Exercise 2

Damian Ke & Kyriakos Papadopoulos

2022-11-29

Assignment 1. Explicit regularization

Question 1: Assume that Fat can be modeled as a linear regression in which absorbance characteristics (Channels) are used as features. Report the underlying probabilistic model, fit the linear regression to the training data and estimate the training and test errors. Comment on the quality of fit and prediction and therefore on the quality of model.

Answer 1:

We chose to use **MSE** as our error function because we have a regression problem.

```
## The mse for training data is:  0.005709117
```

```
## The mse for test data is:  722.4294
```

We can clearly see that there is a big difference between training and test errors. The training error is much smaller compared to the test error. That shows that our model is very complicated and it overfits.

The probabilistic model for linear regression is given by the formula:

$$y|x \ N(\theta^T x, \sigma^2)$$

which is equal if we write it like as equation:

$$y = \theta_0 + \theta + 1x_1 + ...\theta_p x_p + \epsilon$$

The ϵ represents the noise term and accounts for random errors in the data not captured by the model. The noise is assumed to have mean zero and to be independent of x. The zero-mean assumption is nonrestrictive, since any (constant) non-zero mean can be incorporated in the offset term θ_0

Question 2: Assume now that Fat can be modeled as a LASSO regression in which all Channels are used as features. Report the cost function that should be optimized in this scenario.

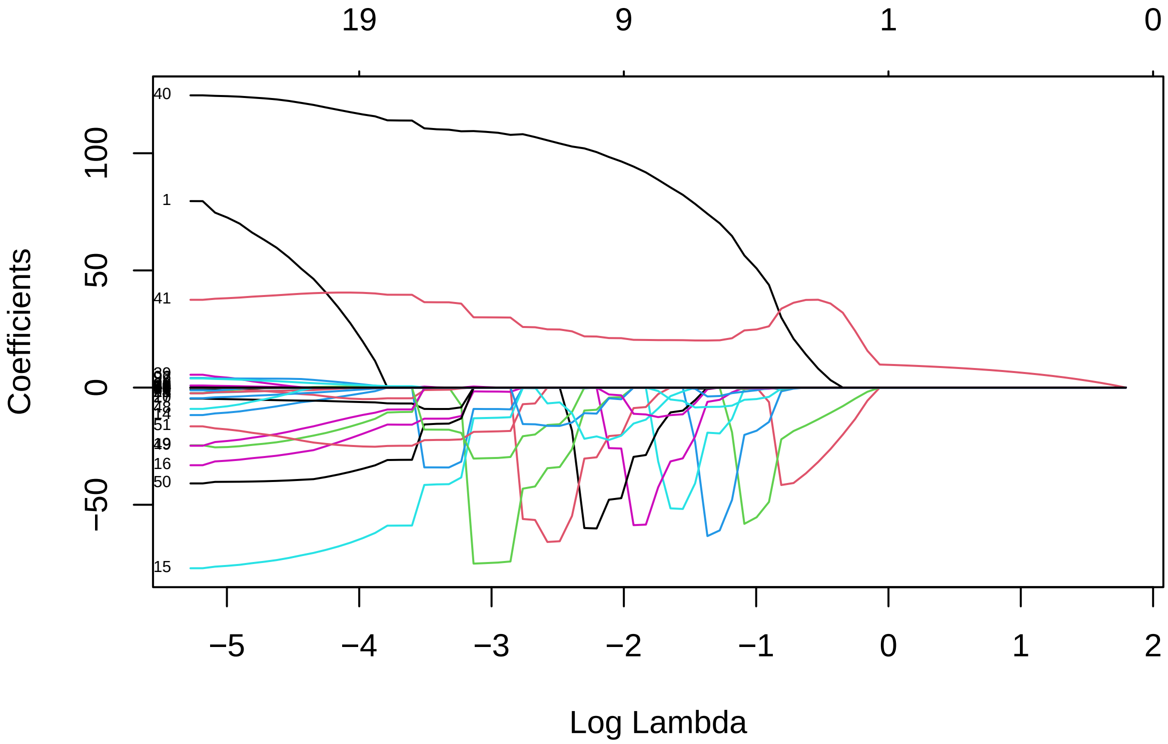
Answer 2: The cost function that we are gonna use for LASSO regression according to the slides:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta 0 - \theta_1 x_1 j - ... - \theta_p x_p j)^2 + \lambda \sum_{j=1}^{p} |\theta_i|$$

Where $p = 100$ in our case.

