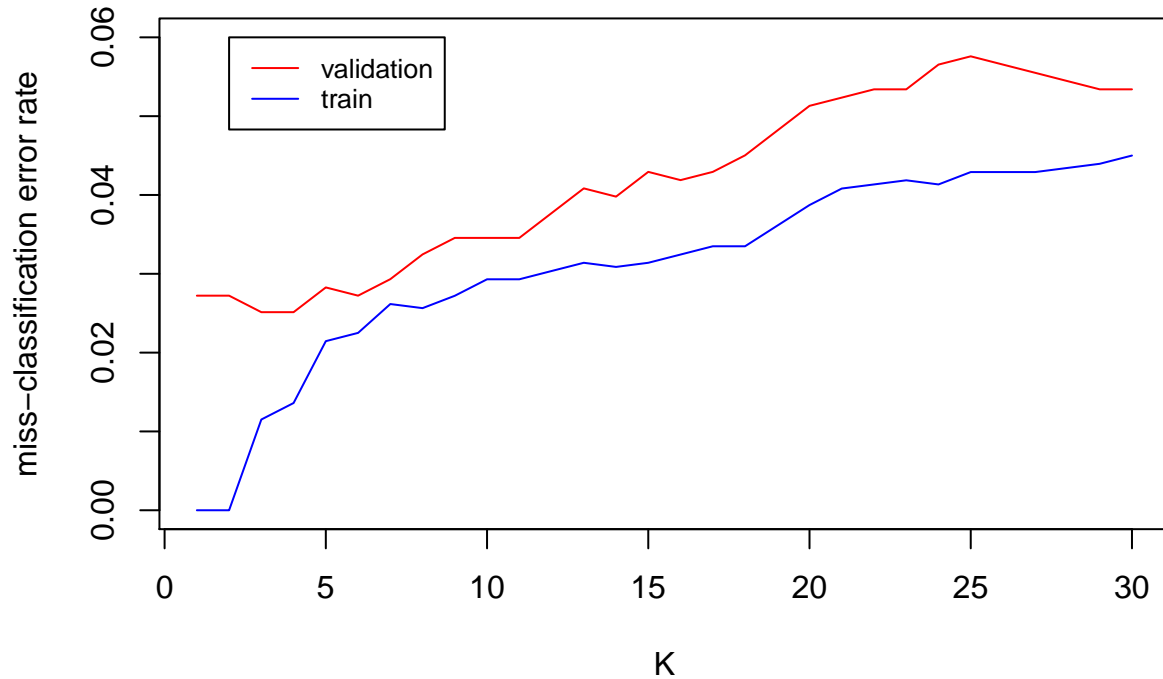
This task is to find cases that are hard to classify and easy to classify and plot them with heatmap plots:



The cases that are hard to see are chosen with lowest correct rates being 0.1000000, 0.1333333 and 0.1666667. While for the easy to identify ones, the rates are both 100%. As the heatmap shows, the cases that has lowest accuracy rates(the top three showing on the image) are hard to recognized as digits 8 when visualizing them as well. All of them are either too blur or the shape is not clear enough to identify as digit 8. While the 100% accurate cases(the bottom two on the image) are very easy to see as 8.

#### 1.4 Use different knn models to fit the train data and analyze with miss-classification rate

This is about fitting the train data with  $K=1..30$  knn models and plot the mis-classification rates for train and validation data. Then observe the plot in different perspective. The plot is shown as below:



As  $k$  increases, the complexity of the model for both data sets also increases overall. As we can see in the plot, the error rate for the validation data decreases first to 0.025 when  $K$  is 3 and stays a bit before going up quickly as  $K$  increases with a final slight drop till 0.05. While the error rate for the train data starts with 0 when  $k$  is 0, and then goes up gradually as  $K$  goes up.

The optimal  $K$  is a little difficult to detect right away since the differences for different  $K$  values are very small, especially when  $K$  is about 3 to 7. But we take 4 as the optimal one based on only the rate we plot here. As at that point, the value for validation rate is the lowest. The difference between validation and train data is not the smallest but also not very high. From the perspective of bias-variance trade-off, the difference could be interpreted as variance and the error rate can be seen as bias. In this case, we want the bias to be as small as possible and compromise with the variance not being perfect. Low bias is not always perfect as it can lead to a higher variance (when error of train data is 0, the error of validation data is relatively higher). When bias is a little bit higher, the variance can be a bit lower as it can generalize new data better. But after certain point, when bias gets too high, the error rate for both train and validation data will be high, which would be not a suitable model as it may be under-fitting.

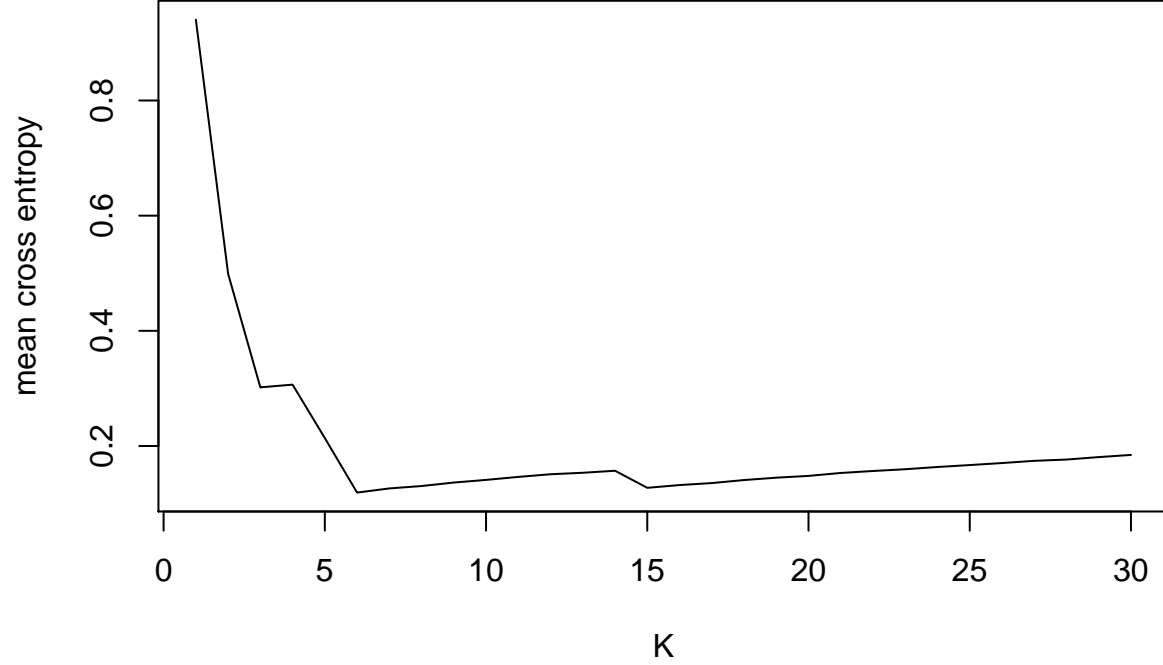
With the optimal  $K$  value selected based on above analysis, the result of the model for test data can be shown as:

```
## Test mis-classification rate when k=4 is: 0.02507837
## Train mis-classification rate when k=4 is: 0.01360544
## Validation mis-classification rate when k=4 is: 0.02513089
```

The test error being 2.51% with  $K=4$  is about the same as the error rate of validation data and a little higher (about 1.15%) higher than train data. We think the quality of the model with  $K=4$  is acceptable because with a similar relatively high fitting rate to both validation and test data (not over-fitting) and well fitted for train data (not under-fitting).

### 1.5 Use different knn models to fit the train data and analyze with empirical risk

This task is also about fitting train data with different  $k$  values. But estimate the result with empirical risk, in this case, cross entropy value. This plot of the result is as follows:



As the plot shows, the optimized K value is 6, because at that point, the risk value is the lowest with a value being 0.1191288. The reason cross entropy is a better approach of estimating optimal model here is because it uses the probabilities of each class for every observation. The lower the the probability the model has for the correct class, the more penalty it will get. But for the mis-classification method, it only outputs 0 or 1, which means that even if the probabilities are slightly different or huge differ from each other, the penalty would be the same. Thus, the way cross-entropy works can lead the model to be trained more correctly.

