**Data 558: Statistical Machine Learning**

**Spring 2023**

**Homework – 4**

**Arjun Sharma**

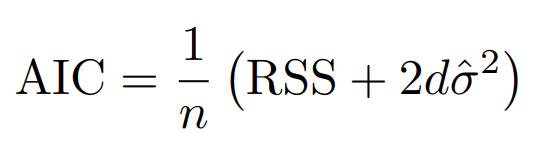
**Conceptual Questions:**

1. Describe the different model selection criterions AIC, BIC, and adjusted  
   R2, and advantages of using one over the other. Here, think about factors such as being able to explain it to general audience, obtaining simple models, theoretical support, and general applicability.

Solution:

AIC:

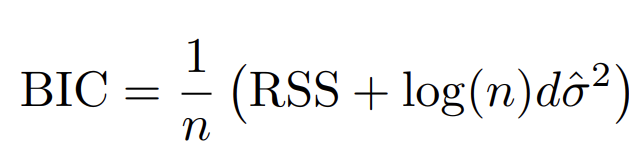
Akaike Information Criterion is a measure used to compare different models fitted onto the same dataset. Mathematically, it is calculated as:



It is evident that a lower value of the AIC represents a ‘better’ model. As in, a lower AIC value represents a balance between the model complexity and the fit of the model. A higher value of the AIC would imply that the model is potentially too complex for the data. The AIC is a value that isn’t absolute, in the sense that it would help one obtain insight as to which model fitted for a particular dataset has been fit best, relative to other models fitted on the same data.

BIC:

Bayesian Information Criterion is a measure, like the AIC used to compare different models fitted on to the same dataset. Mathematically, it is calculated as:

**

Like AIC, BIC is also interpreted as “the lower, the better”. However, BIC tends to penalize higher complexity models more harshly when compared to the AIC. In addition, it appears as though a large sample size would also lead to further penalization of the model. Like the AIC, the BIC also gives you a value that gives you insights on how each model compares to others fitted on the same data.

Adjusted R-squared:

The adjusted R-squared value is an absolute metric, which typically ranges from 0 to 1. Mathematically, Adjusted R-squared is represented as:

Adjusted R-squared = 1 – [(1-R2)(n-1)/(n-k-1)]

Where R2 = 1 – (RSS/TSS)

The adjusted R-squared is preferred over the R-squared metric since it takes the dimensionality of the model into account and penalizes the model for excessive features. The Adjusted R-squared will always take on values less than or equal to R2. It is interpreted as the percentage of variance in the target that is explained by the set of predictors. For instance, Adjusted R-squared of 0.75 would mean that the model explains 75% of the variance in the target. What makes a good Adjusted R-squared value is context dependent, but the interpretation of the metric is objective, unlike AIC and BIC.

AIC vs Adjusted R-squared:

Points in favor of AIC:

* AIC has rigorous theoretical justifications which rely on asymptotic arguments (scenarios where the sample size n is very large). Despite its popularity, and even though it is quite intuitive, the adjusted R2 is not as well motivated in statistical theory as AIC.
* AIC can be used for higher complexity models. Since R-squared is the square of the correlation coefficient which can only capture linear relationships, The Adjusted R-squared may not wholly capture higher degree polynomial relationships.

Points in favor of Adjusted R-squared:

* Adjusted R-squared is objective, and easy to interpret.
* It is also easier to explain to the general audience who may not have much domain knowledge in statistics.

AIC vs BIC:

Points in Favor of AIC:

* It is relatively ‘relaxed’ in terms of penalizing model complexity. Therefore, it also allows for higher complexity models to be included in comparison with lower complexity models, when the model fit is concerned.
* It places relatively higher weightage on the goodness of fit of the model. It is favorable in situations where we are more concerned about the fit of the model.
* If we aren’t concerned about sample size, the AIC again is an appropriate metric.

Points in favor of BIC:

* BIC is appropriate in a situation where we might not have much data at our disposal and do not wish to use models of a higher complexity.
* It heavily penalizes higher complexity models, and hence is appropriate when we do not want a complex model.

BIC vs Adjusted R-squared:

Points in favor of BIC:

* BIC has rigorous theoretical justifications which rely on asymptotic arguments (scenarios where the sample size n is very large). Despite its popularity, and even though it is quite intuitive, the adjusted R2 is not as well motivated in statistical theory as BIC.
* BIC can be used for higher complexity models. Since R-squared is the square of the correlation coefficient which can only capture linear relationships, The Adjusted R-squared may not wholly capture higher degree polynomial relationships.

Points in favor of Adjusted R-squared:

* Adjusted R-squared is objective, and easy to interpret.
* It is also easier to explain to the general audience who may not have much domain knowledge in statistics.

1. Please determine whether following practices are good or not. Please  
   explain why in either case. [Grading criterion: each question subpart is worth 4 points with 1 point assigned for the answer and 3 points for the explanation.]  
   (a) We want to model a housing dataset consisting of 500 features. We decide to use Ridge  
   Regression (ℓ2 regularization) in order to better obtain more prominent features and how  
   they determine a house’s price.  
   (b) We are aware of the time cost of finding the globally optimal subset of features for a prediction task, so we resort to greedy algorithms such as forward step-wise selection. We think that the run time will be significantly reduced, even though the solution may not be optimal.  
   (c) We notice that for a certain value of λ in the LASSO solution path, the training MSE is  
   almost 0 but the cross-validation MSE is 100. To fix this, we choose a lower value of the  
   regularization parameter λ.  
   (d) Suppose we are consulting for a company that wants to know which set of features is relevant for predicting house sales. We know that some of the predictors are highly correlated. We run the LASSO and choose the regularization parameter λ via the validation set approach. We are happy because the LASSO solution at this value of λ yields only 2 non-zero coefficient predictors out of the possible 100. We go and report it to the company and tell them that these 2 predictors are relevant, while the others are irrelevant for prediction because their estimated coefficients are zero.

Solution:

(a) This is bad practice. We have 500 features in this dataset, and the goal is to obtain more prominent features and understand how they determine house price. Ridge regression does not eliminate any features, the coefficients for all features will be > 0 and hence, the process of choosing the more prominent predictors out of the said 500 will be cumbersome. Furthermore, it is practically a reasonable assumption that not all of the 500 features are important, and that it would be computationally inexpensive to remove some of the less important features. To do this while retaining the ability to clearly comprehend the impact of each feature, lasso regression (l1 regularization) is a more appropriate approach. If the focus was purely on dimensionality reduction, I would have suggested using PCA, however that would compromise on the goal to interpret and explain the impact of each feature on the model behavior. Hence, Lasso Regression is suggested as an alternative.

(b) If we prefer simplicity in implementation and reduced computational power and time over the goodness of fit of the model – this is considered good practise. Forward step-wise selection allows us to choose each feature sequentially after assessment of each feature’s performance relative to others. In this case, we obtain a reasonably well fit model while using a fraction of the computational resources and time.

If the goal was to obtain an optimal model with the best combination of parameters irrespective of time and computational resources, then this would be considered bad practise.

(c) This is very bad practise. If the training MSE is almost 0, and the testing MSE is 100, it means that the model is overfitting to the training set – this point is strengthened with the knowledge that the said metrics were obtained after cross validation. The model thus has high variance. Reducing the value of λ would lead to higher variance since the bias term would decrease. This would exacerbate the overfitting problem. If anything, the λ value must be increased in order to regularize the model by decreasing the variance as a result of increased bias. Alternatively, it is possible that the model is too complex for the data and is overfitting to the noise, the user may also use a model with lower complexity.

(d) This is bad practice. Prior to performing LASSO regression, there should have been some emphasis on removing redundant parameters, particularly by looking at the correlated features. After which, the decision to remove some of the features must have been made. In addition, coefficients being reduced to zero is not necessarily conducive to retaining only the important features. It is possible that the scale of different features is also different, and hence the smaller coefficient for one feature is a result of the scale of the concerned feature being much higher than that of other features, in which case, standardization or normalization of the features should be performed (this hasn’t been mentioned as part of the procedure taken). Furthermore, the holdout validation process has been used – the inference of relevant features has been made through only one set of train and test sets.

The problem here is that we are using correlated features, which could cause redundancy, then only use the holdout validation approach. We also do not know if the data has been standardized.

It is advisable to use other methods to split the data such as k-fold cross validation or HOOCV to have multiple sets of validation and training sets, this would give us insight that is more generalized and less impacted by noise or bias. Furthermore, it is advisable to look at other processes to perform feature selection aside from just lasso. Also, it is important to standardize or normalize the data since Lasso regression is scale dependent.

1. Is there always a unique solution, bβ to ordinary least squares? If not,  
   come up with two scenarios where there may not be a unique solution, i.e., you can obtain bβ1 and bβ2, where both estimates minimize the RSS, but with bβ1 ̸ = bβ2. What about for Ridge regression with λ > 0? Why or why not?

Solution:

Ordinary least Squares does not have unique solution for the following situations:

a. When p>n; when the number of predictors is greater than the number of training examples.

b. When multicollinearity is observed in the features.

In the case of Ridge Regression, for λ > 0, there is always a unique solution.

The addition of the penalty term λ∑ βi2­ where I = 1,2,3,…p, to the covariance matrix ensures that all eigenvalues are greater than 0 and hence it becomes invertible. Invertibility necessitates unique solution, and hence, in the case of ridge regression for λ > 0, there is always a unique solution

4. (BONUS Question)  
(a)

One useful heuristic for stopping could be to monitor the BIC of the model as the predictors are added. If the BIC score increases over a certain threshold, we can stop adding predictors to the model since the BIC penalizes more complex models very heavily.

Hypothetically, consider a threshold like; if the BIC rises by over 10% over 2 consecutive feature additions.

(b) For the full model, the BIC can be intuitively assumed high, because of the large number of features, and because the question says that only a few are relevant. Therefore we can theoretically expect to stop when

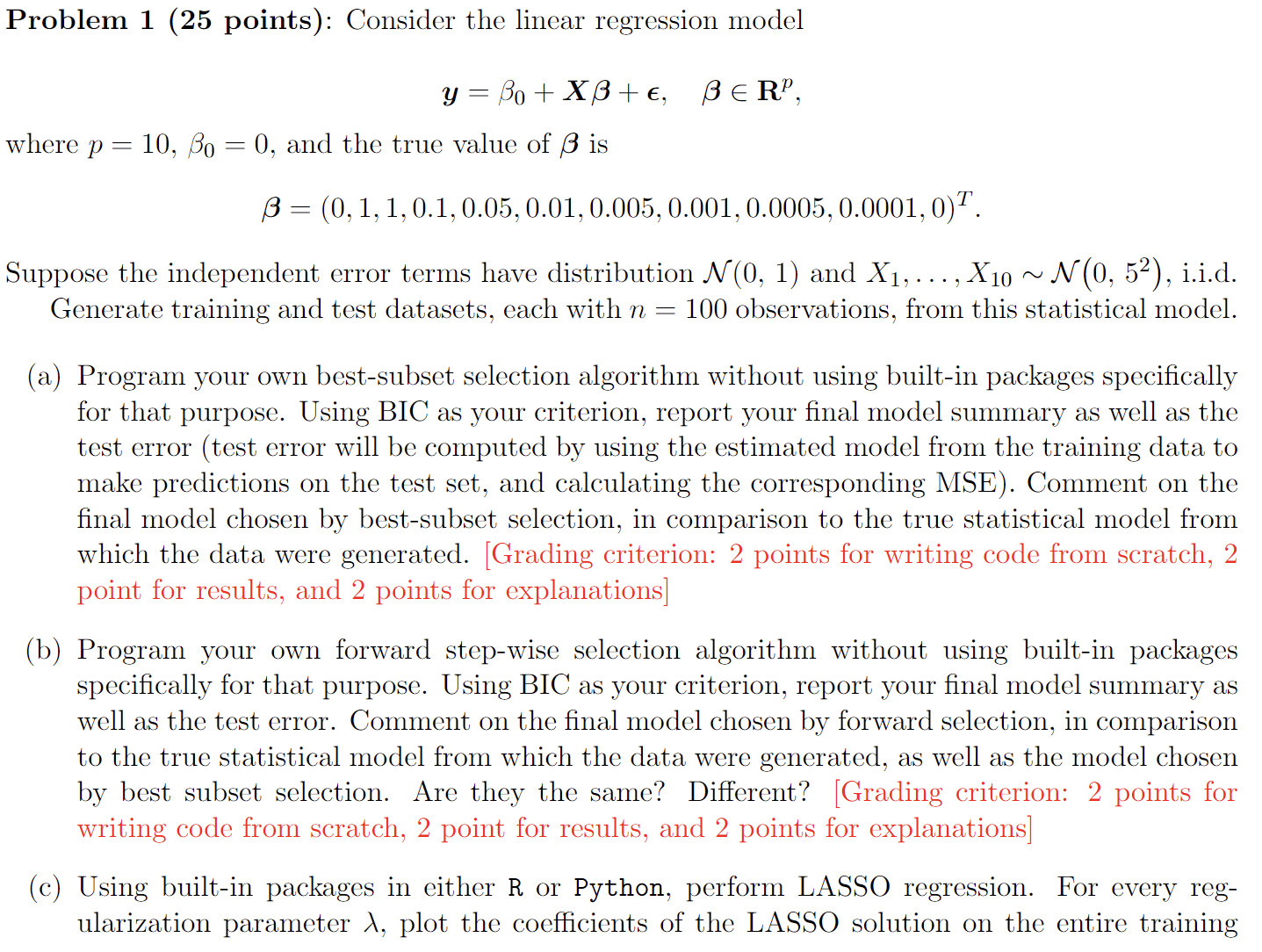
1. RSScurrent - RSSfull starts increasing and,