Question 3: Fit the LASSO regression model to the training data. Present a plot illustrating how the regression coefficients depend on the log of penalty factor (log λ) and interpret this plot. What value of the penalty factor can be chosen if we want to select a model with only three features?

Answer 3:

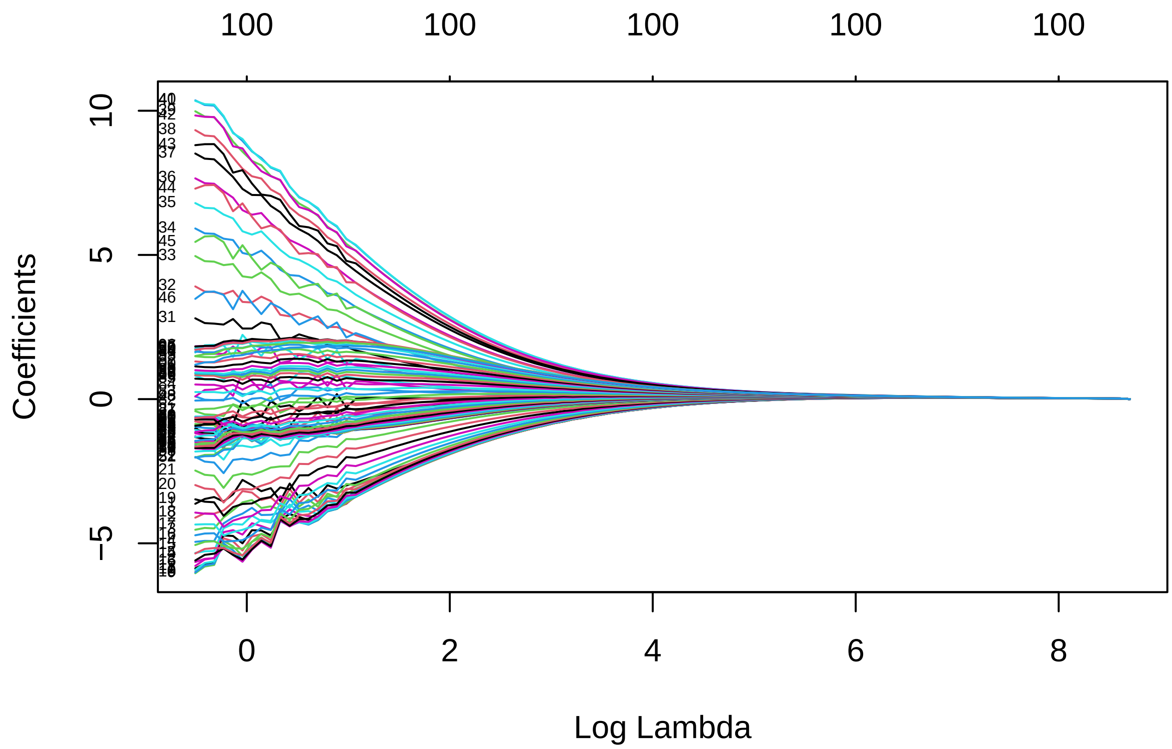



```
## The values of lambda that we get 3 features are:  0.8530452 0.777263 0.7082131
```

We can also do it visually by going to the above plot and find the value of lambda that only 3 lines are != 0.

Question 4: Repeat step 3 but fit Ridge instead of the LASSO regression and compare the plots from steps 3 and 4. Conclusions?