## 2. Assignment 2 - Ridge regression and model selection

### 2.1-3 Bayesian Ridge regression model, Data Preparation and Functions to run Ridge regression

Bayesian ridge regression model for the data (without the intercept parameter) is:

$$\mathbf{y}|\mathbb{X}, w_0, \mathbf{w} \sim N(w_0 + \mathbb{X}\mathbf{w}, \sigma^2\mathbb{I}),$$

where:

- $\mathbf{y}$  is the vector of Parkinson's disease symptom scores, "motor UPDRS", the dependent variable
- $\mathbb{X}$  is the design matrix containing the observed values of all the independent variables corresponding to  $\mathbf{y}$  (voice characteristics in this case)
- $\mathbf{w}$  is the vector of bayesian ridge regression parameters
- $\sigma$  is a scalar such that  $\sigma^2 \cdot \mathbb{I}$  is the covariance matrix of the linear regression error terms

The Bayesian prior for  $\mathbf{w}$  is:

$$\mathbf{w} \sim N(0, \frac{\sigma^2}{\lambda} \mathbb{I}),$$

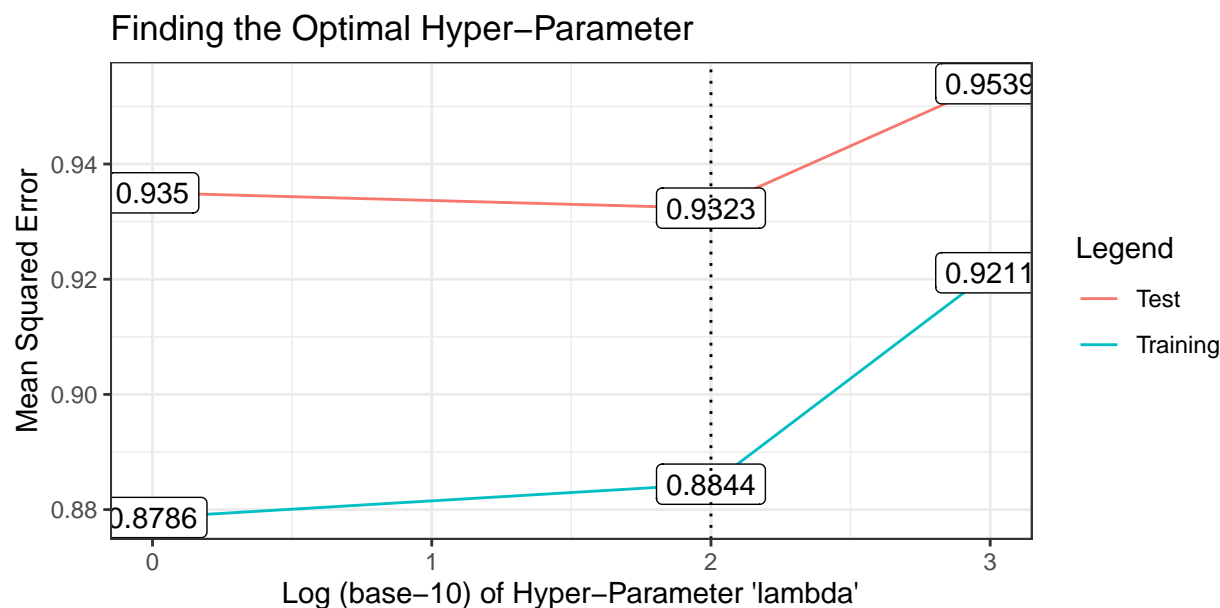
where:

- $\lambda$  is the ridge penalty hyper-parameter

Data and functions were prepared as required (see appendix for code)

### 2.4.1 Training and test MSE values

Fitting the above mentioned ridge regression model on the training dataset, with  $\lambda = 1, 100$  and  $1,000$ , we see the training and test MSE values as shown in the plot below. Note that the test data was scaled using the same means and standard deviations used to scale training data (this is as recommended here)



### 2.4.2 Most appropriate value of penalty parameter lambda

We can see that the optimal value of  $\lambda$  within the set  $\{1, 100, 1000\}$  is 100. At  $\lambda = 1$ , there is overfitting, evidenced by low training error and large test error; at  $\lambda = 1000$ , on the other hand, there is underfitting, evidenced by high training and test error.

### 2.4.3 MSE is a more appropriate measure here than other empirical risk functions

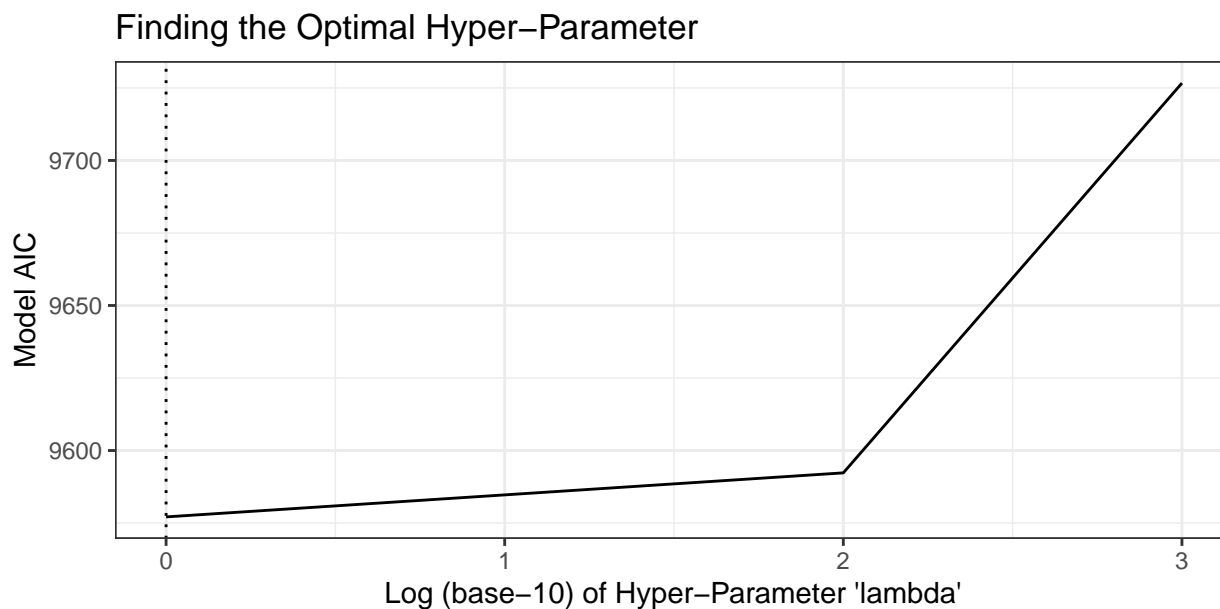
MSE is a more appropriate measure here than other empirical risk functions such as MAE (mean absolute error), because in regression problems, MSE can, unlike MAE and similar measures, also serve as an estimator of the variance of the error terms. Such an estimate would be useful in validating the underlying assumptions of a regression model (such as independence of error terms). Of course empirical risk functions such as mis-classification rate or cross-entropy do not apply in regression problems such as this, making MSE more appropriate than such measures.

### 2.5.1 Optimal model according to AIC criterion

In terms of AIC, the optimal value of  $\lambda$  within the set  $\{1, 100, 1000\}$  is 1 (see plot below); note that degrees of freedom are calculated according to Equation 3.50 in the book *The Elements of Statistical Learning*, Hastie et al, Second Edition, Springer, in page 68.

### 2.5.2 Theoretical advantage of AIC based model selection compared to holdout model selection

The theoretical advantage of using AIC for model selection is that it penalises models for high complexity (high degrees of freedom), which holdout model selection does not do; in reality, however, AIC is influenced both by data likelihood given model and degrees of freedom. In the plot below, we see evidence of AIC being dominated by the likelihood factor and in effect, we end up selecting the model  $\lambda = 1$ , as it has lowest AIC.