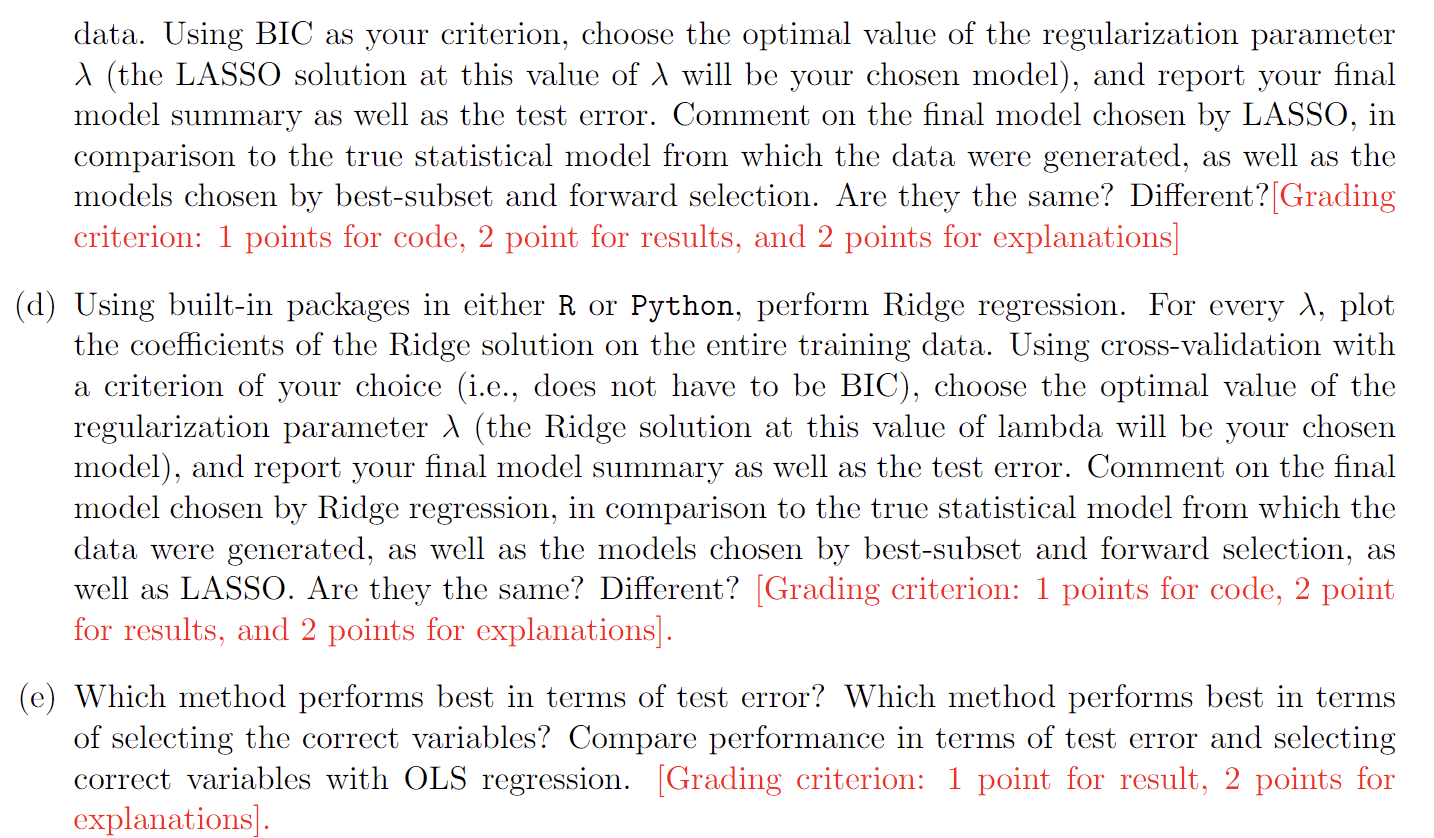
2. BICfull – BICcurrent starts decreasing

Thus, we can use a theoretical optimization that simultaneously attempts to find:  
a. MAX(BICfull – BICcurrent) and

b. MIN(RSScurrent - RSSfull)

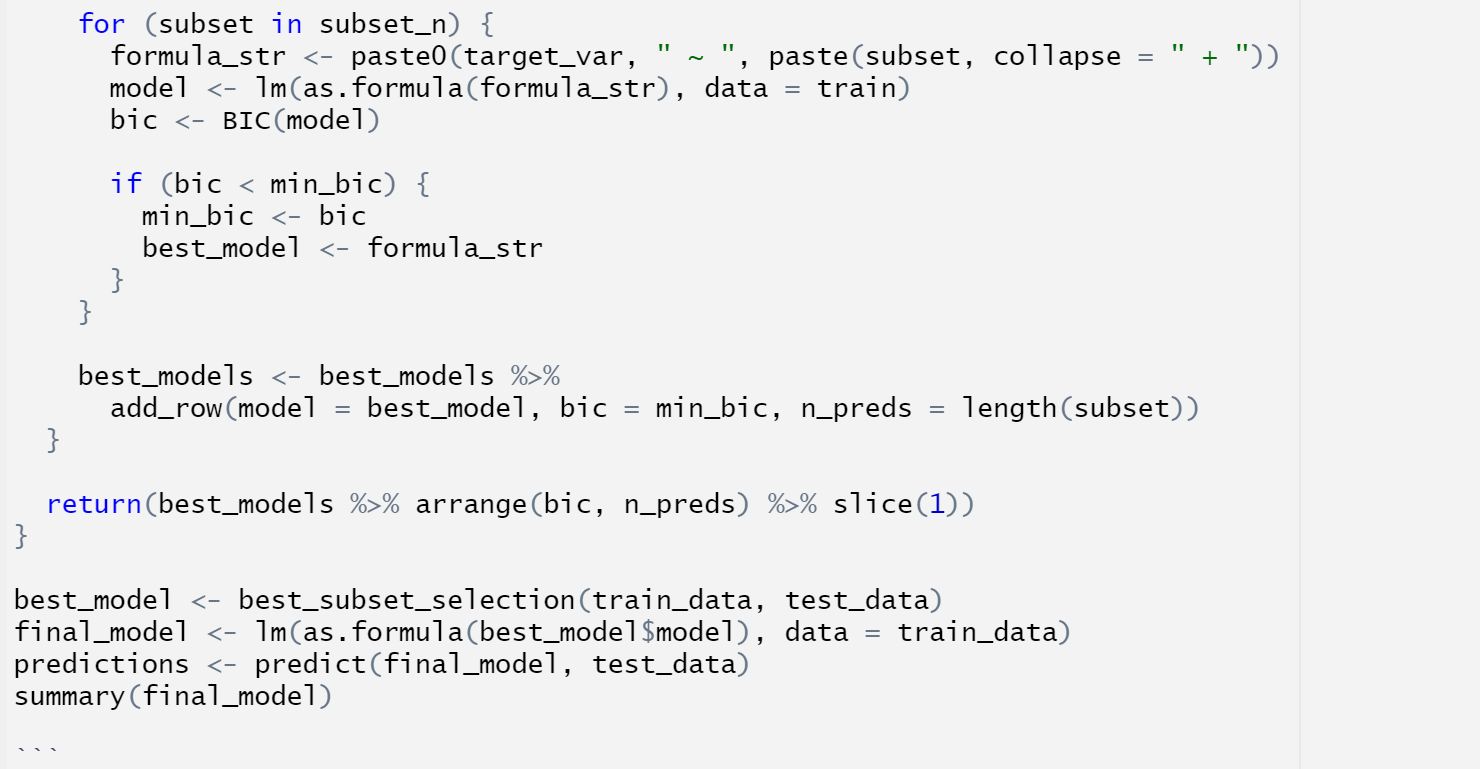
**Applied Questions:**

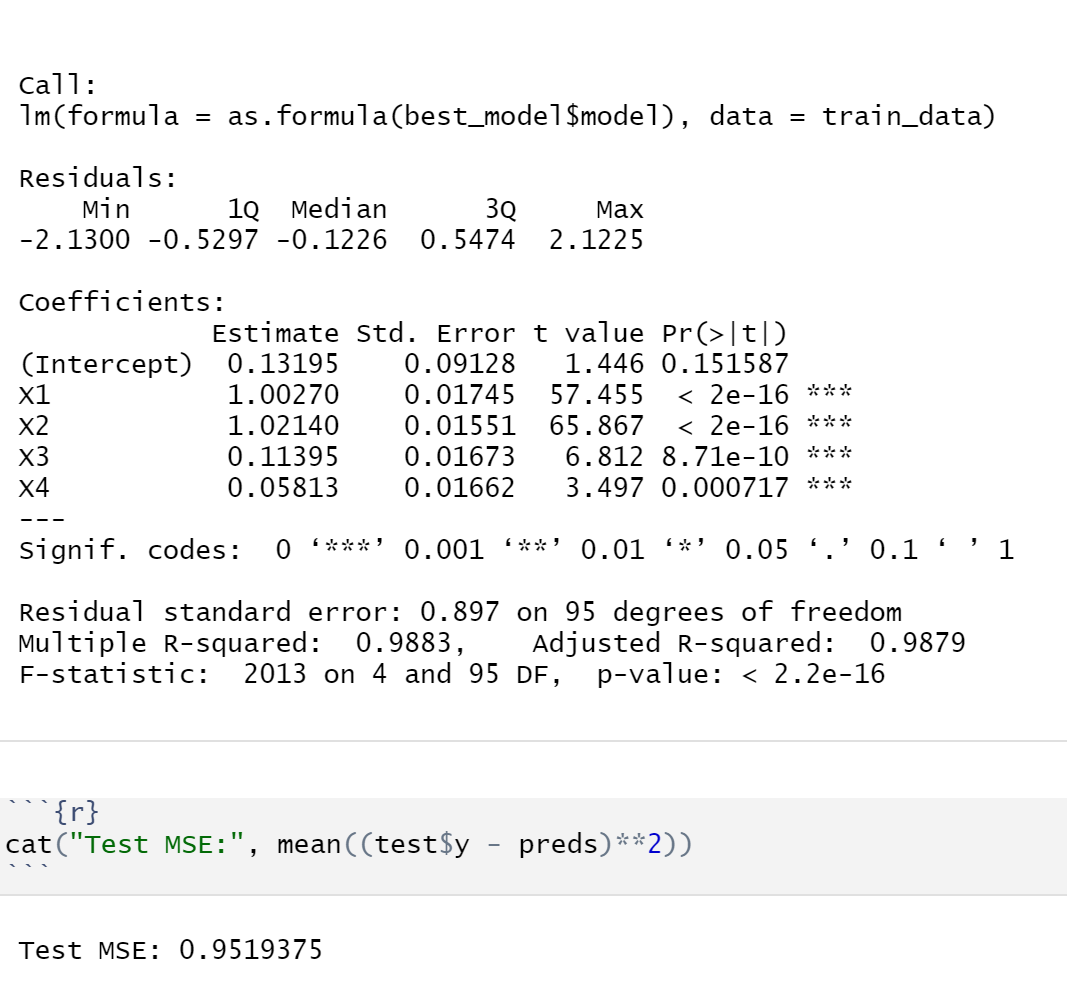




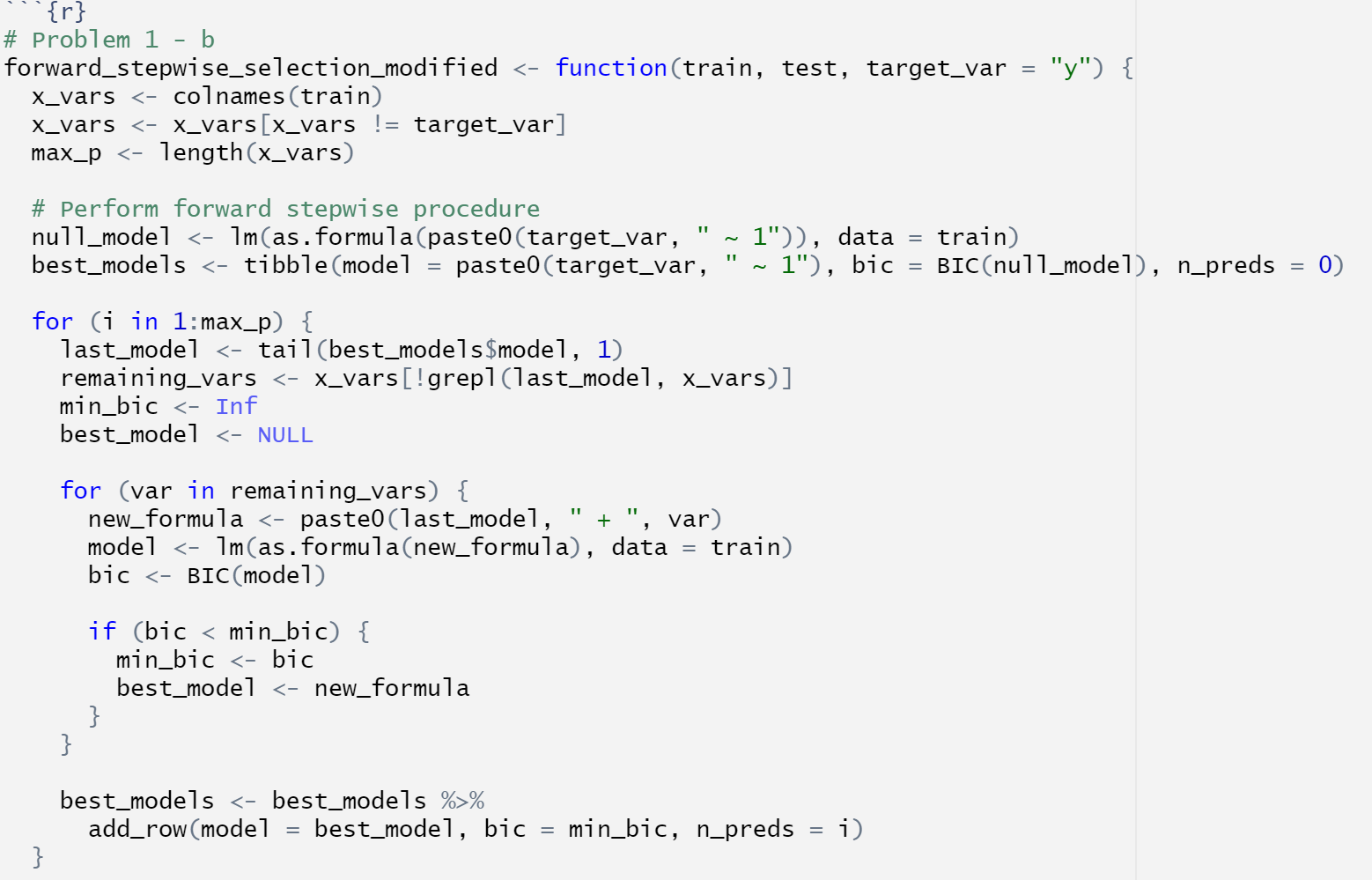
Solution:

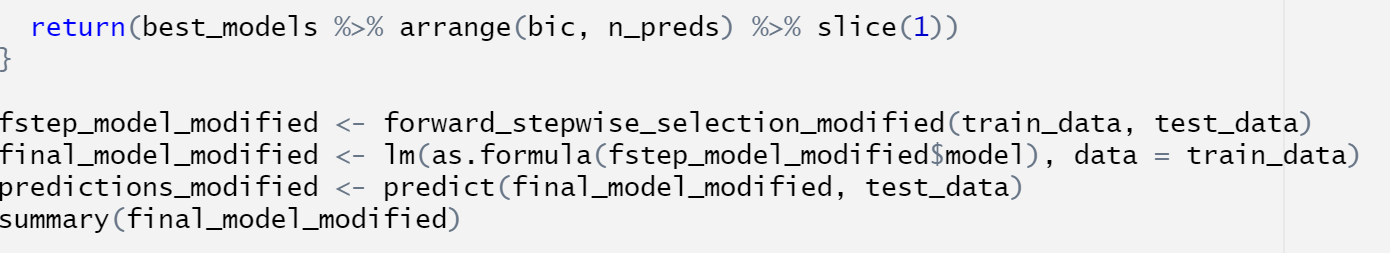


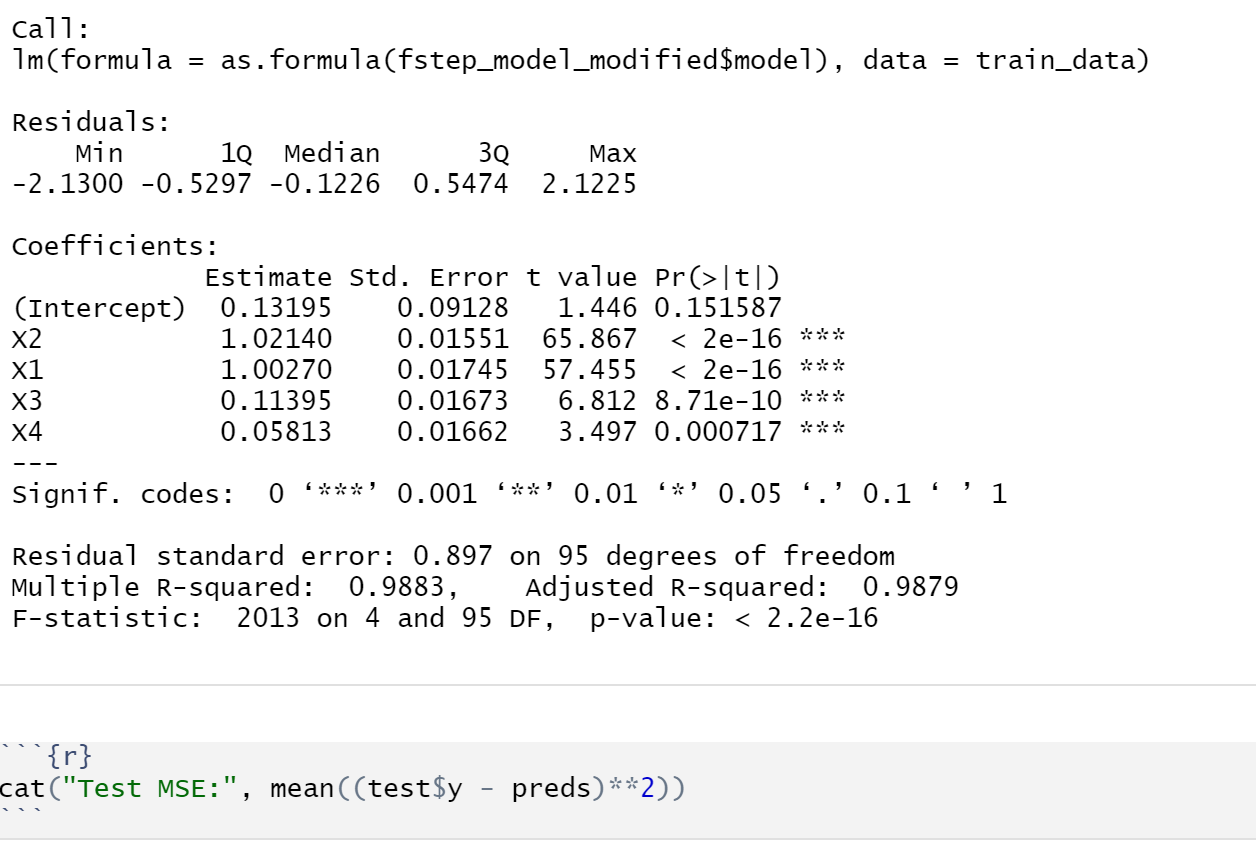




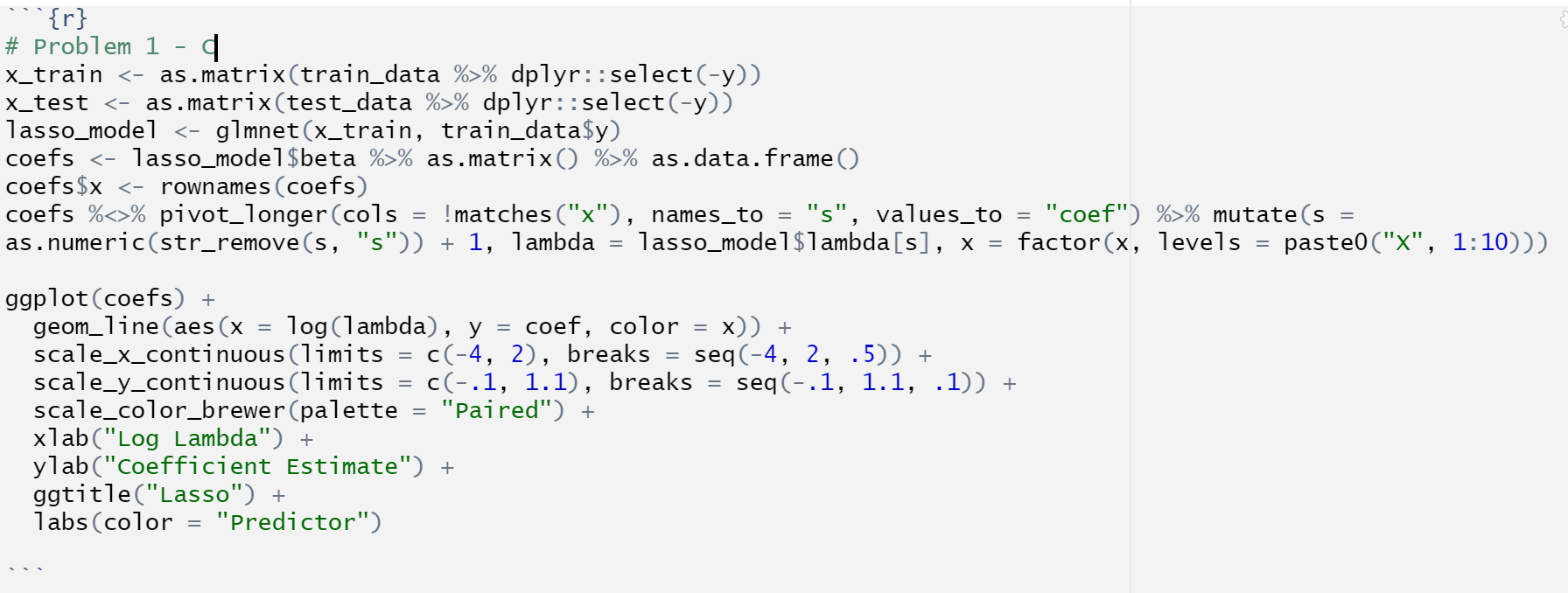
We observe that the model obtained after applying the best subset process, that we only retain 4 of our features namely X1, X2, X3, and X4. There is an error term in y, and that error term could potentially contribute to noise that leads to only the 4 said features being selected. The test MSE is 0.95159375

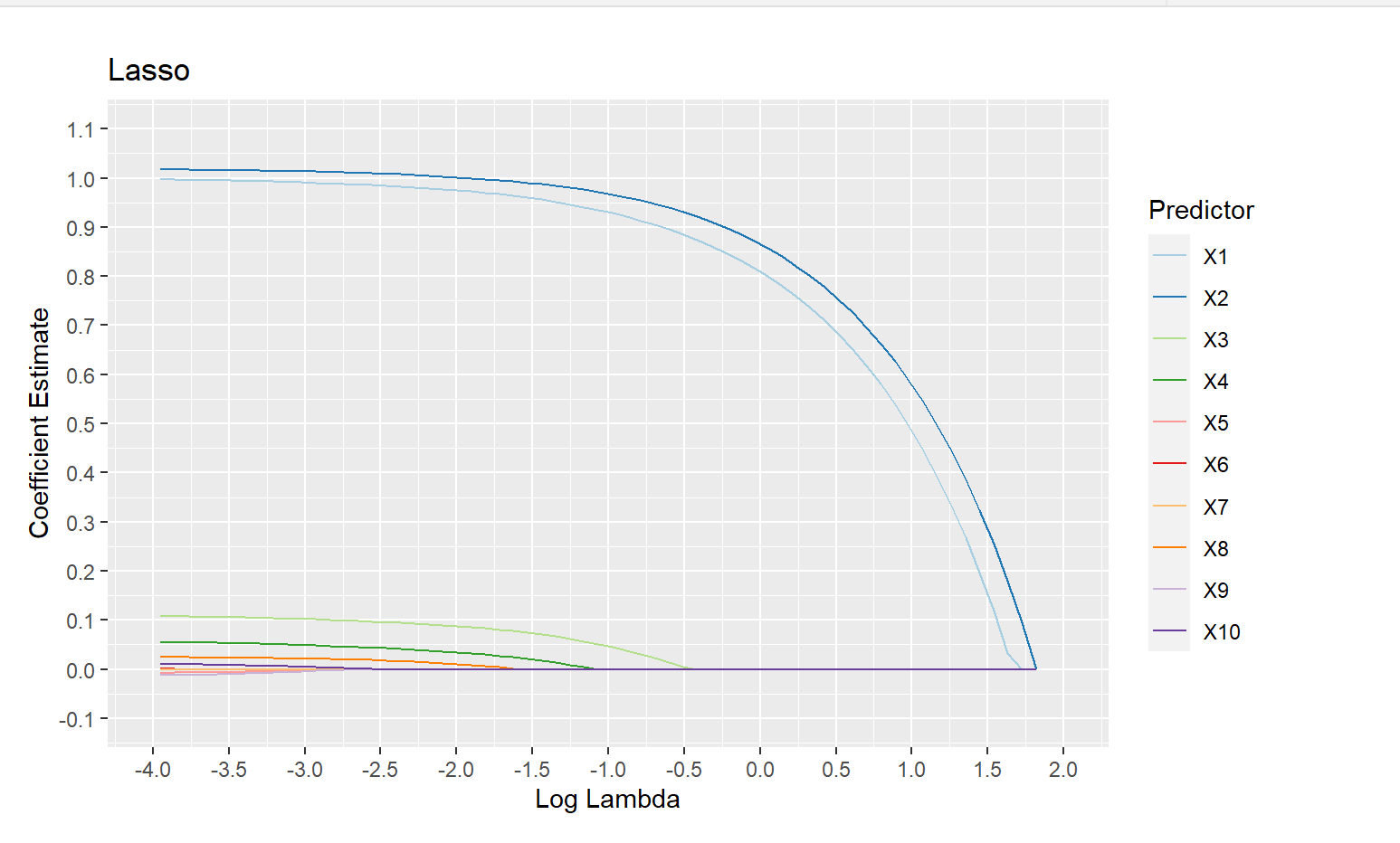


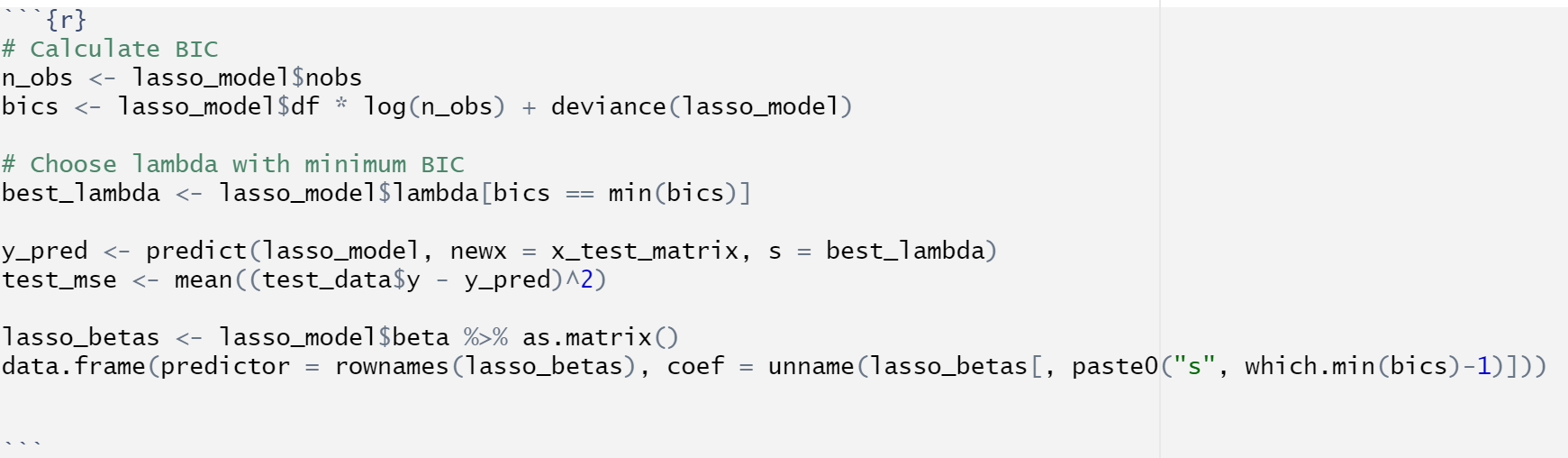


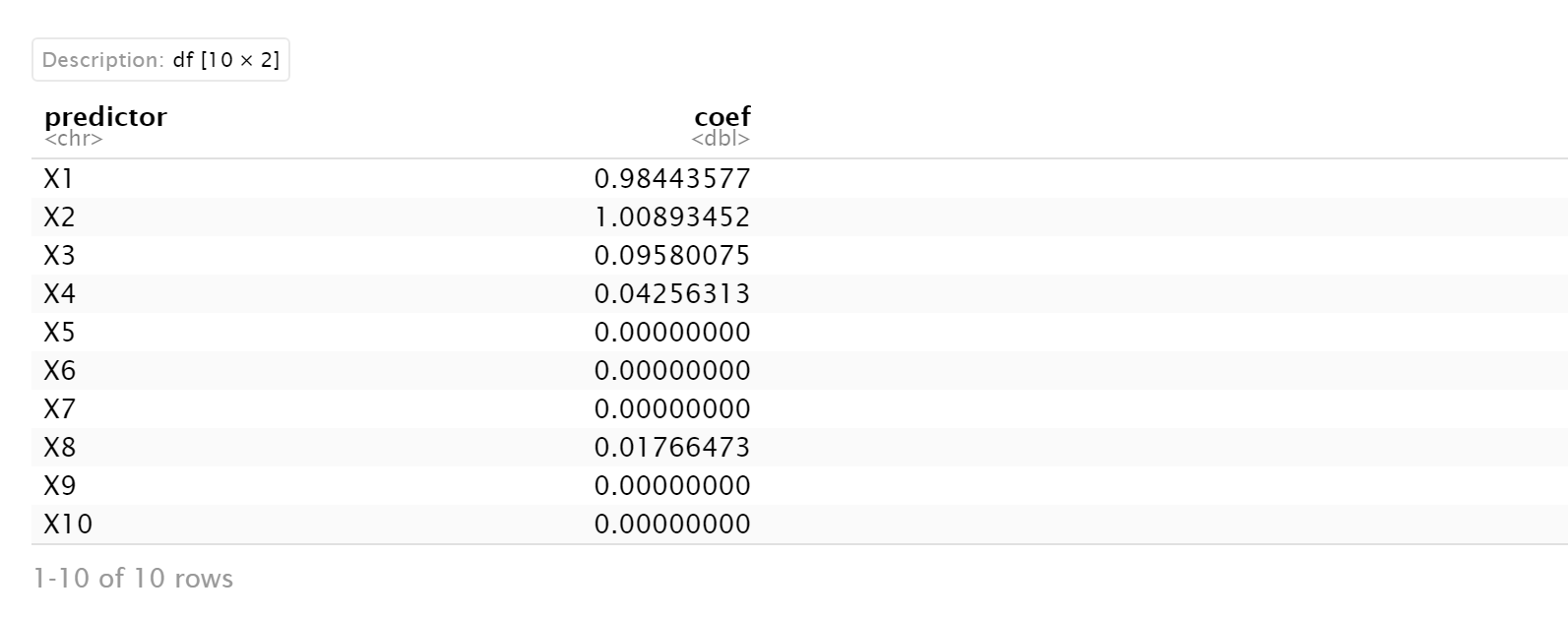


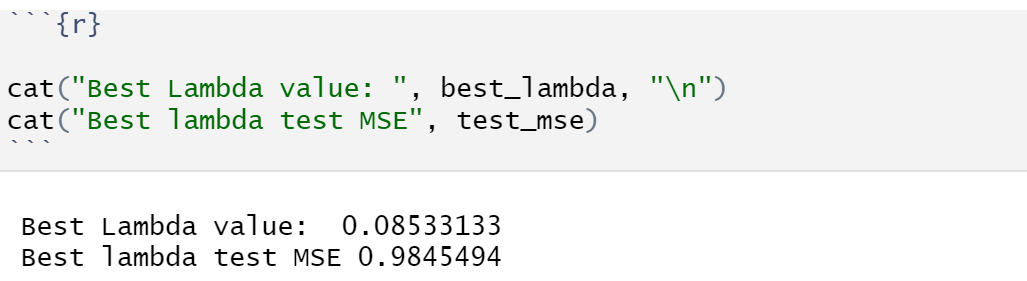
Observing the results of the forward stepwise selection, it is noticed that the test MSE is exactly the same. Furthermore, the selected features are also the same as in the previous case.