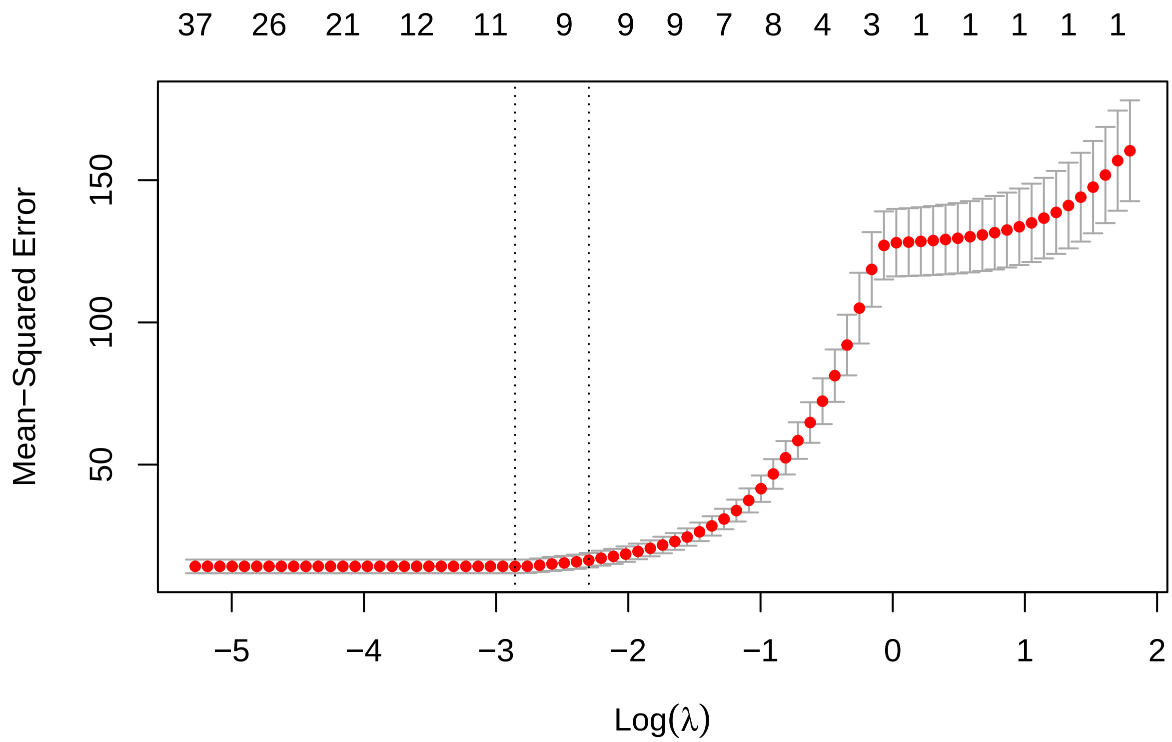
Answer 4:



We can see that for the same value of lambda LASSO gets rid of much more coefficients compared to Ridge. That's because LASSO takes the magnitude of the coefficients and ridge takes the square.

Question 5: Use cross-validation with default number of folds to compute the optimal LASSO model. Present a plot showing the dependence of the CV score on log λ and comment how the CV score changes with log λ. Report the optimal λ and how many variables were chosen in this model. Does the information displayed in the plot suggests that the optimal λ value results in a statistically significantly better prediction than log λ = -4? Finally, create a scatter plot of the original test versus predicted test values for the model corresponding to optimal lambda and comment whether the model predictions are good

Answer 5:



We can see that the cv score is stable until the optimal lambda(in that point it gets it's minimum value). After the optimal lambda the error starts to be increased.

We can see that the confidence intervals are equal so for the optimal λ it's not significantly better in predictions compared to $\lambda = log(-4)$. The optimal λ is where is the first dotted line in the plot.