## Part done by Tore so far

So 3.6 and 3.7, are not finished need to do the hypothesis test, and was abit unsure as what to plot thought it to be abit unclear for 3.6.

### 3.1

```
##
## Call:
## lm(formula = Fat ~ ., data = train_fat)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.201500	-0.041315	-0.001041	0.037636	0.187860

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-1.815e+01	5.488e+00	-3.306	0.01628	*
Channel1	2.653e+04	1.126e+04	2.357	0.05649	.
Channel2	-5.871e+04	3.493e+04	-1.681	0.14385	
Channel3	1.154e+05	7.373e+04	1.565	0.16852	
Channel4	-2.432e+05	1.175e+05	-2.070	0.08387	.
Channel5	3.026e+05	1.193e+05	2.536	0.04430	*
Channel6	-2.365e+05	8.160e+04	-2.898	0.02741	*
Channel7	1.090e+05	3.169e+04	3.440	0.01380	*
Channel8	-6.054e+04	1.508e+04	-4.015	0.00700	**
Channel9	7.871e+04	2.160e+04	3.643	0.01079	*
Channel10	-1.730e+04	1.640e+04	-1.055	0.33215	
Channel11	9.562e+04	3.529e+04	2.710	0.03512	*
Channel12	-2.114e+05	6.198e+04	-3.410	0.01431	*
Channel13	9.725e+04	4.424e+04	2.198	0.07026	.
Channel14	5.296e+04	4.666e+04	1.135	0.29968	
Channel15	-7.855e+04	5.245e+04	-1.498	0.18491	
Channel16	-8.209e+03	1.893e+04	-0.434	0.67969	

## Channel17	3.769e+04	1.987e+04	1.897	0.10666	
## Channel18	3.306e+04	7.934e+03	4.167	0.00590	**
## Channel19	-8.405e+04	1.929e+04	-4.358	0.00478	**
## Channel20	1.510e+05	3.361e+04	4.492	0.00414	**
## Channel21	-2.069e+05	4.256e+04	-4.862	0.00282	**
## Channel22	1.348e+05	3.824e+04	3.526	0.01243	*
## Channel23	-4.094e+04	3.546e+04	-1.154	0.29222	
## Channel24	2.023e+04	2.761e+04	0.733	0.49134	
## Channel25	3.269e+03	1.071e+04	0.305	0.77045	
## Channel26	-1.297e+04	7.636e+03	-1.699	0.14028	
## Channel27	4.131e+03	1.422e+04	0.291	0.78120	
## Channel28	-4.548e+03	2.988e+04	-0.152	0.88402	
## Channel29	1.089e+04	1.768e+04	0.616	0.56072	
## Channel30	-7.985e+04	2.653e+04	-3.010	0.02371	*
## Channel31	1.756e+05	5.279e+04	3.326	0.01589	*
## Channel32	-1.107e+05	2.904e+04	-3.813	0.00883	**
## Channel33	-6.525e+04	5.407e+04	-1.207	0.27294	
## Channel34	1.007e+05	6.589e+04	1.528	0.17738	
## Channel35	-2.841e+03	1.214e+04	-0.234	0.82266	
## Channel36	-2.268e+04	2.295e+04	-0.988	0.36127	
## Channel37	-4.479e+04	1.292e+04	-3.468	0.01334	*
## Channel38	3.209e+04	1.843e+04	1.742	0.13221	
## Channel39	1.992e+04	2.067e+04	0.964	0.37246	
## Channel40	-9.833e+03	2.431e+04	-0.404	0.69988	
## Channel41	1.659e+04	3.648e+04	0.455	0.66531	
## Channel42	-1.829e+04	3.528e+04	-0.519	0.62260	
## Channel43	-2.423e+04	2.427e+04	-0.998	0.35669	
## Channel44	3.246e+04	2.013e+04	1.613	0.15793	
## Channel45	-8.089e+03	4.023e+04	-0.201	0.84728	
## Channel46	7.065e+03	2.810e+04	0.251	0.80990	
## Channel47	-4.062e+04	1.007e+04	-4.034	0.00685	**
## Channel48	9.080e+04	2.618e+04	3.469	0.01332	*
## Channel49	-6.647e+04	2.372e+04	-2.803	0.03105	*
## Channel50	-4.196e+04	2.856e+04	-1.469	0.19213	
## Channel51	1.097e+05	5.572e+04	1.968	0.09661	.
## Channel52	-1.148e+05	6.376e+04	-1.800	0.12196	
## Channel53	9.525e+04	7.450e+04	1.278	0.24830	
## Channel54	-4.534e+04	7.363e+04	-0.616	0.56067	
## Channel55	-1.535e+03	4.933e+04	-0.031	0.97618	
## Channel56	-2.377e+03	2.109e+04	-0.113	0.91394	
## Channel57	3.174e+04	1.005e+04	3.158	0.01961	*
## Channel58	2.221e+03	1.048e+04	0.212	0.83915	
## Channel59	-8.504e+04	2.574e+04	-3.304	0.01634	*
## Channel60	6.382e+04	1.607e+04	3.972	0.00735	**
## Channel61	2.151e+04	1.234e+04	1.742	0.13211	
## Channel62	-2.859e+04	1.065e+04	-2.685	0.03631	*
## Channel63	1.796e+04	9.187e+03	1.955	0.09838	.
## Channel64	5.759e+04	3.526e+04	1.633	0.15354	
## Channel65	-1.470e+05	6.911e+04	-2.127	0.07752	.
## Channel66	9.121e+04	4.461e+04	2.045	0.08688	.
## Channel67	-5.733e+03	2.197e+04	-0.261	0.80288	
## Channel68	-6.290e+04	2.192e+04	-2.870	0.02843	*
## Channel69	6.421e+04	2.074e+04	3.096	0.02121	*
## Channel70	-1.749e+04	1.581e+04	-1.106	0.31111	