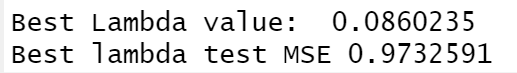




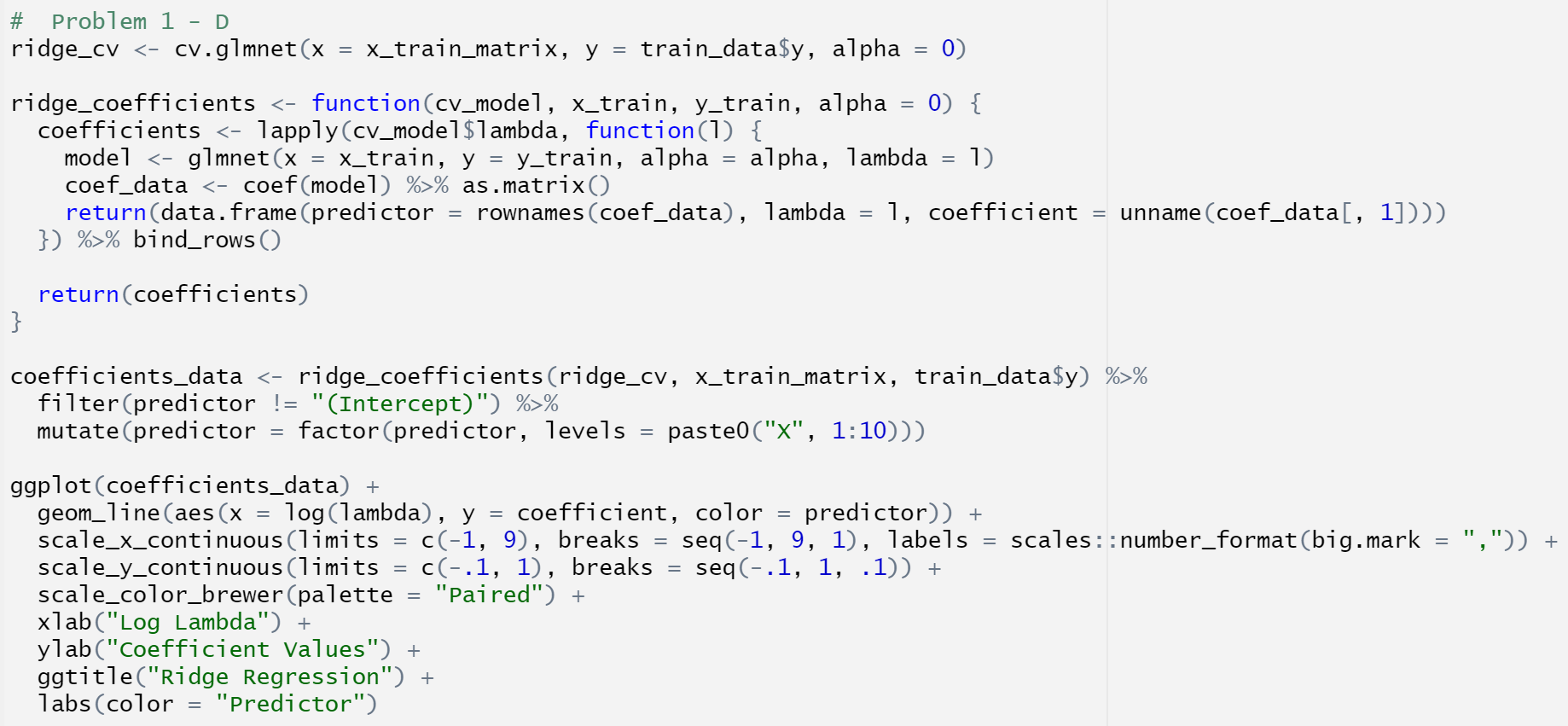


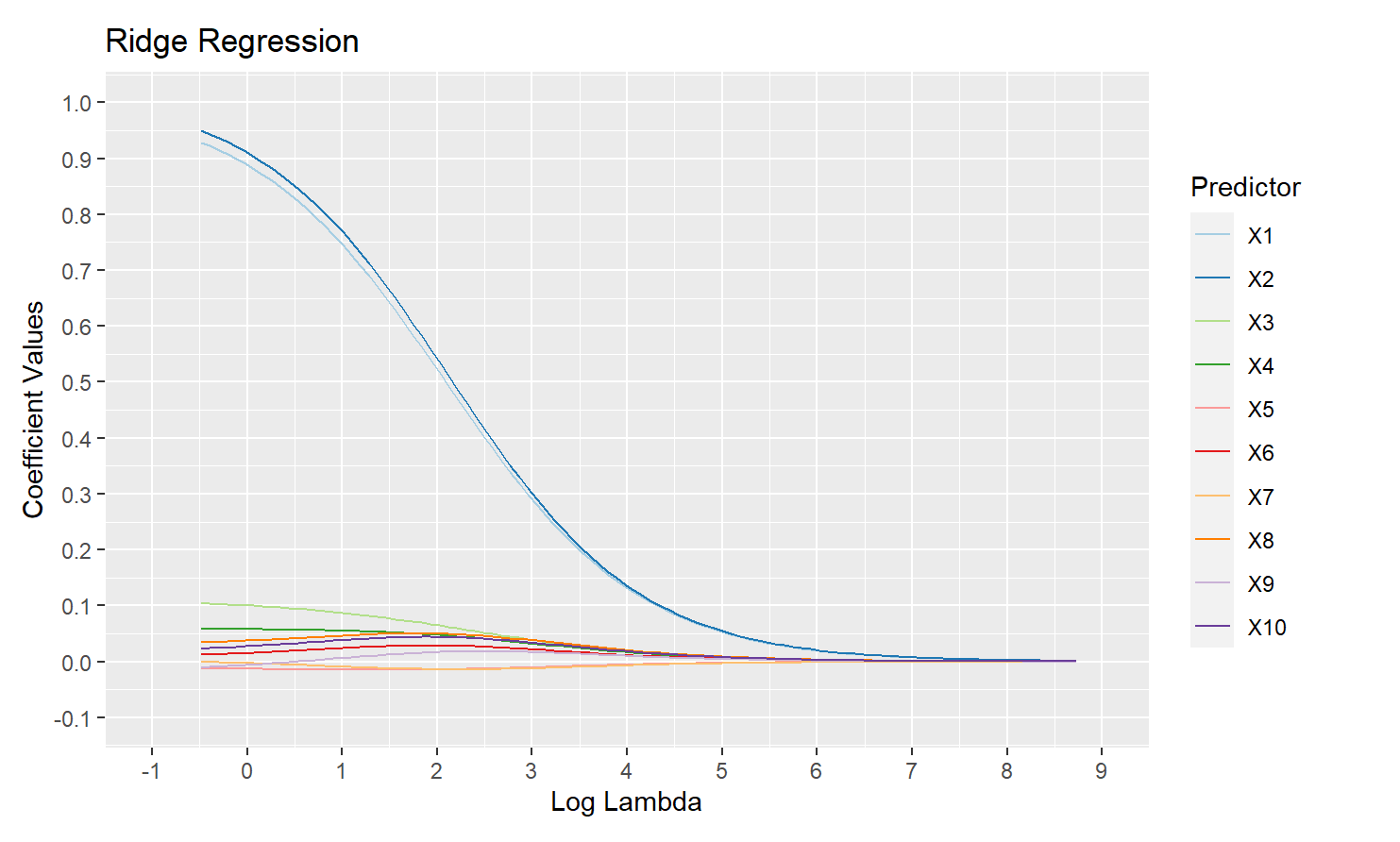


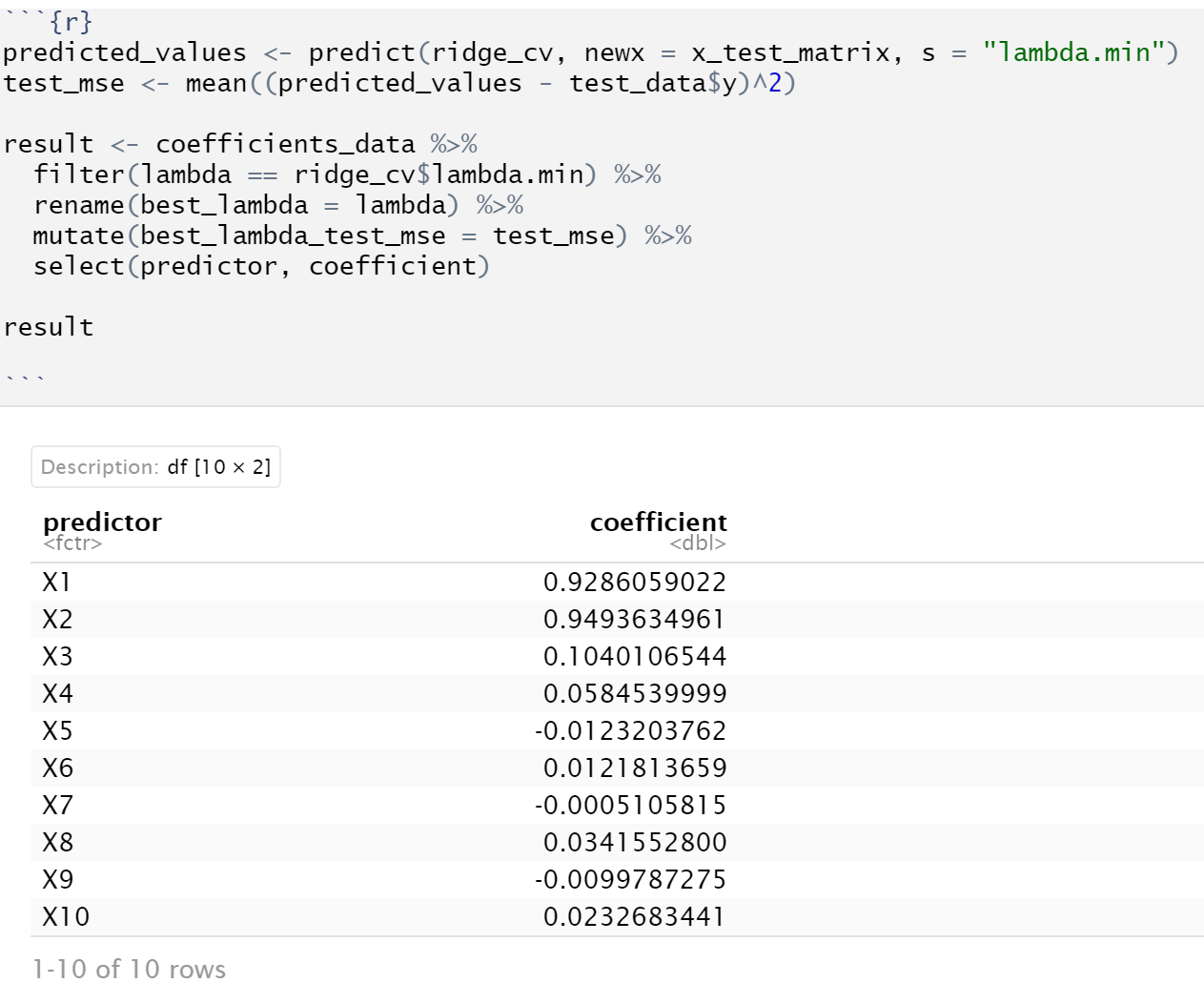


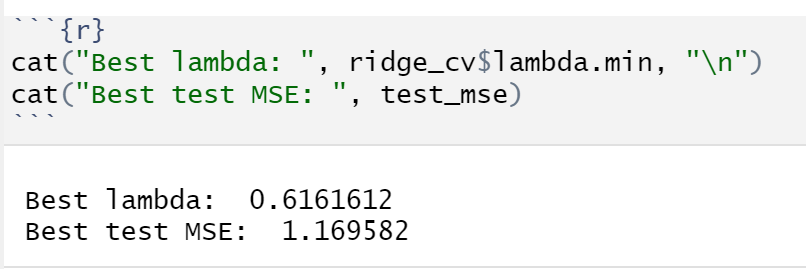


It is observed that the model selects features X1, X2, X3, X4, and X8. The best lambda value is 0.860235, and the best test MSE is 0.9732591. The coefficients for all the features X1 to X4 are similar to the previous model.