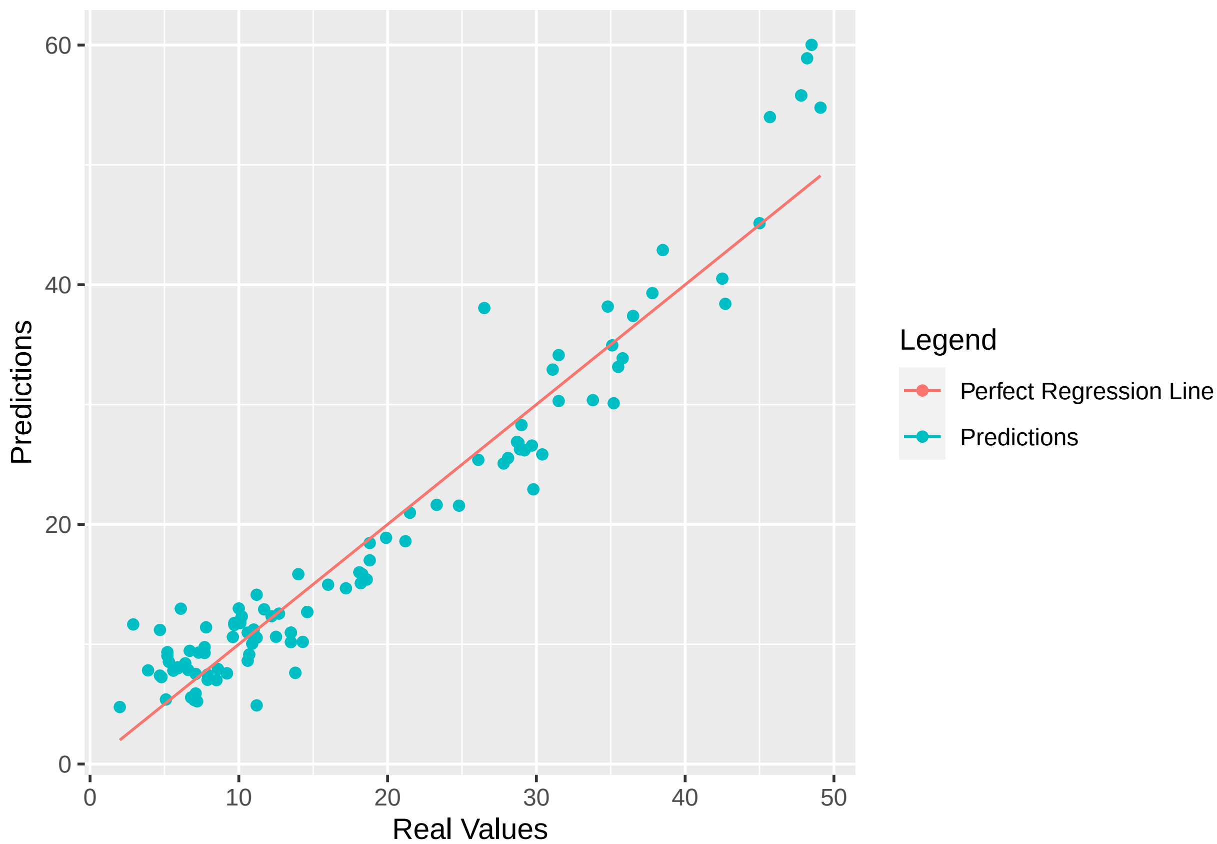
```
## 100 x 1 sparse Matrix of class "dgCMatrix"
##                    s0
## Channel1     .
## Channel2     .
## Channel3     .
## Channel4     .
## Channel5     .
## Channel6     .
## Channel7     .
## Channel8     .
## Channel9     .
## Channel10    .
## Channel11    .
## Channel12    .
## Channel13  -11.85508
## Channel14  -44.98479
## Channel15  -27.78654
## Channel16  -11.93569
## Channel17    .
## Channel18    .
```

```
## Channel19     .
## Channel20     .
## Channel21     .
## Channel22     .
## Channel23     .
## Channel24     .
## Channel25     .
## Channel26     .
## Channel27     .
## Channel28     .
## Channel29     .
## Channel30     .
## Channel31     .
## Channel32     .
## Channel33     .
## Channel34     .
## Channel35     .
## Channel36     .
## Channel37     .
## Channel38     .
## Channel39     .
## Channel40  123.14896
## Channel41   17.81151
## Channel42     .
## Channel43     .
## Channel44     .
## Channel45     .
## Channel46     .
## Channel47     .
## Channel48     .
## Channel49     .
## Channel50     .
## Channel51  -52.09615
## Channel52     .
## Channel53     .
## Channel54     .
## Channel55     .
## Channel56     .
## Channel57     .
## Channel58     .
## Channel59     .
## Channel60     .
## Channel61     .
## Channel62     .
## Channel63     .
## Channel64     .
## Channel65     .
## Channel66     .
## Channel67     .
## Channel68     .
## Channel69     .
## Channel70     .
## Channel71     .
## Channel72     .
```

```
## Channel73     .
## Channel74     .
## Channel75     .
## Channel76     .
## Channel77     .
## Channel78     .
## Channel79     .
## Channel80     .
## Channel81     .
## Channel82     .
## Channel83     .
## Channel84     .
## Channel85     .
## Channel86     .
## Channel87     .
## Channel88     .
## Channel89     .
## Channel90     .
## Channel91     .
## Channel92     .
## Channel93     .
## Channel94     .
## Channel95     .
## Channel96     .
## Channel97     .
## Channel98     .
## Channel99     .
## Channel100    .
```

We can that the model with the optimal λ uses in total 7 features. These are:

- Channel13
- Channel14
- Channel15
- Channel16
- Channel40
- Channel41
- Channel52



The red line represents a perfect model and the blue dots are our predictions. We can see that our model is not perfect but it's predictions are close to the real values.