```

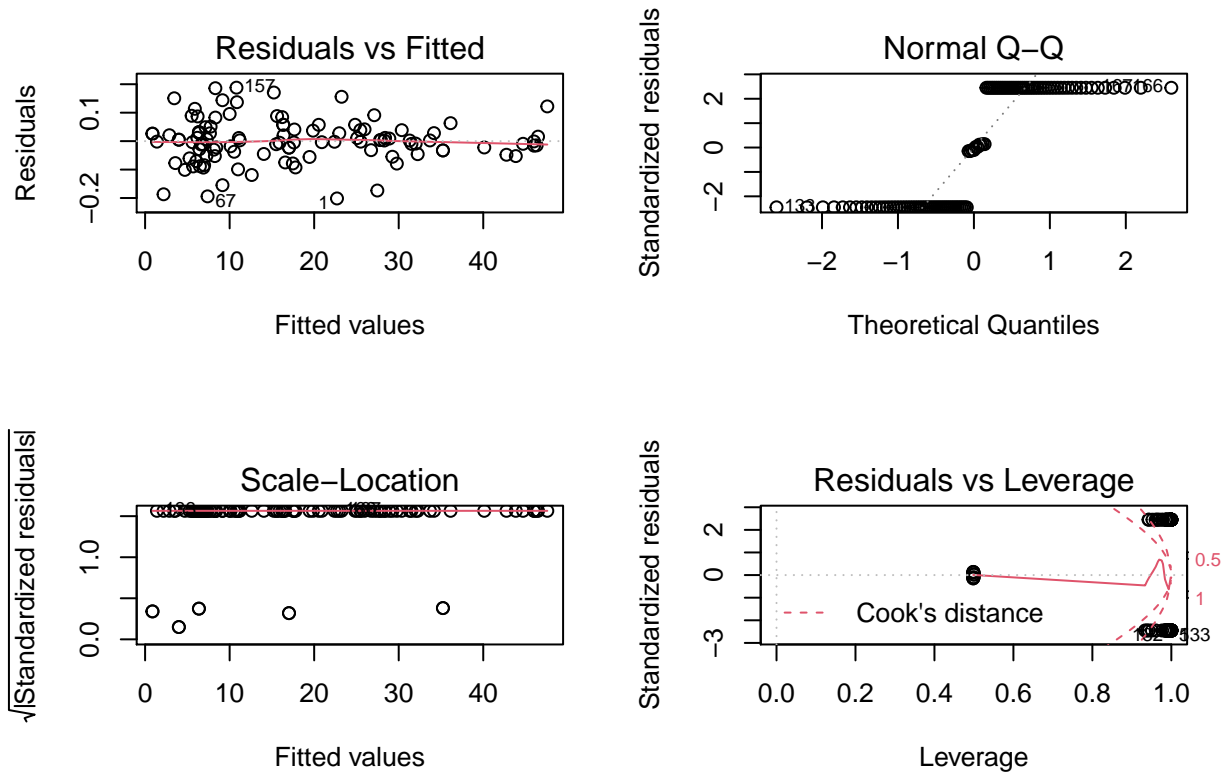
## Channel71    -7.248e+03  1.934e+04  -0.375  0.72075
## Channel72     3.406e+04  1.185e+04   2.873  0.02830 *
## Channel73    -2.100e+04  1.132e+04  -1.855  0.11308
## Channel74    -3.314e+04  1.220e+04  -2.717  0.03480 *
## Channel75     7.039e+04  2.054e+04   3.427  0.01402 *
## Channel76    -3.187e+04  1.736e+04  -1.836  0.11597
## Channel77     2.061e+04  1.810e+04   1.138  0.29832
## Channel78    -1.180e+04  2.273e+04  -0.519  0.62225
## Channel79     2.669e+04  2.997e+04   0.890  0.40750
## Channel80    -6.051e+04  1.483e+04  -4.080  0.00650 **
## Channel81     1.386e+03  2.628e+04   0.053  0.95966
## Channel82     1.020e+05  4.694e+04   2.173  0.07275 .
## Channel83    -1.706e+05  4.688e+04  -3.640  0.01083 *
## Channel84     1.097e+05  2.892e+04   3.792  0.00905 **
## Channel85    -1.294e+05  3.600e+04  -3.594  0.01145 *
## Channel86     2.130e+05  4.345e+04   4.903  0.00270 **
## Channel87    -1.198e+05  3.818e+04  -3.139  0.02011 *
## Channel88    -2.199e+04  6.085e+04  -0.361  0.73021
## Channel89     7.974e+04  5.077e+04   1.571  0.16733
## Channel90    -1.711e+05  5.499e+04  -3.112  0.02079 *
## Channel91     2.107e+05  6.406e+04   3.289  0.01663 *
## Channel92    -1.959e+05  7.171e+04  -2.733  0.03407 *
## Channel93     2.874e+05  9.937e+04   2.892  0.02762 *
## Channel94    -3.064e+05  9.601e+04  -3.191  0.01881 *
## Channel95     2.048e+05  6.220e+04   3.292  0.01656 *
## Channel96    -5.600e+04  2.929e+04  -1.912  0.10441
## Channel97    -1.318e+04  3.050e+04  -0.432  0.68065
## Channel98    -2.724e+04  2.107e+04  -1.292  0.24375
## Channel99     3.556e+04  1.382e+04   2.573  0.04218 *
## Channel100   -1.206e+04  4.264e+03  -2.828  0.03006 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3191 on 6 degrees of freedom
## Multiple R-squared:      1, Adjusted R-squared:  0.9994
## F-statistic: 1651 on 100 and 6 DF, p-value: 1.058e-09
## Training MSE:  0.005709117 . Test MSE:  722.4294

With the very low training MSE and comparably high test MSE we have reason to believe that there is
overfitting occurring with the model. From viewing the summary for the training model it was found to give
a adjusted R-square of 0.9994% which gives an clear indication that the model is overfit.