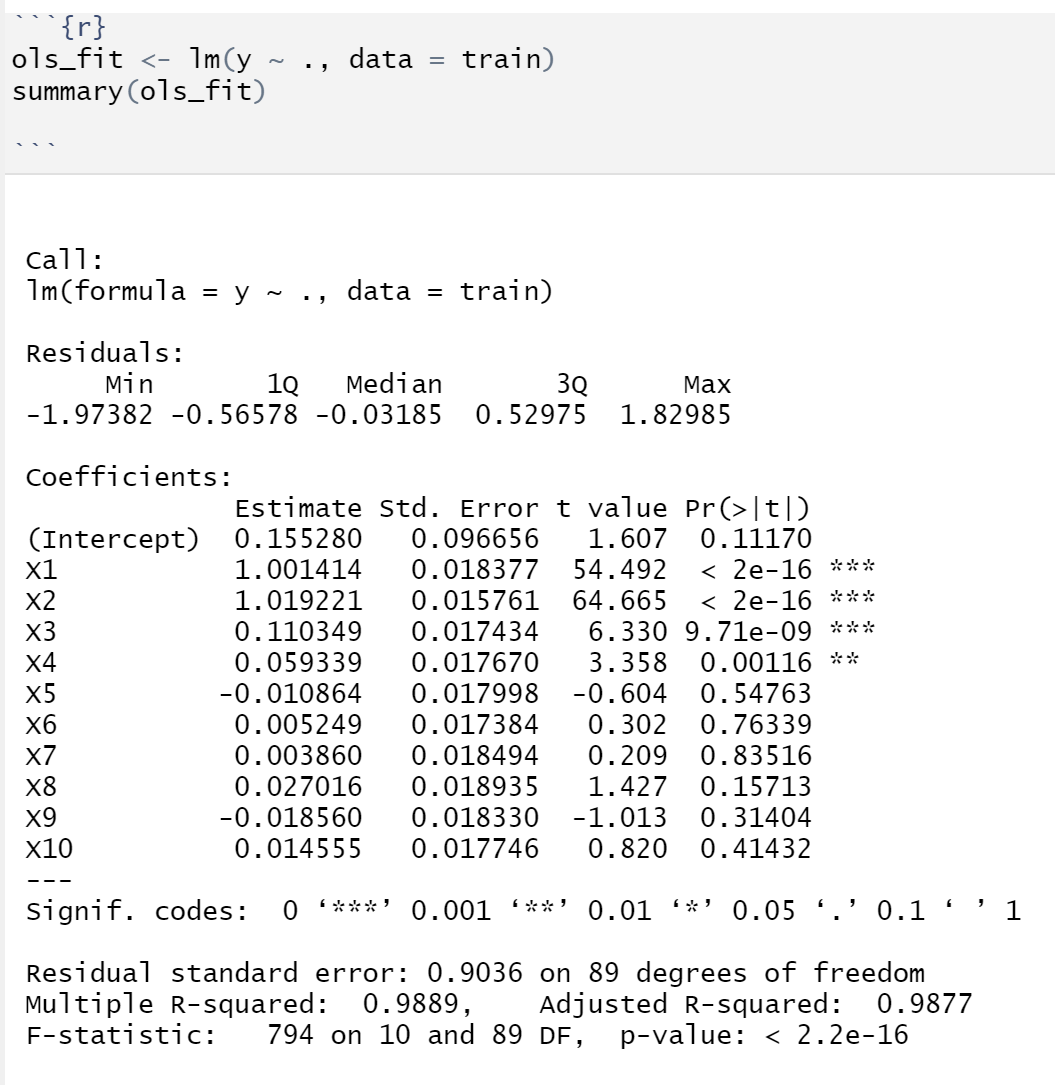




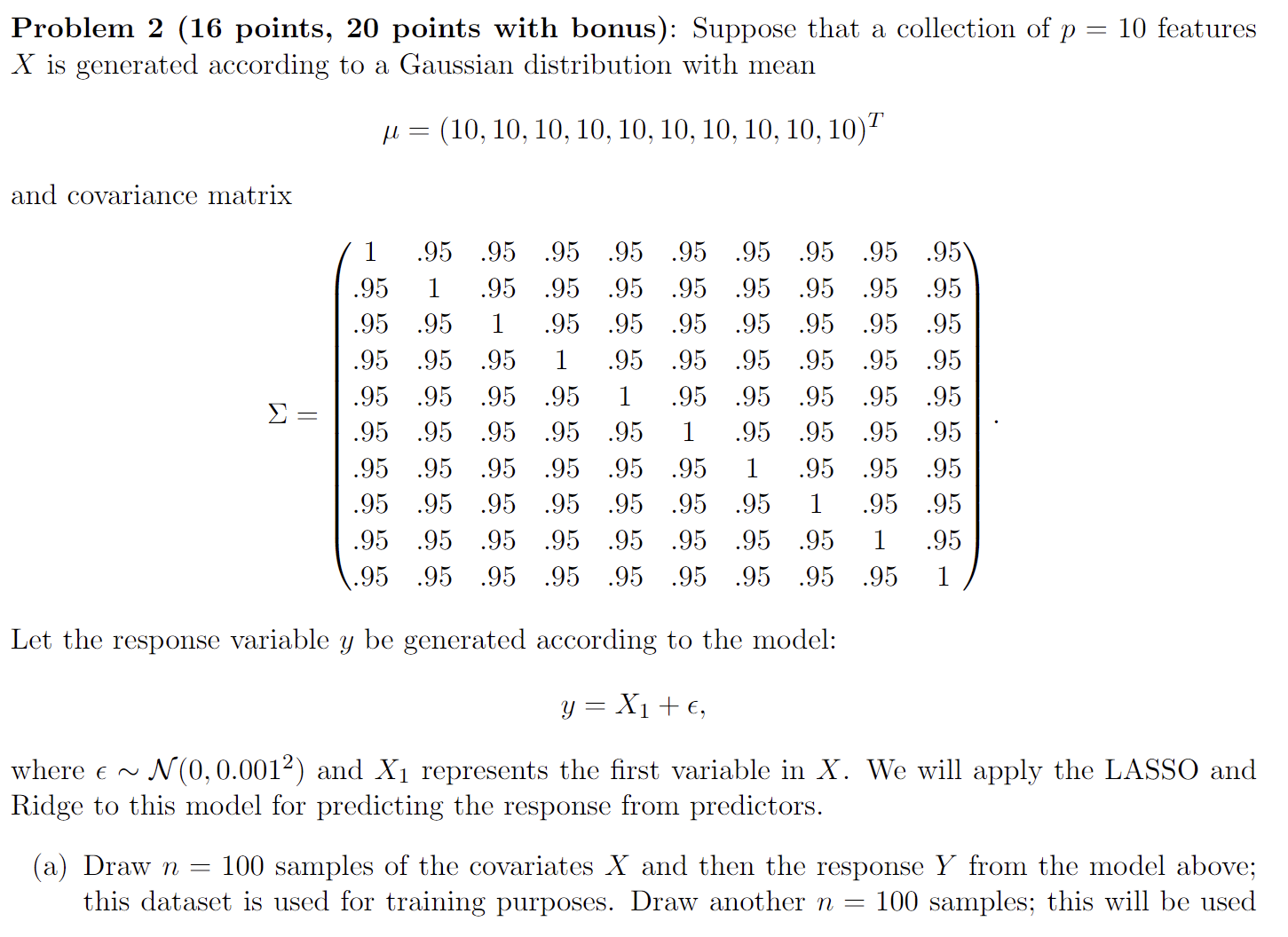


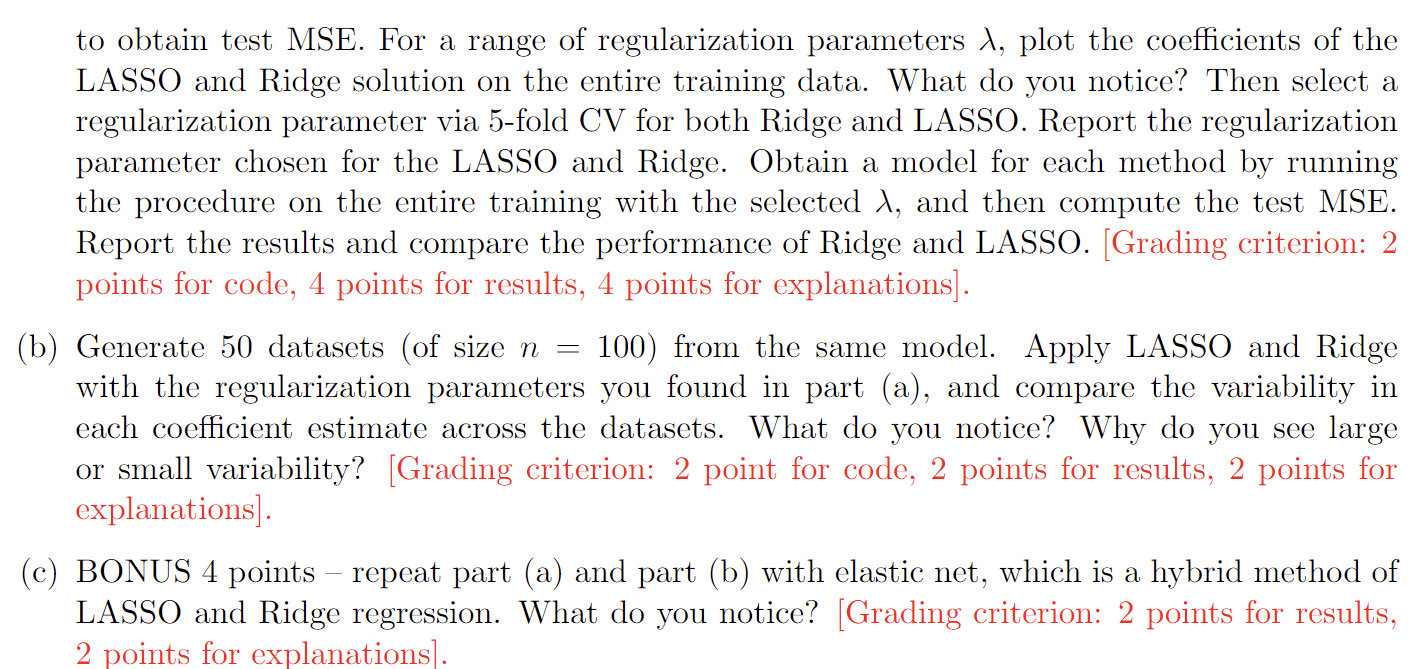


The best lambda observed here is 0.6161612, while the best test MSE is 1.169582. It is observed that the model does not set any coefficients to 0, since it is a ridge regression model, and hence produces a more complex model. Some of the coefficients are different from the previous models as well.

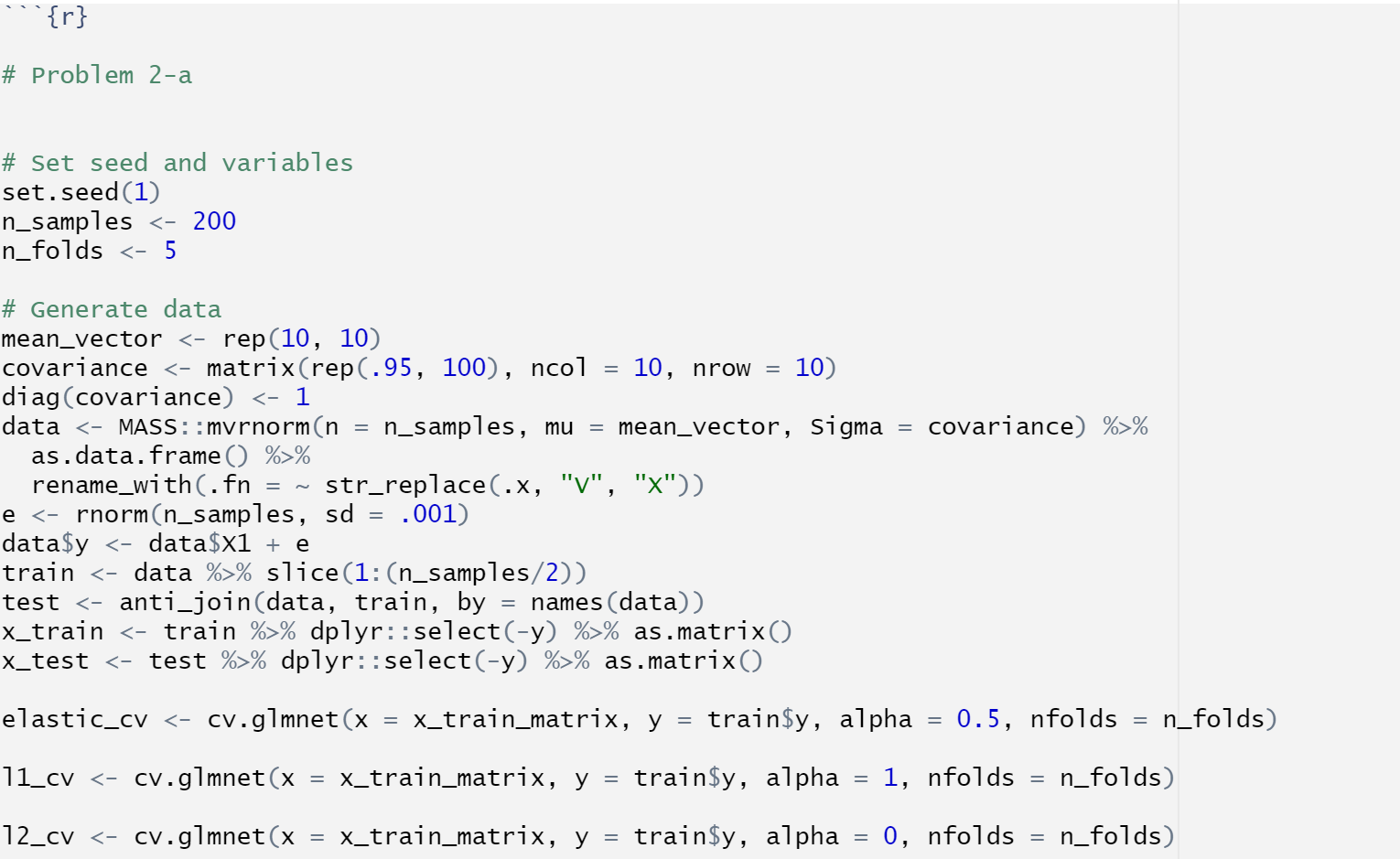


From the model summary above, we see which variables are statistically significant. It is observed that the subset selection and the forward step selection processes both selected X1, X2, X3 and X4, while Ridge and Lasso Regression selected more variables in addition to the ones earlier. Furthermore, the MSE in the case of Ridge and Lasso is also higher than that of the best subset selection and the forward step selection processes.

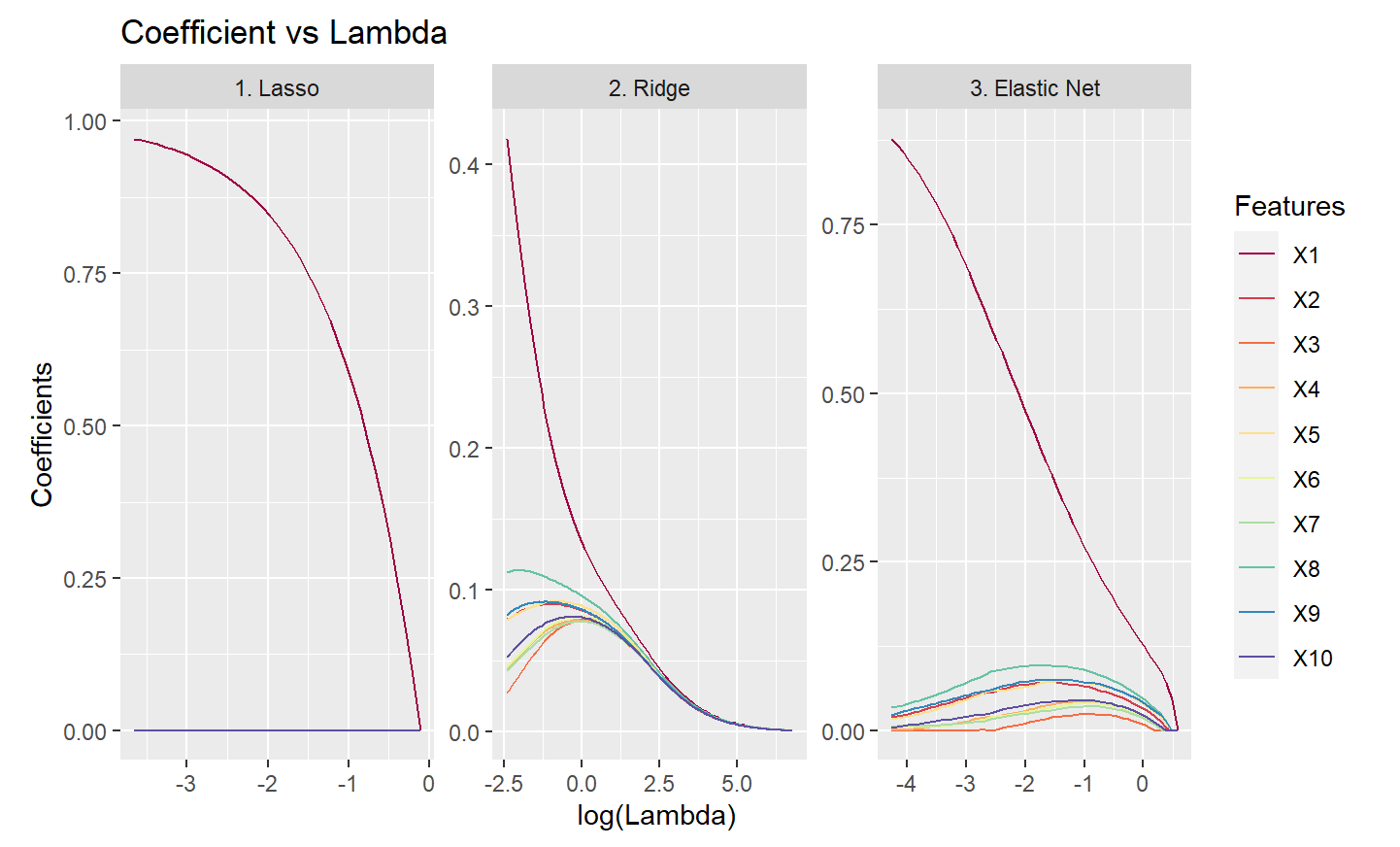




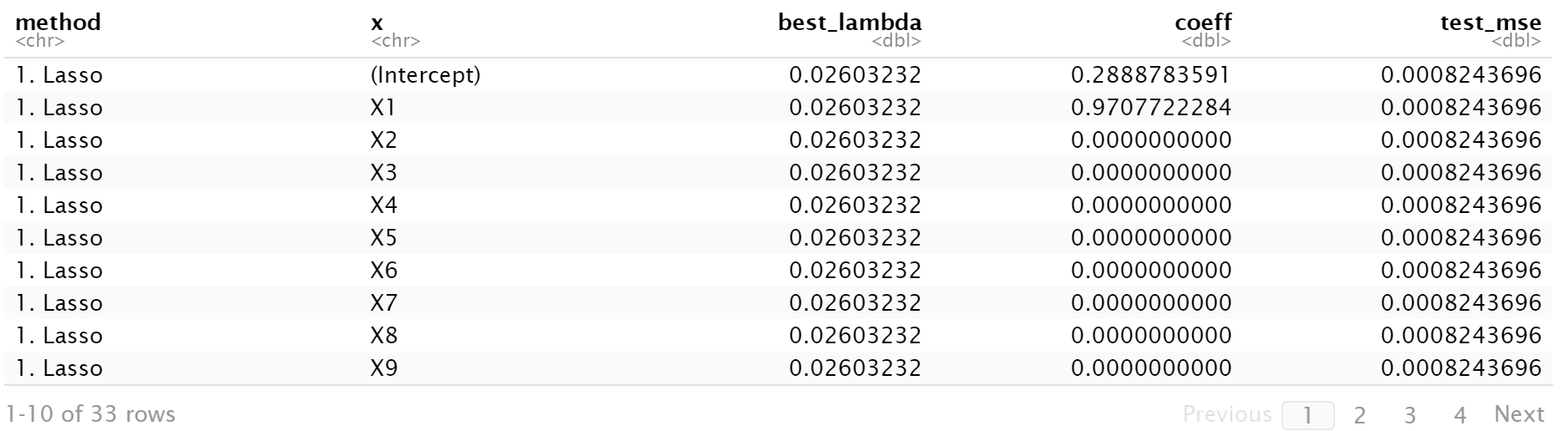
Solution:

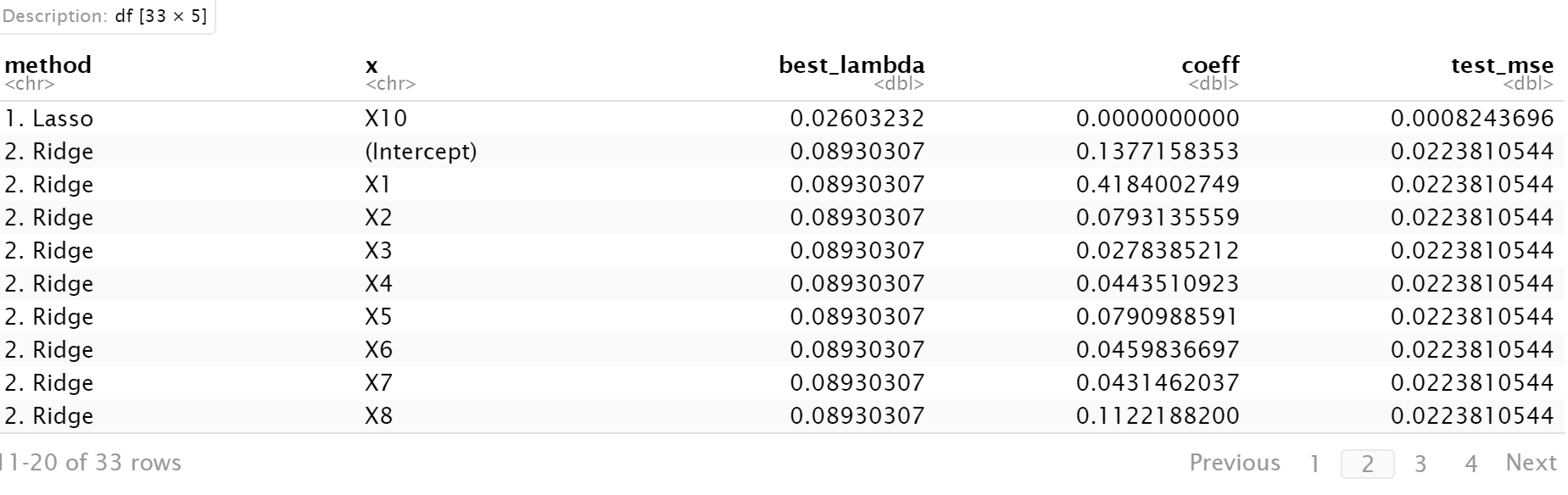


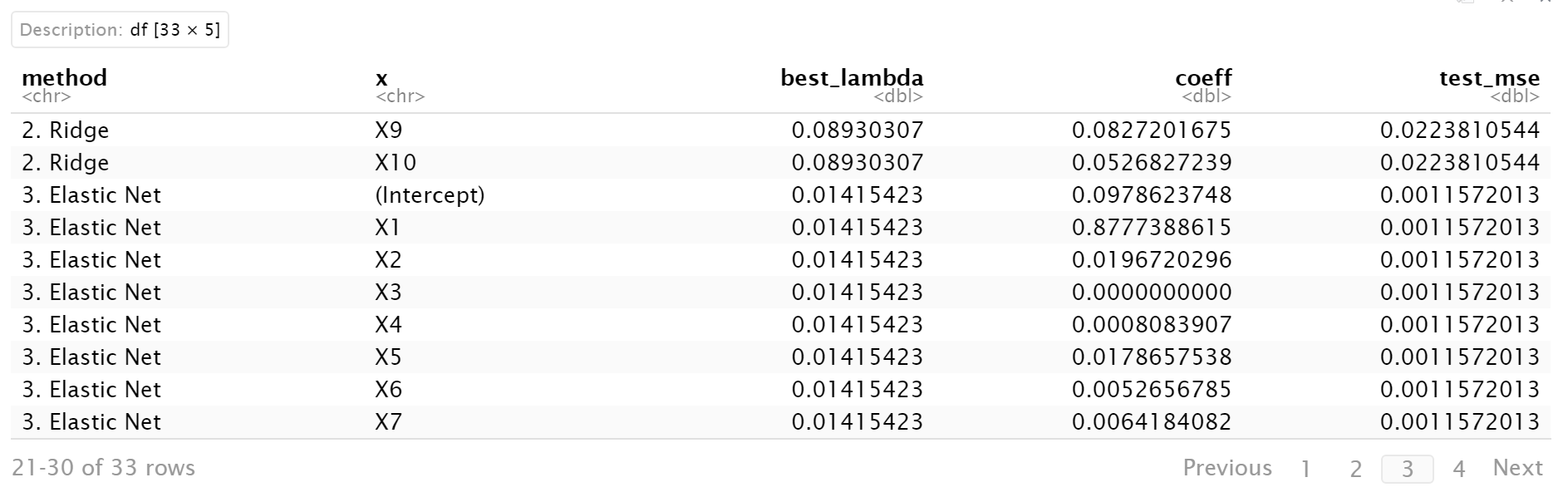








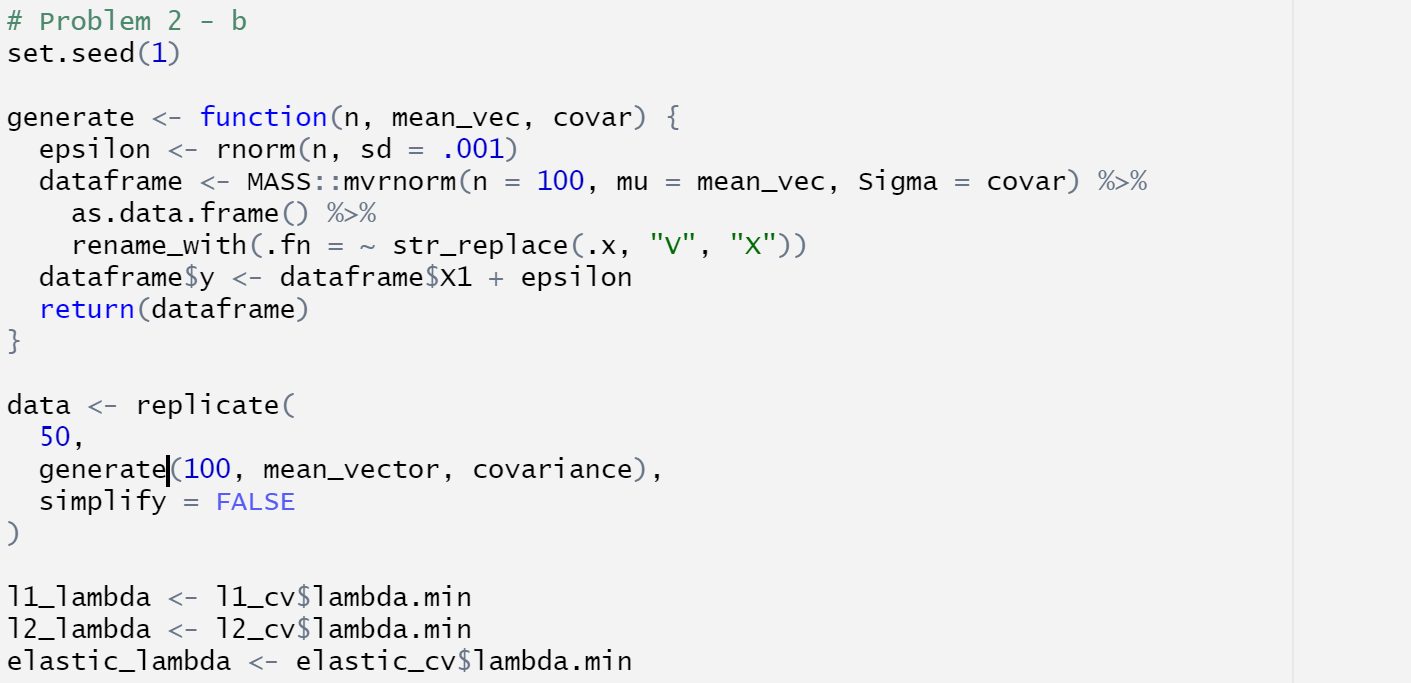


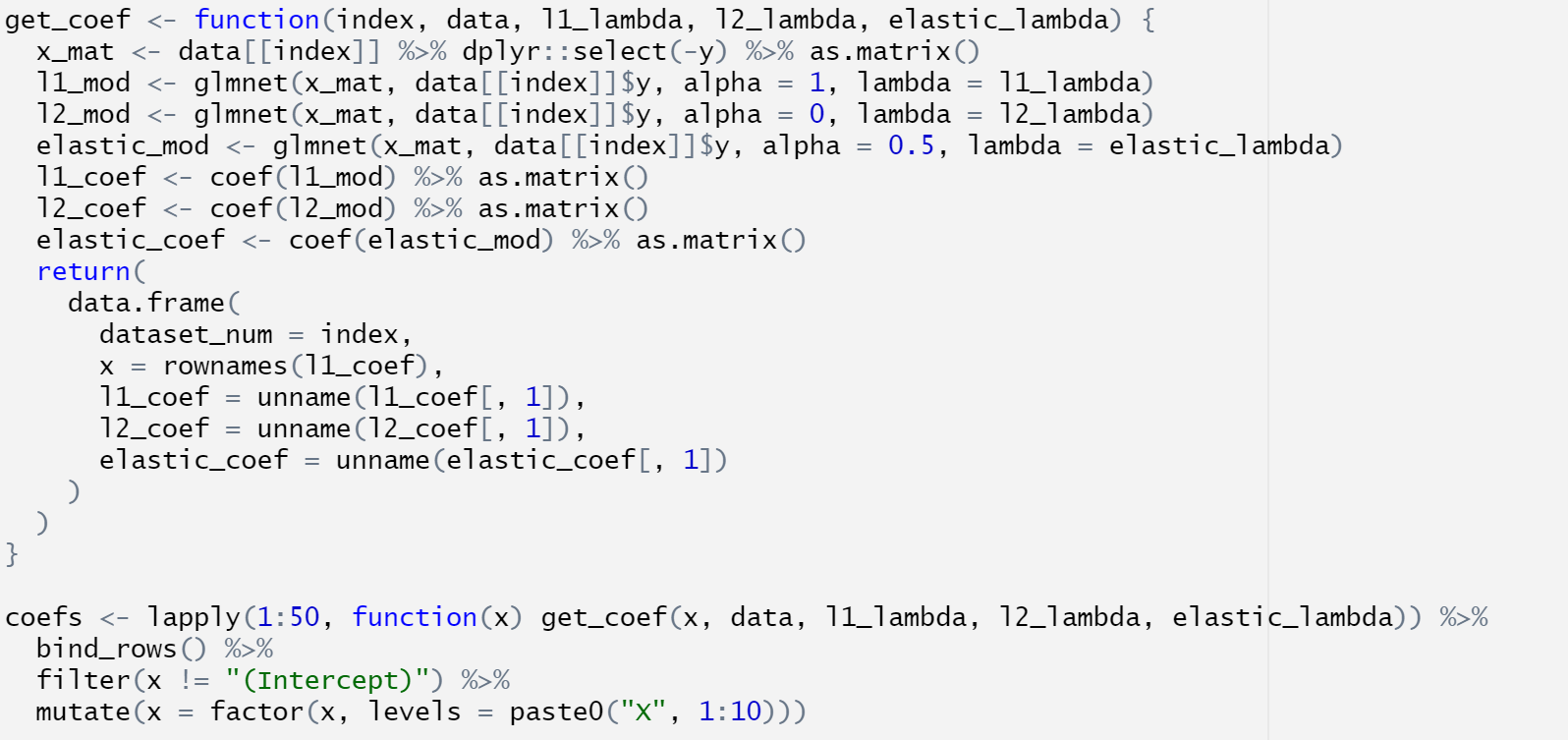


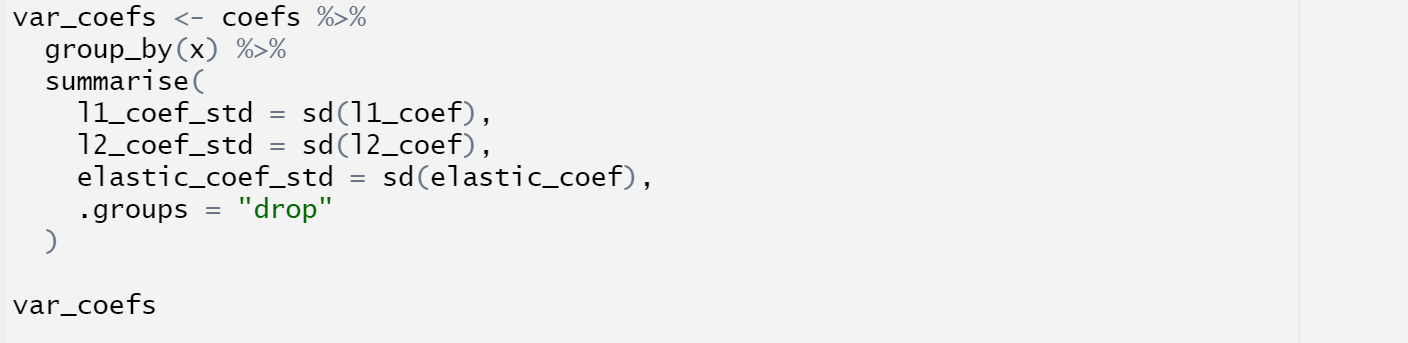


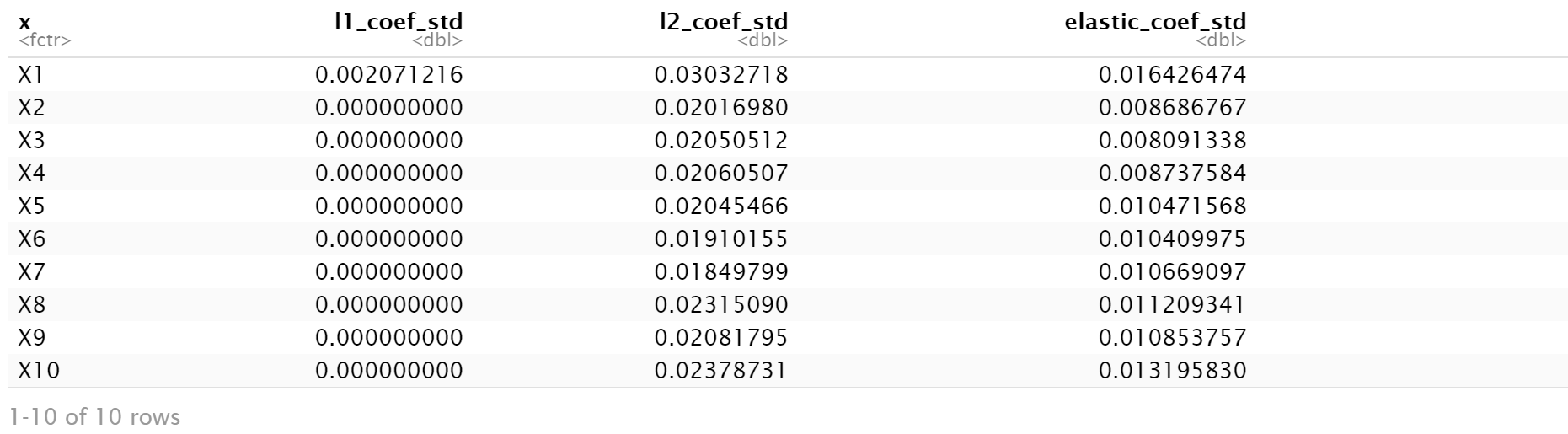
The Lasso Regression Method has a non-zero coefficient for only X1, as it encourages all other coefficients towards zero – resulting in a much less complex model compared to Ridge and Elastic Net Methods. It also has the lowest test\_MSE among the 3 methods. In these considerations, Lasso regression performs the best.