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

```



Analyzing the plots we find that the assumptions that the model residuals are i.i.d can be assumed to be approximately fulfilled, it's clear that the residuals do not follow anything similar to a normal distribution with mean 0 and variance 1 when observing the QQ-plot. Hence the assumptions for the model are not fulfilled so we should not draw any conclusions from the model.

### 3.2

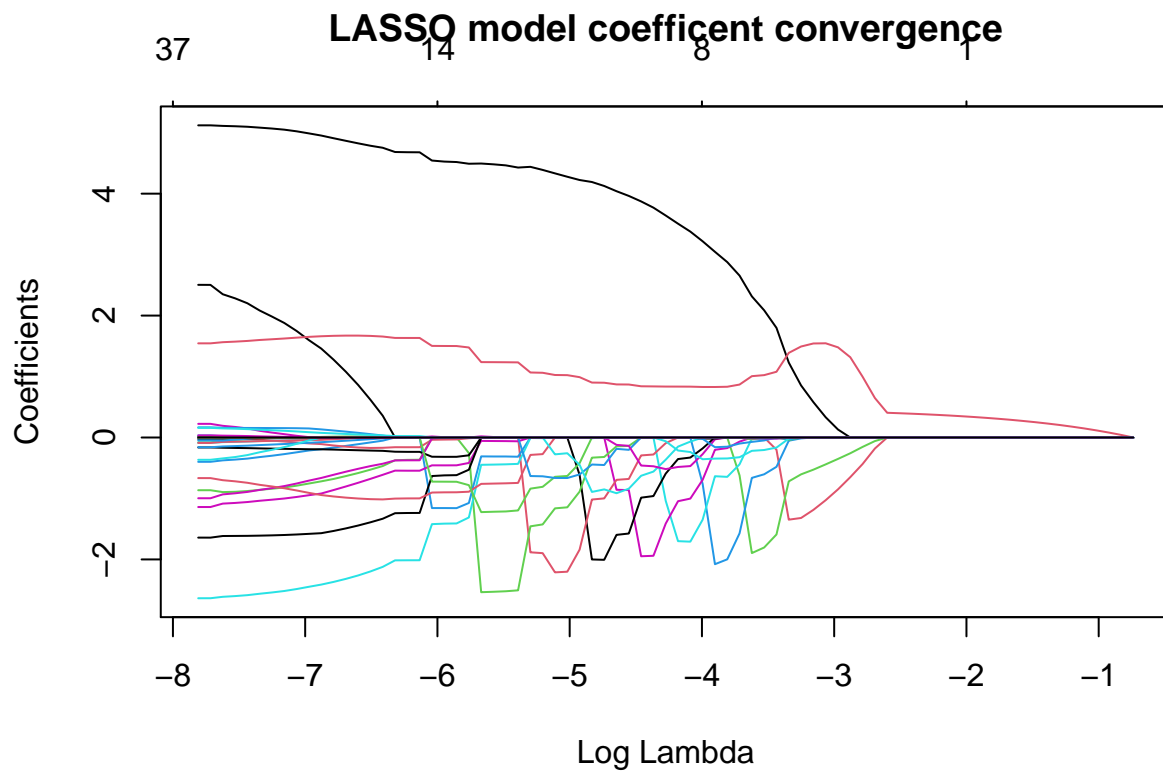
The function we are looking to optimize in this case is loss function of the LASSO model.

$$\sum_{i=1}^N (y_i - \mathbb{X}_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Where  $N$  is the number of observations and  $\mathbb{X}$  is the matrix of features and  $y$  is the dependent variable.  $\beta$  are the parameters that the shrinkage function right part of the plus sign. With  $\lambda$  being the shrinkage parameter and  $P$  the number of features. Where the chosen  $\lambda$  will shrink the  $\beta$  parameters.

### 3.3

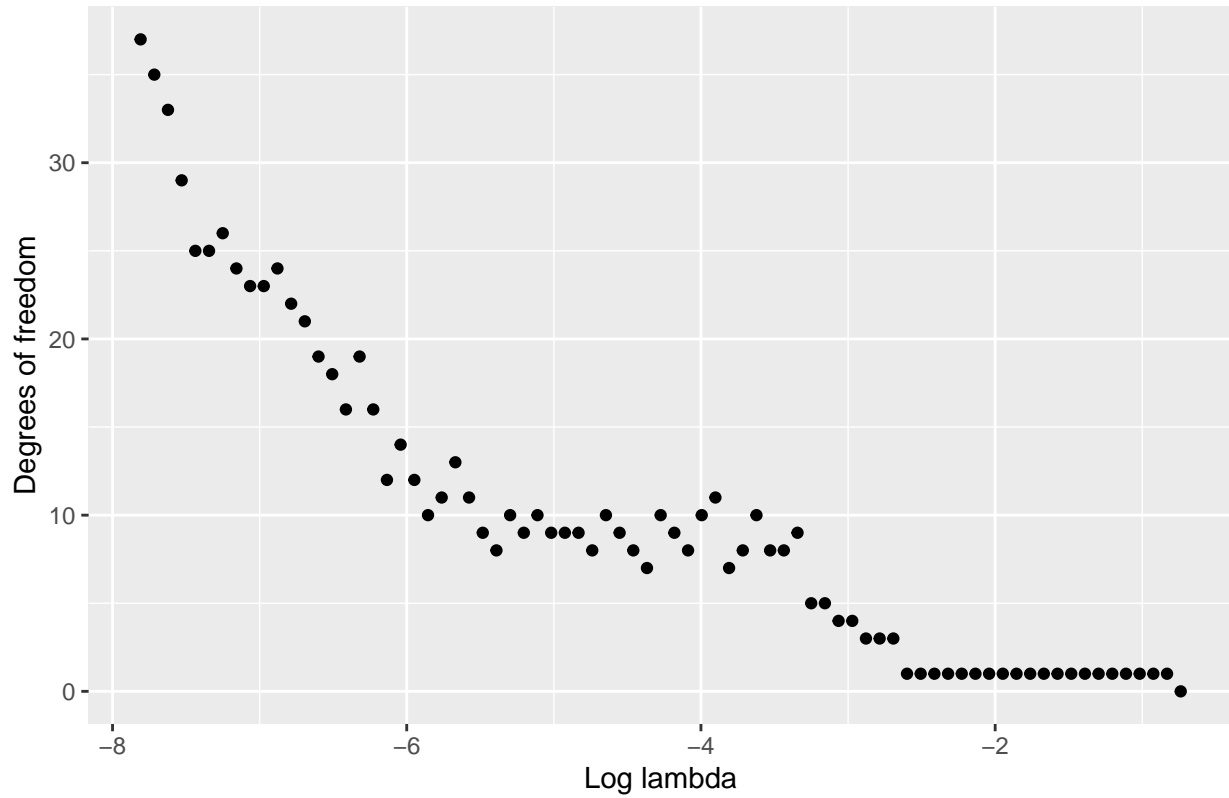
```
## [1] -2.688248
```



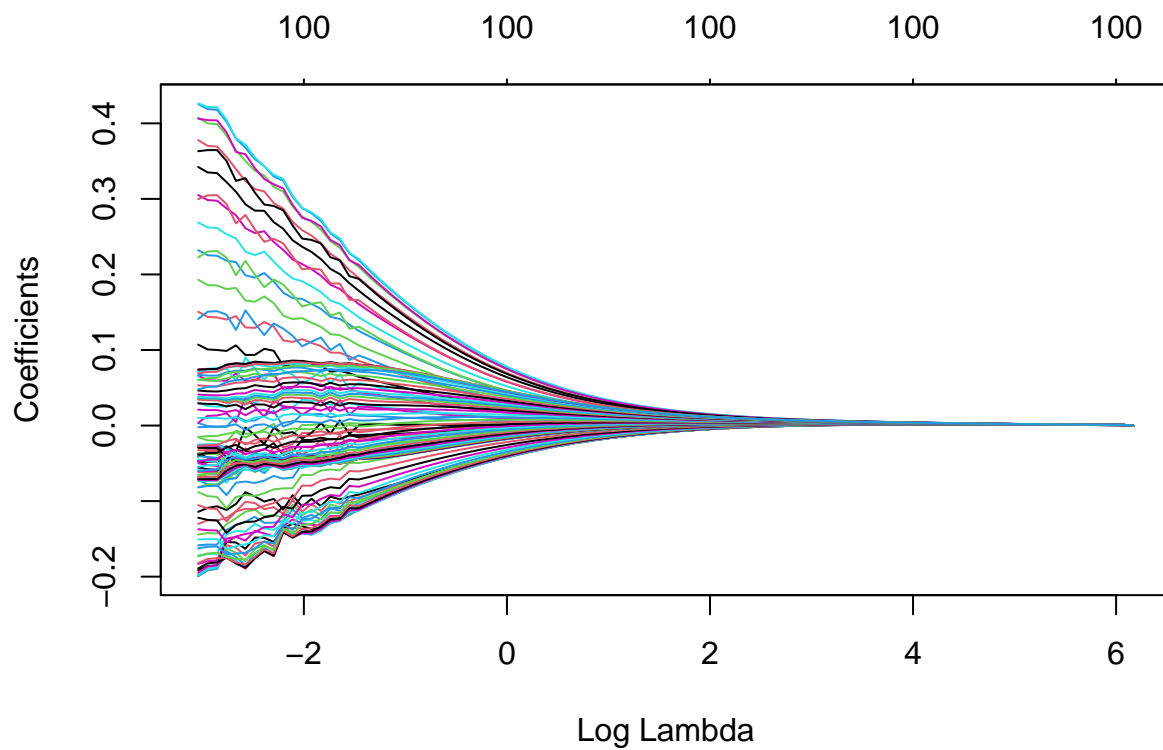
From the plot we can see that where the function converges to 3 features that have non zero coefficients is approximately at  $e^{-2.6}$  and it was numerically shown that at  $\log(\lambda) = -2.688248$  the function would have a df of 3 and 3 features with non zero coefficients.

## 3.4

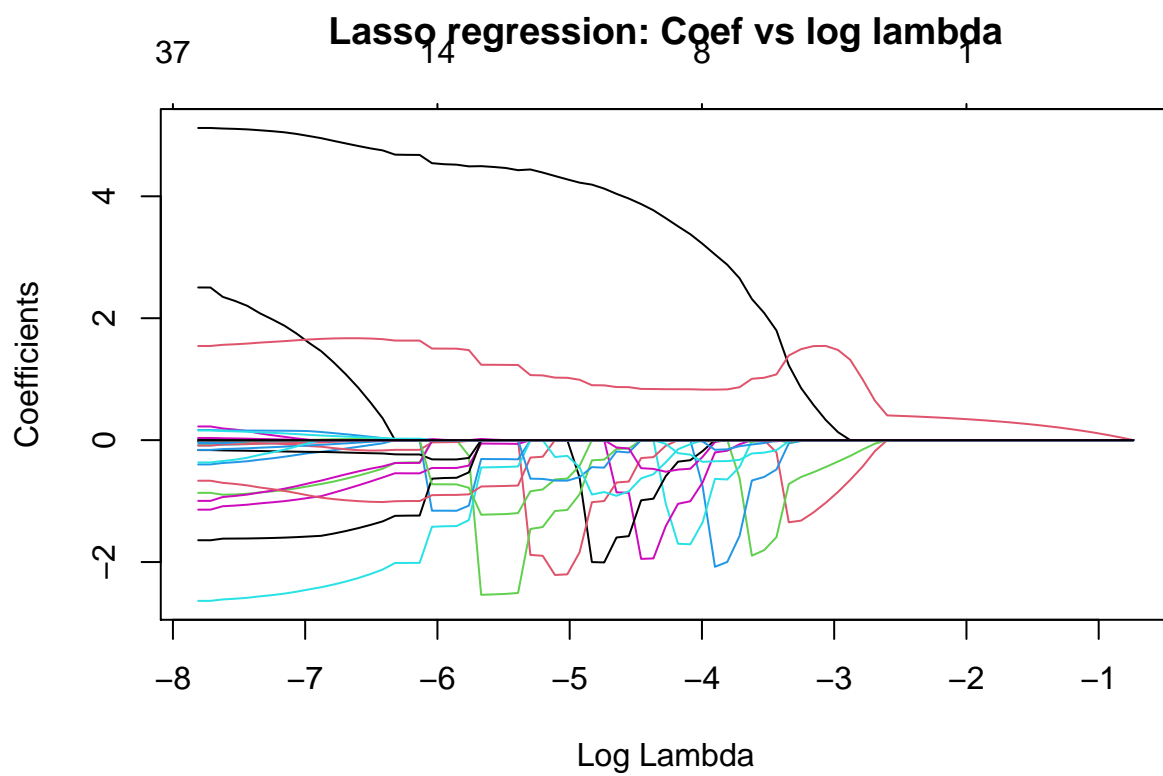
Degrees of freedom vs Log lambda

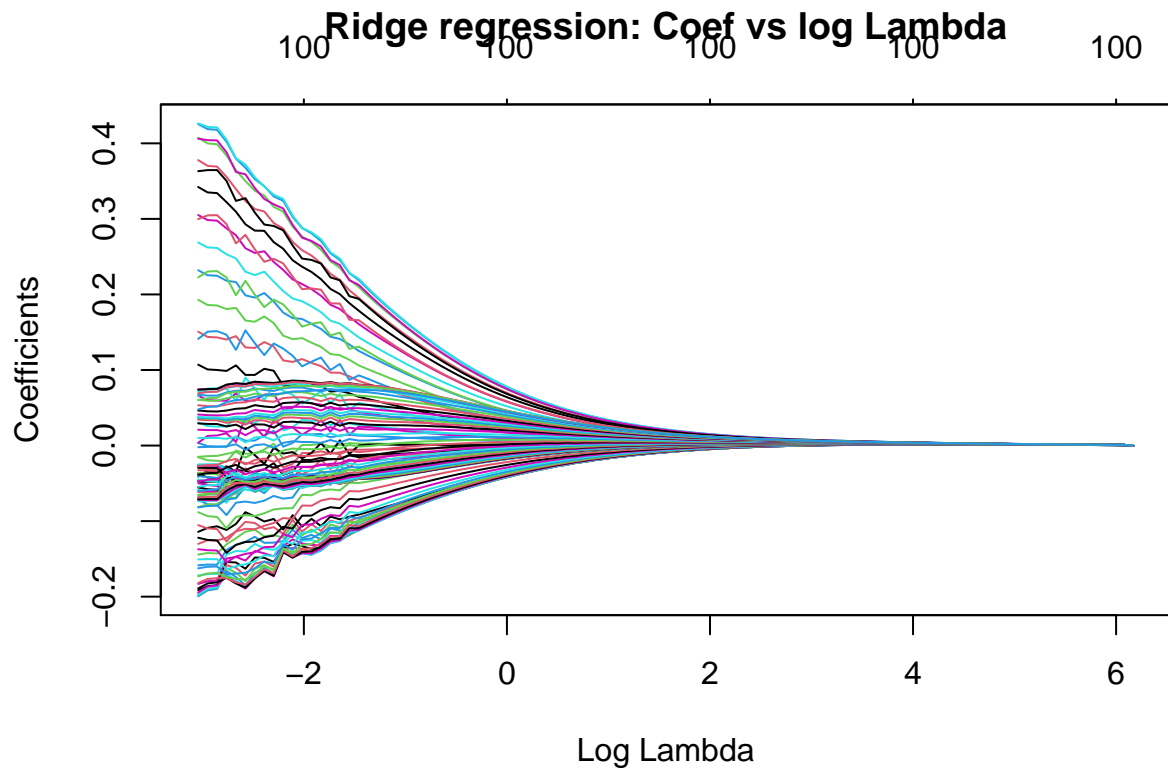


The graph shows that as the value of the shrinkage parameter  $\lambda$  increases it reduces the number of features in the model. Which is exactly as expected from the given function with an sufficiently large  $\lambda$  it will reduce the number of features included to 0 as such decreasing the degrees of freedom.



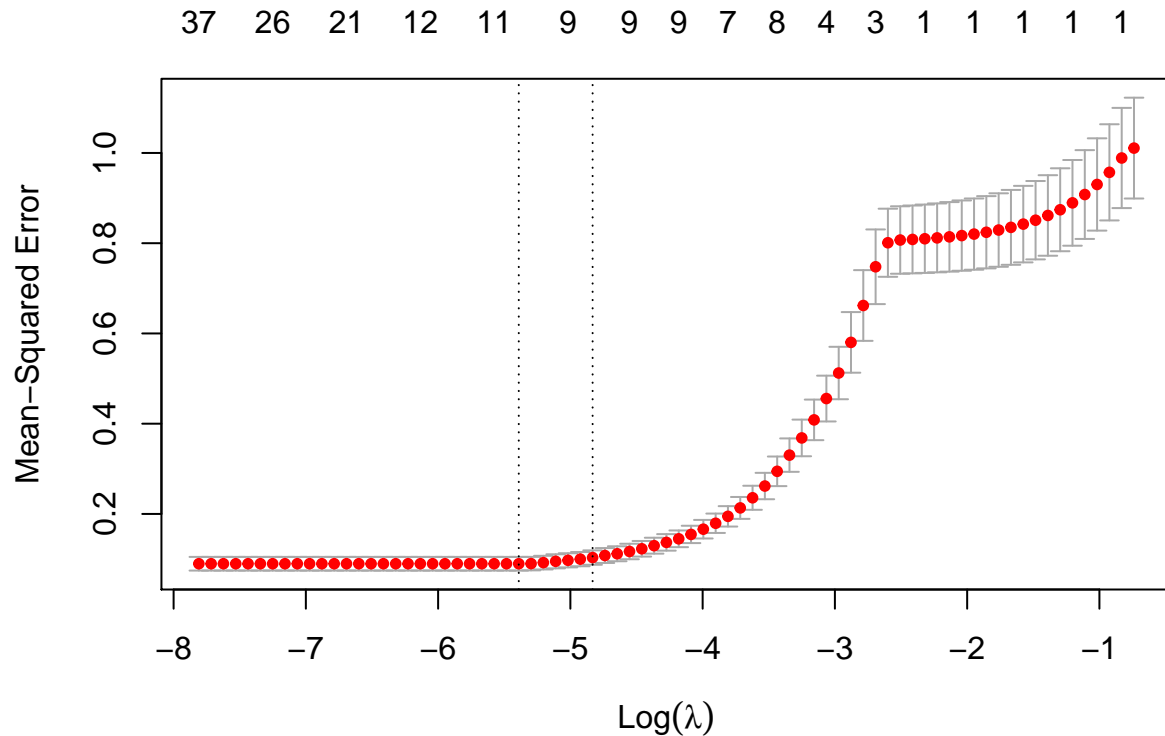
## [1] -2.688248





As seen in the graphs the lasso function will converge the coefficients of the model to 0 whereas the ridge regression model will reduce the coefficients close to zero but never to zero.