In order of performance: Lasso > Elastic Net > Ridge

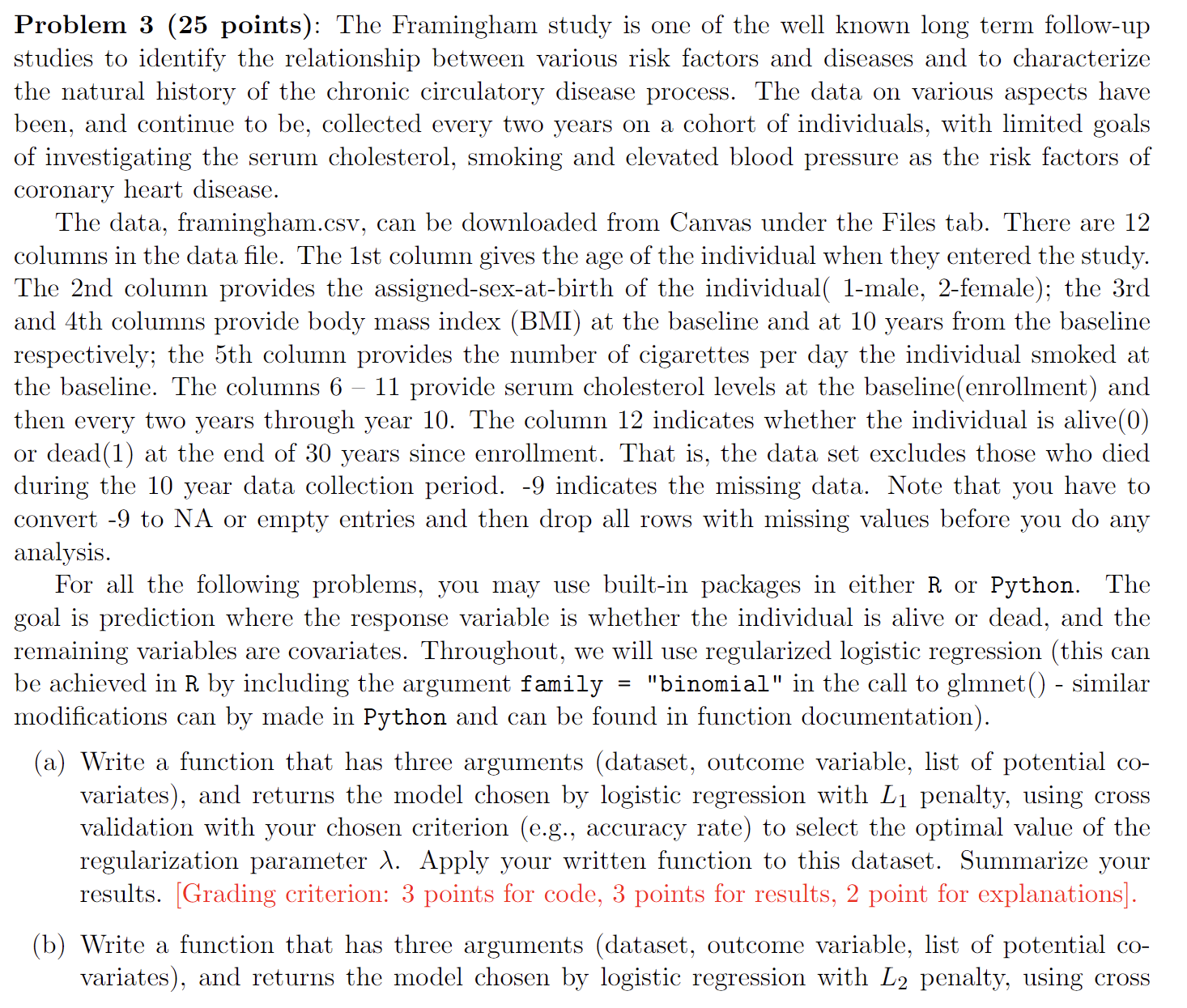


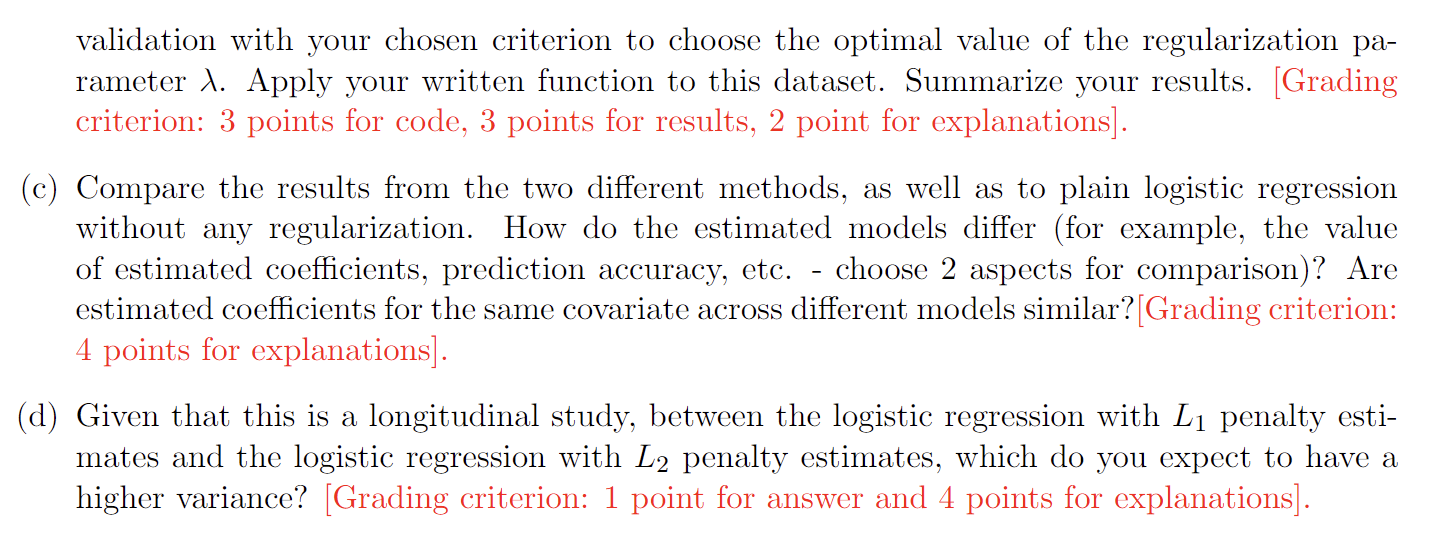




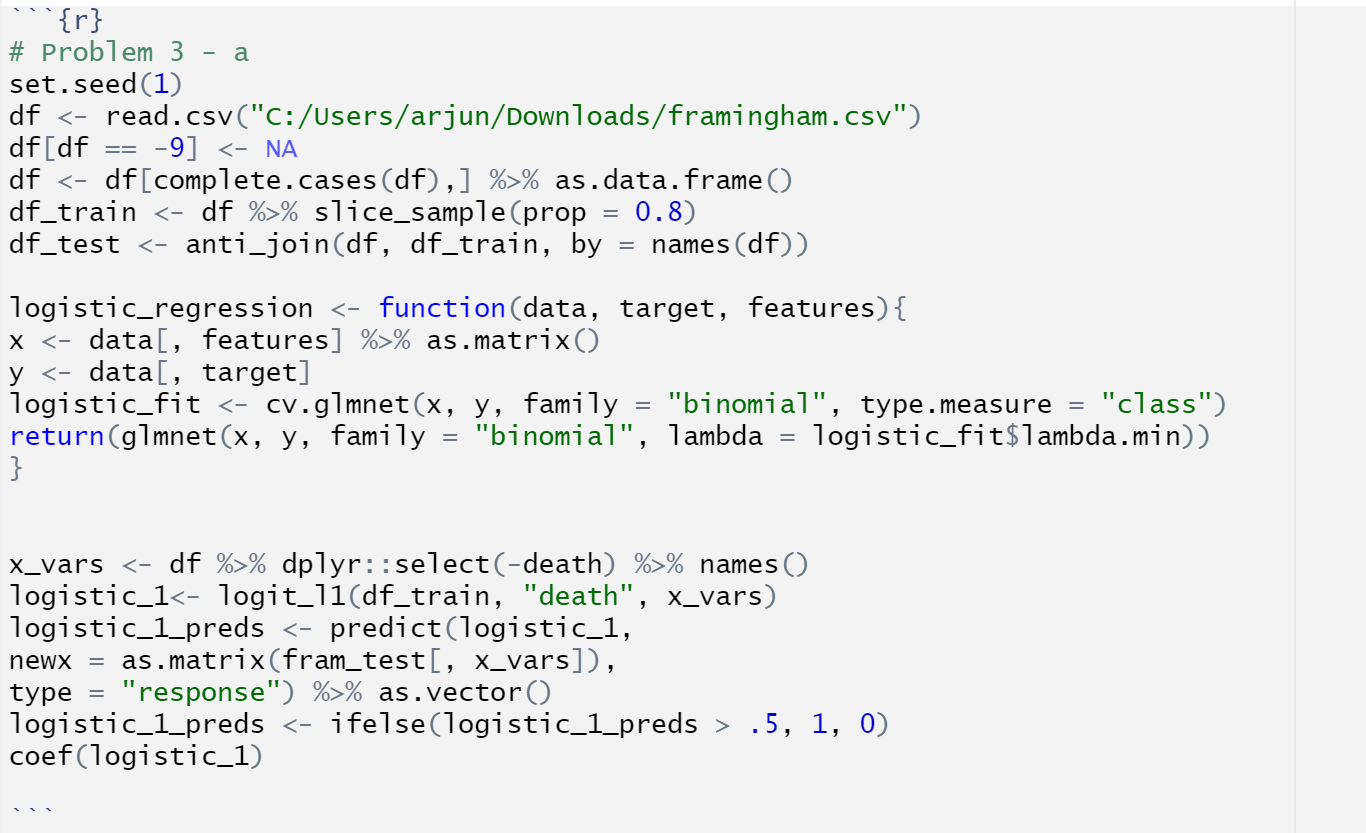


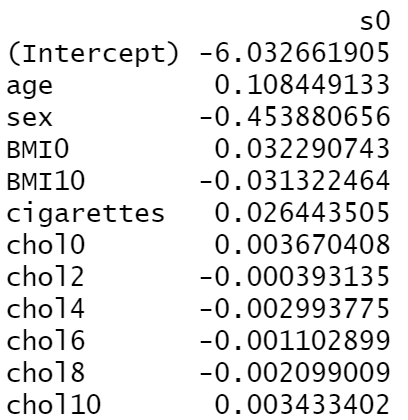
The lasso regression method had the least variability. Elastic Net is ‘in-between’ lasso and ridge in that it showcases intermediate levels of variability relative to ridge and lasso. Ridge has the highest levels of variability in the coefficients



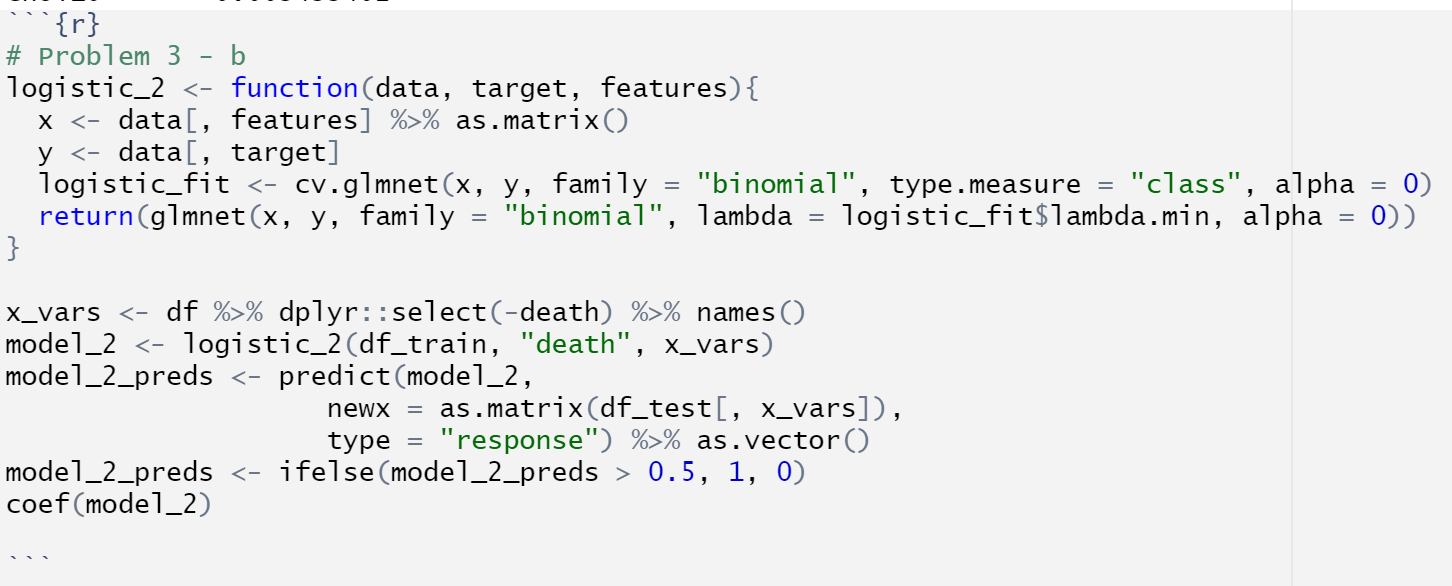


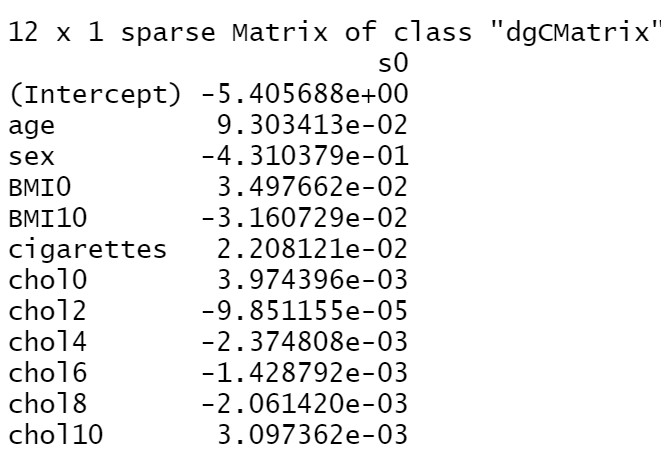
Solution:



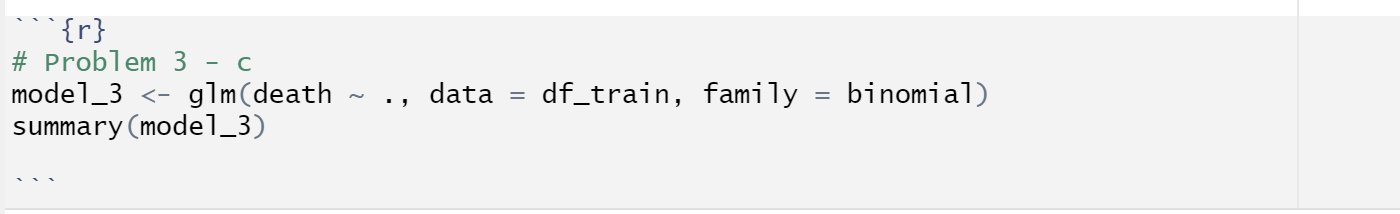


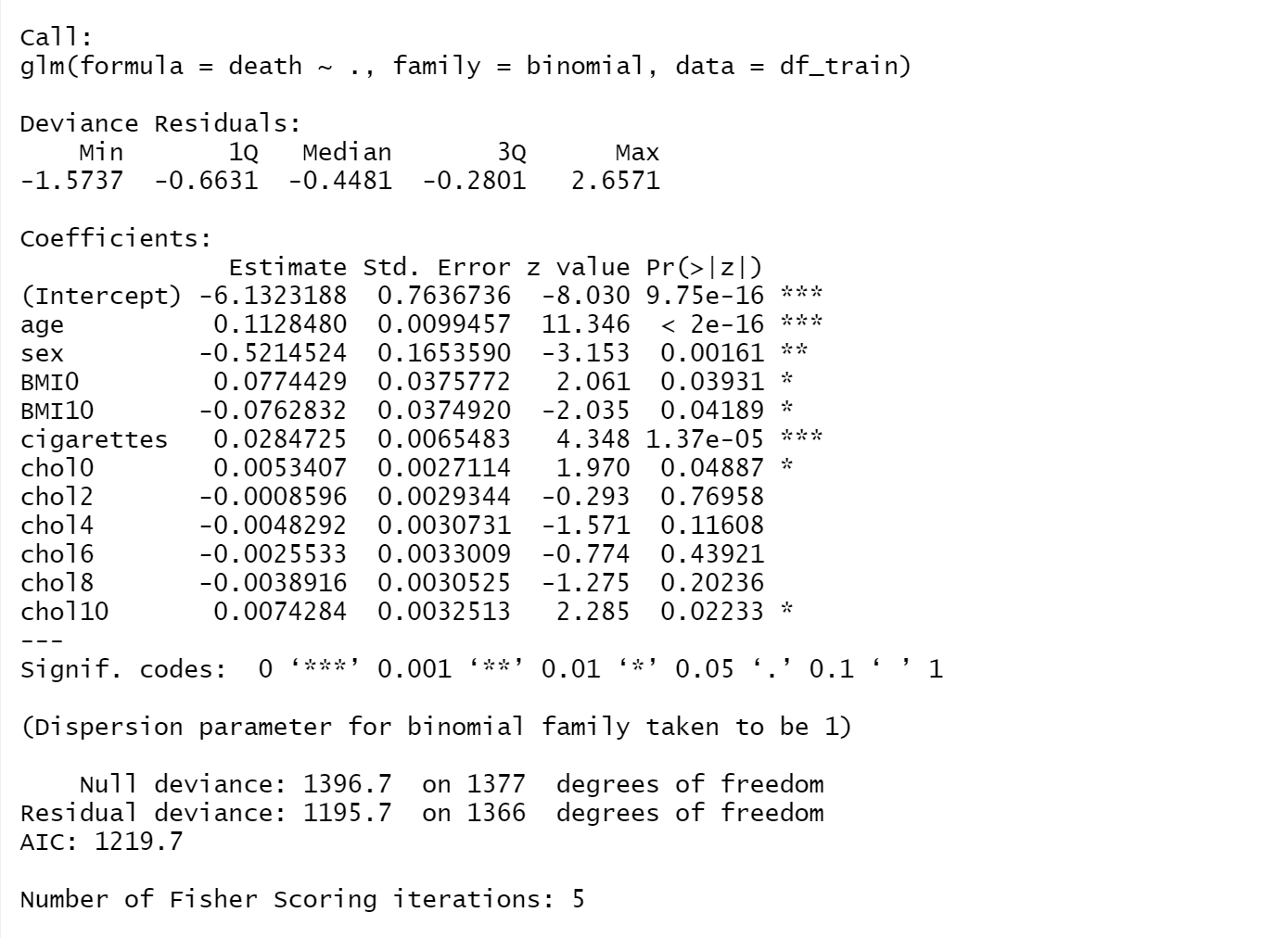
It is observed that the model produces a non-zero coefficient for all the features in the dataset. The highest coefficient belongs to age, while the lowest coefficient is from sex.

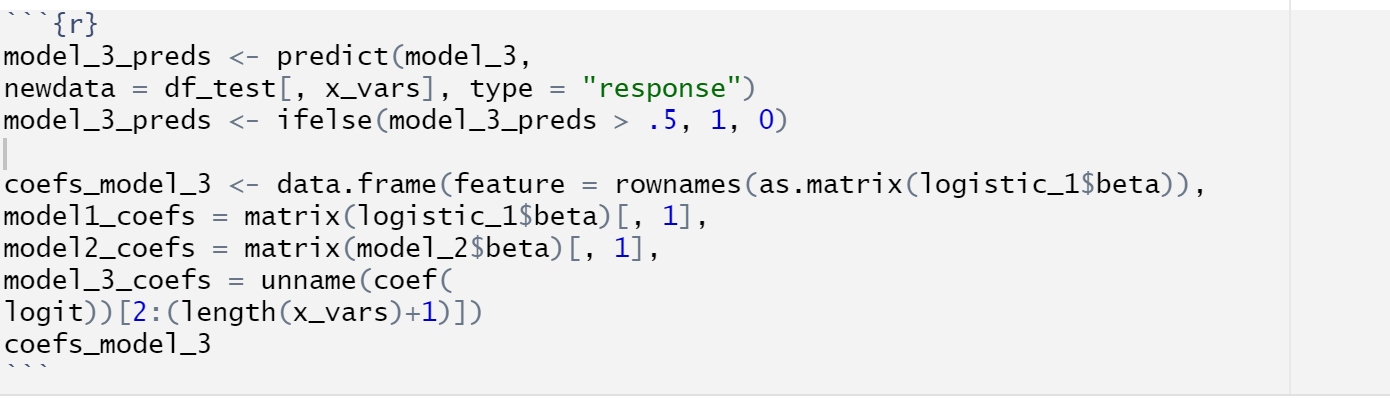


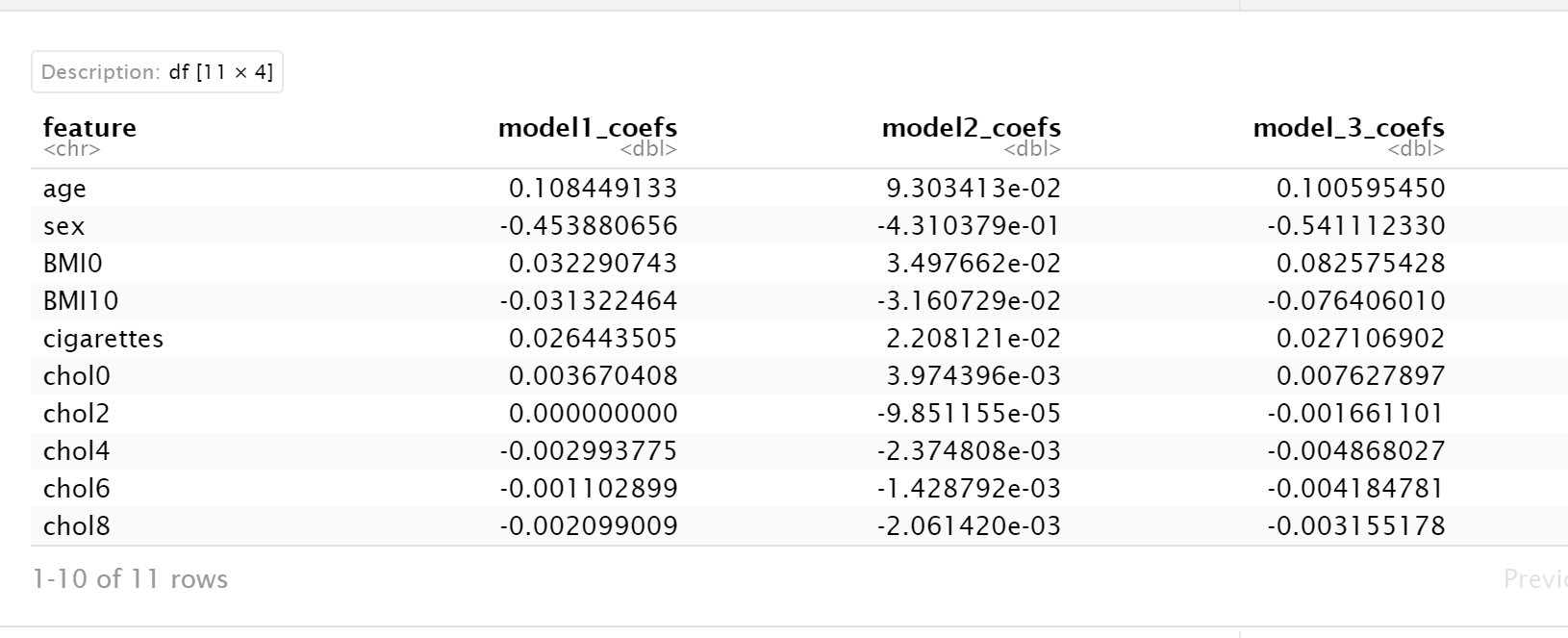


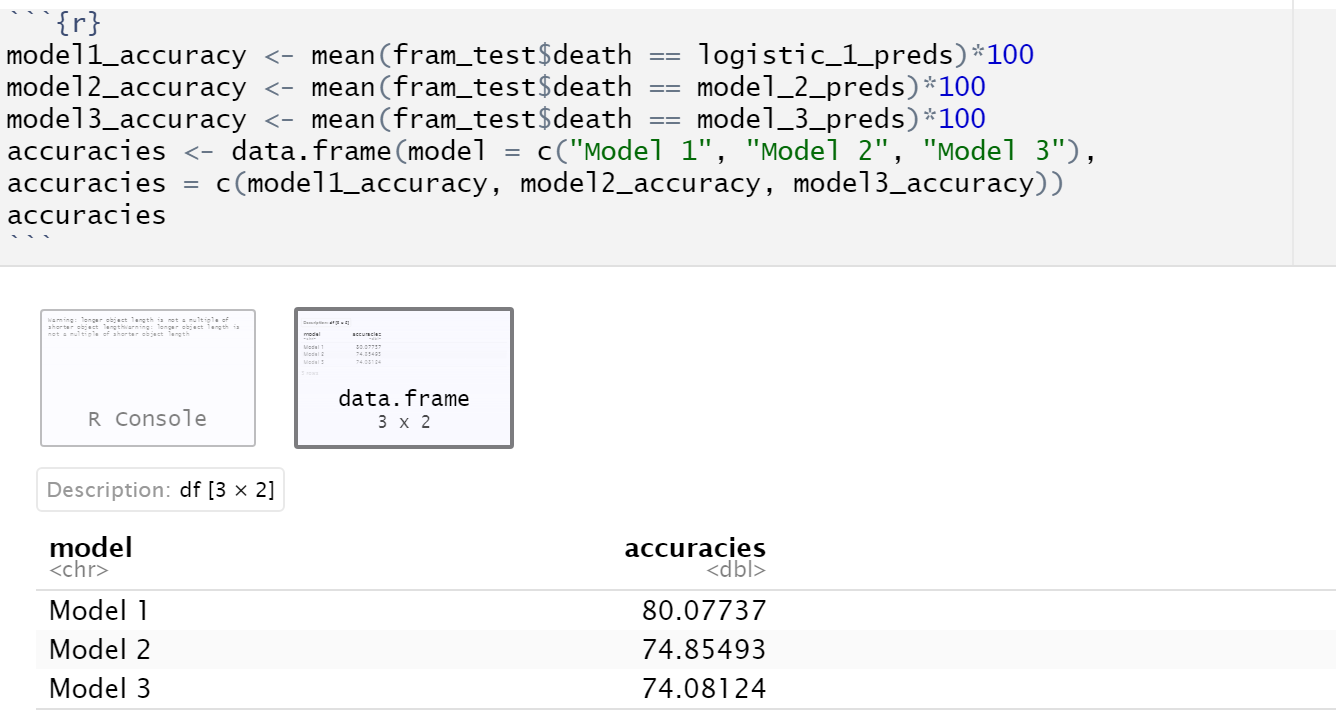
Like in the case of problem 3 – a, the model produced non-zero coefficients for all predictors. Once again, age has the highest coefficient, while chol2 has the lowest coefficient.











Model 3 is the plain logistic regression model. It is observed that the coefficients for all models are quire similar for the most part, barring some minor differences on the same scale, in which case the coefficients for the models 1 and 2 are closer. In addition, model 3 has the least accuracy of all the models, while Model 1 has the highest accuracy. It is observed that the model with L1 (lasso) Regularization performs the best.

Problem 3 – d:

I would expect the model with L1 penalty estimates to have a higher variance. This is because:

1. The l1 regularization penalty term is of a lower scale than that of the l2 regularization term. Considering this, by the bias-variance tradeoff, we expect the l1 penalty to have lesser impact on the increase of bias compared to l2.

2. The l1 regularization approach would encourage some coefficients to be zero, while the l2 regularization does not do that. Hence, L2 regularization also has more terms that contribute to the penalty.

Hence, the logistic regression model with l2 penalty estimates has lesser variance than a logistic regression model with the l1 penalty estimate.