```
## [1] 0.004561105
```



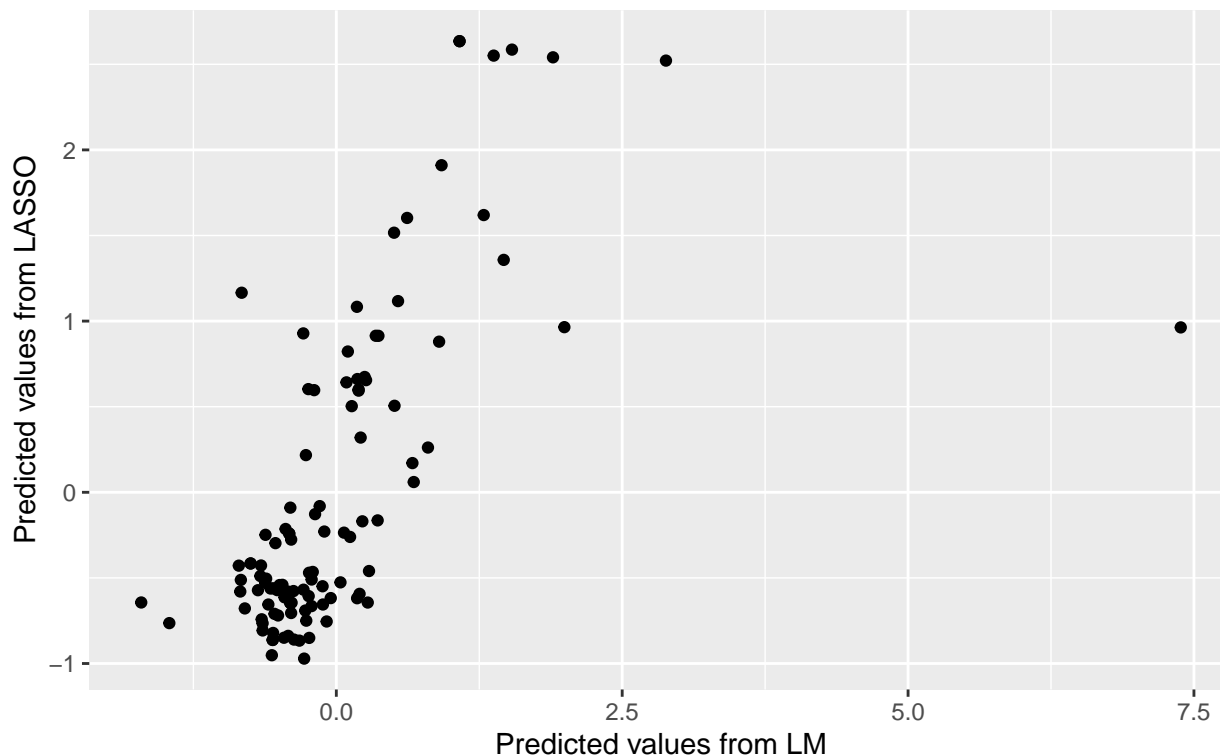
The graph shows that the lowest MSE is found at  $\lambda = 0.004561105$  which is at  $\log(\lambda) = -5.39019$  as shown in the graph. As the  $\log(\lambda)$  increases the MSE will increase with a linear rate from -8 to approx -5 and then

with a non-linear increase until approximately -2.8 and then where it will slowdown the increasing rate of MSE.

```
## [1] 7
```

```
## [1] 1
```

Comparison of Predicted values with LASSO optimal vs  
predicted values from linear regression

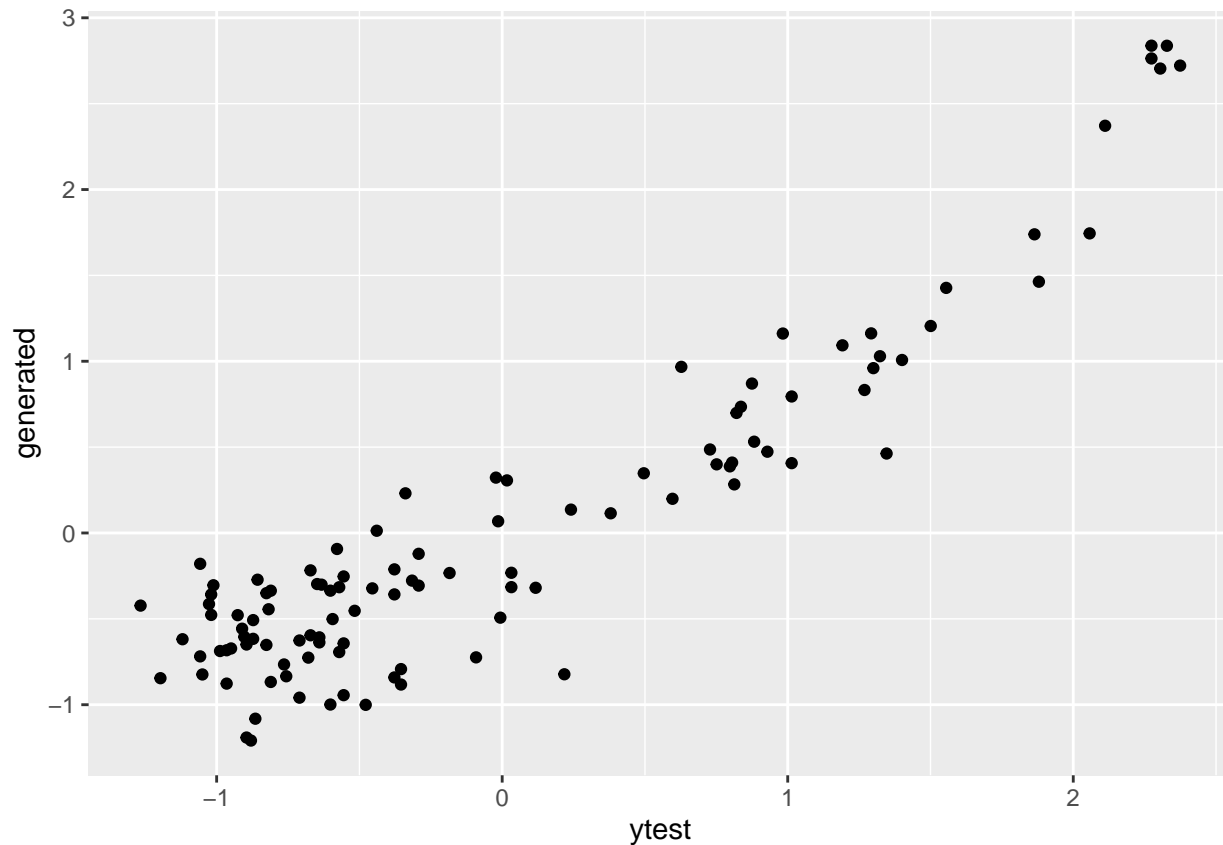


### 3.7

```
preds <- predict(lasso3_6_opt, test_scale[, -101])

sigma <- sqrt(mean((train_scale[, 101] - predict(lasso3_6_opt, train_scale[, -101]))^2))
new_targets <- rnorm(n = nrow(test_scale), preds, sigma)

generated_data <- data.frame("ytest" = test_scale[, 101], "generated" = new_targets)
ggplot(generated_data, aes(x = ytest, y = generated)) + geom_point()
```



## Appendix: Code

```
library(readr)
library(kknn)
library(ggplot2)
library(data.table)
library(glmnet)

#Assignment 1
#1.1
digits <- read.csv("optdigits.csv", header = F)
colnames(digits)[ncol(digits)] <- "number"
digits$number <- as.factor(digits$number)
n <- dim(digits)[1]
set.seed(12345)
id <- sample(1:n, floor(n*0.5))
train <- digits[id,]
id1 <- setdiff(1:n, id)
set.seed(12345)
id2 <- sample(id1, floor(n*0.25))
validation <- digits[id2,]
id3 <- setdiff(id1, id2)
test <- digits[id3,]

#1.2
knn_train <- kknn(number ~., train, train, k = 30, kernel = "rectangular")
```



```

knn_test <- kknn(number ~., train, test, k = 30, kernel = "rectangular")
#Confusion matrix
pred_train<- fitted(knn_train)
confusion_train <- table(train$number, pred_train)
pred_test <- fitted(knn_test)
confusion_test <- table(test$number, pred_test)
print(confusion_train)
print(confusion_test)
print(diag(confusion_train)/rowSums(confusion_train))
print(diag(confusion_test)/rowSums(confusion_test))
#Misclassification errors
missclass <- function(X,X1) {
  n = length(X)
  return(1-sum(diag(table(X,X1)))/n)
}
mismatch_rate_train <- missclass(train$number, pred_train) #0.04500262
mismatch_rate_test <- missclass(test$number, pred_test) #0.05329154
cat("The mismatch rate for train data is:", mismatch_rate_train,
    "\nThe mismatch rate for test data is:", mismatch_rate_test)

#1.3
train_eight <- train[which(train$number==8),]
train_eight$prob <- knn_train$prob[which(train$number==8), "8"]
sortedResult <- sort(train_eight[, "prob"], index.return = T)
sortedResult$x[1:3] #0.1000000 0.1333333 0.1666667
hard_ids <- sortedResult$ix[1:3] #50 43 136
tail(sortedResult$x, n=c(2)) #1 1
easy_ids <- tail(sortedResult$ix, n=c(2)) #179 183
plotHeatmap <- function(index){
  heatmap(matrix(as.numeric(train_eight[index, 1:64])), 8, 8, byrow = T), Colv = NA, Rowv = NA)
}
for (i in c(hard_ids, easy_ids)) {
  plotHeatmap(i)
}

#1.4
e_t <- 0
e_v <- 0
for (k in 1:30) {
  kt <- kknn(number ~., train, train, k = k, kernel = "rectangular")
  e_t[k] <- missclass(train$number, fitted(kt))
  kv <- kknn(number ~., train, validation, k = k, kernel = "rectangular")
  e_v[k] <- missclass(validation$number, fitted(kv))
}
plot(1:30, e_v, ylim = c(0,0.06), type = "l", col = "red", xlab = "K", ylab = "miss-classification error")
lines(1:30, e_t, ylim = c(0,0.06), col = "blue")
legend(2, 0.06, legend=c("validation", "train"),
      col=c("red", "blue"), lty=1:1, cex=0.8)

k_test_optik <- kknn(number ~., train, test, k = 4, kernel = "rectangular")
cat("Test mis-classification rate when k=4 is:", missclass(test$number, fitted(k_test_optik)),
    "\nTrain mis-classification rate when k=4 is:", e_t[4],
    "\nValidation mis-classification rate when k=4 is:", e_v[4])

```

```

#1.5
r_v <- 0
for (k in 1:30) {
  kv <- kknncv(number ~., train, validation, k = k, kernel = "rectangular")
  s <- 0
  for (i in 0:9) {
    s <- s + sum(log(kv$prob[which(validation$number==i), toString(i)] + 1e-15)/nrow(kv$prob))
  }
  r_v[k] <- -s
}
plot(1:30, r_v, type = "l", xlab = "K", ylab = "mean cross entropy")

### Assignment 2
### Question 2

# read in data
parkinsons <- read_csv("parkinsons.csv", col_types = cols())

# drop columns not to be used for prediction
parkinsons <- parkinsons[, -c(1, 2, 3, 4, 6)]

# bayesian model

##  $y | X, w_0, w \sim N(w_0 + Xw, \sigma^2 \cdot I)$ 
##  $w \sim N(0, \sigma^2 / \lambda \cdot I)$ 

# create train/test partitions
all_indices <- 1:nrow(parkinsons)
set.seed(12345)
train_indices <- sample(all_indices, ceiling(0.6*nrow(parkinsons)))
test_indices <- all_indices[!(all_indices %in% train_indices)]
train_data <- parkinsons[train_indices, ]
test_data <- parkinsons[test_indices, ]

# scale data -- see https://sebastianraschka.com/faq/docs/scale-training-test.html
# for why training mu/sigma are used for scaling test data
n_col <- ncol(parkinsons)
train_mu <- sapply(1:n_col, function(x) mean(unlist(train_data[x])))
train_sigma <- sapply(1:n_col, function(x) sd(unlist(train_data[x])))

train_data <- as.matrix(as.data.frame(
  sapply(1:n_col, function(x) (train_data[x] - train_mu[x])/train_sigma[x])))
# as.matrix has no effect without as.data.frame!!!
test_data <- as.matrix(as.data.frame(
  sapply(1:n_col, function(x) (test_data[x] - train_mu[x])/train_sigma[x])))

# set up helper functions

model_log_likelihood <- function(w_vector, sigma, X) {
  # built for no intercept models, as asked

```

```

n <- nrow(X)
term1 <- -n/2*log(2*pi*sigma^2)
term2 <- -sum(sapply(1:n, function(x) X[x, 1] - X[x, -1] %*% w_vector)^2)/2/sigma^2
# assumes y = X[, 1]

return(term1 + term2)
}

ridge <- function(w_vector_sigma, X, lambda) {
  # built for no intercept models, as asked

  n <- ncol(X)
  sigma <- w_vector_sigma[n]
  w_vector <- w_vector_sigma[1:(n-1)]
  return(lambda*sum(w_vector^2) - model_log_likelihood(w_vector, sigma, X))
}

ridge_opt <- function(X, lambda) {

  n <- ncol(X)
  par <- rnorm(n-1) # initial parameter values (to seed optimisation); random
  par[n] <- 1 # initialise with positive value
  return(optim(par = par, fn = ridge, method = "BFGS", X = X, lambda = lambda)$par)
}

degrees_of_freedom <- function(X, lambda) {

  mat <- tcrossprod(X %*% solve(crossprod(X) + diag(lambda, ncol(X), ncol(X))), X)

  return(sum(diag(mat)))
}

MSE <- function(X, optimal_weights) { # based on definition here:
  # https://en.wikipedia.org/wiki/Mean_squared_error

  n <- nrow(X)
  return(sum((X[, 1] - X[, -1] %*% optimal_weights)^2)/n)
  # built for no-intercept models as asked
}

# model 1: lambda = 1
modell1_optimal_weights_sigma <- ridge_opt(train_data, lambda = 1)
modell1_optimal_weights <- modell1_optimal_weights_sigma[1:(n_col - 1)]
modell1_optimal_sigma <- modell1_optimal_weights_sigma[n_col]
modell1_train_MSE <- MSE(train_data, modell1_optimal_weights)
modell1_test_MSE <- MSE(test_data, modell1_optimal_weights)
modell1_AIC <- -2*model_log_likelihood(
  modell1_optimal_weights, modell1_optimal_sigma, as.matrix(train_data)) +

```

```

2*degrees_of_freedom(as.matrix(train_data), lambda = 1)

# model 2: lambda = 100
model2_optimal_weights_sigma <- ridge_opt(train_data, lambda = 100)
model2_optimal_weights <- model2_optimal_weights_sigma[1:(n_col - 1)]
model2_optimal_sigma <- model2_optimal_weights_sigma[n_col]
model2_train_MSE <- MSE(train_data, model2_optimal_weights)
model2_test_MSE <- MSE(test_data, model2_optimal_weights)
model2_AIC <- -2*model_log_likelihood(
  model2_optimal_weights, model2_optimal_sigma, as.matrix(train_data)) +
  2*degrees_of_freedom(as.matrix(train_data), lambda = 100)

# model 3: lambda = 1000
model3_optimal_weights_sigma <- ridge_opt(train_data, lambda = 1000)
model3_optimal_weights <- model3_optimal_weights_sigma[1:(n_col - 1)]
model3_optimal_sigma <- model3_optimal_weights_sigma[n_col]
model3_train_MSE <- MSE(train_data, model3_optimal_weights)
model3_test_MSE <- MSE(test_data, model3_optimal_weights)
model3_AIC <- -2*model_log_likelihood(
  model3_optimal_weights, model3_optimal_sigma, as.matrix(train_data)) +
  2*degrees_of_freedom(as.matrix(train_data), lambda = 1000)

# plot performance
MSE_data <- data.frame(log_10_lambda = c(0, 0, 2, 2, 3, 3),
  Legend = rep(c("Training", "Test"), 3),
  value = c(model1_train_MSE, model1_test_MSE, model2_train_MSE,
    model2_test_MSE, model3_train_MSE, model3_test_MSE))

ggplot(MSE_data) + geom_line(aes(log_10_lambda, value, colour = Legend)) + theme_bw() +
  geom_vline(xintercept = 2, linetype = "dotted") +
  xlab("Log (base-10) of Hyper-Parameter 'lambda'") + ylab("Mean Squared Error") +
  ggtitle("Finding the Optimal Hyper-Parameter")

AIC_data <- data.frame(log_10_lambda = c(0, 2, 3),
  value = c(model1_AIC, model2_AIC, model3_AIC))

ggplot(AIC_data) + geom_line(aes(log_10_lambda, value)) + theme_bw() +
  geom_vline(xintercept = 2, linetype = "dotted") +
  xlab("Log (base-10) of Hyper-Parameter 'lambda'") + ylab("Model AIC") +
  ggtitle("Finding the Optimal Hyper-Parameter")

```