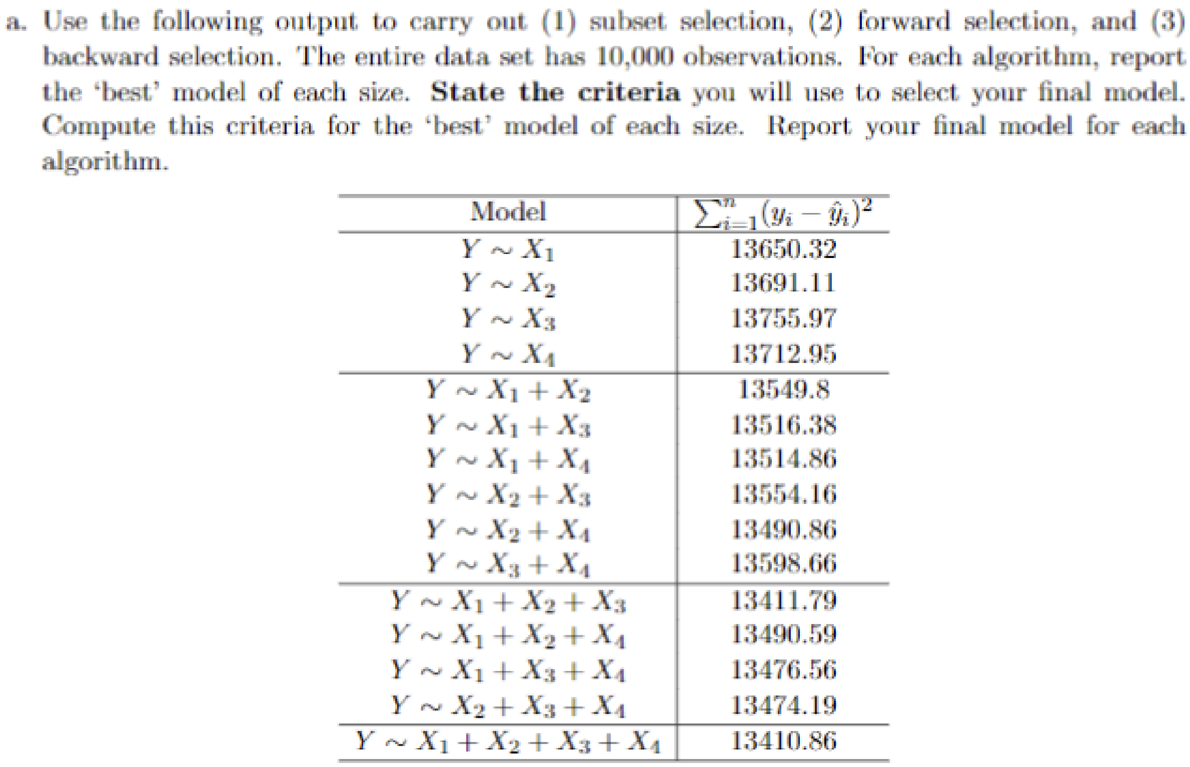
Keven Lin
851529703
Keven Lin

Question 1: Concept Review

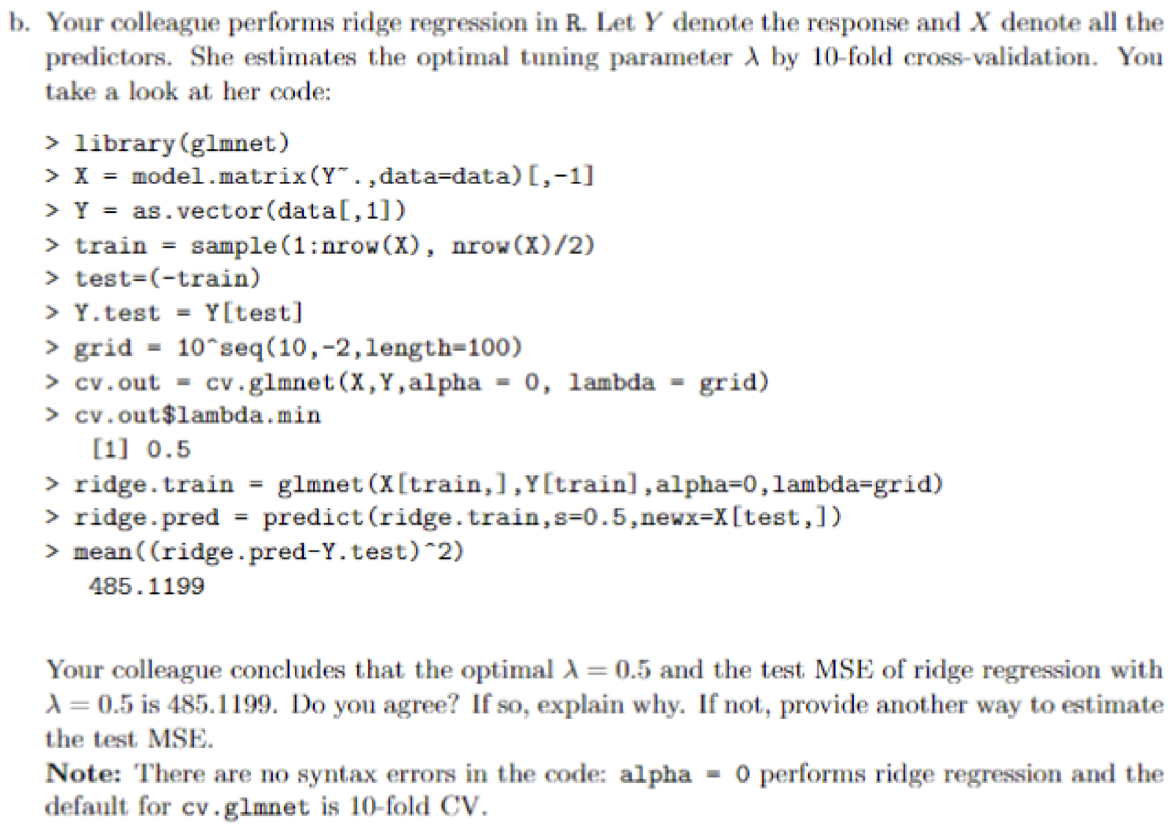


a. Use the following output to carry out (1) subset selection, (2) forward selection, and (3) backward selection. The entire data set has 10,000 observations. For each algorithm, report the 'best' model of each size. **State the criteria** you will use to select your final model. Compute this criteria for the 'best' model of each size. Report your final model for each algorithm.

Model	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$Y \sim X_1$	13650.32
$Y \sim X_2$	13691.11
$Y \sim X_3$	13755.97
$Y \sim X_4$	13712.95
$Y \sim X_1 + X_2$	13549.8
$Y \sim X_1 + X_3$	13516.38
$Y \sim X_1 + X_4$	13514.86
$Y \sim X_2 + X_3$	13554.16
$Y \sim X_2 + X_4$	13490.86
$Y \sim X_3 + X_4$	13598.66
$Y \sim X_1 + X_2 + X_3$	13411.79
$Y \sim X_1 + X_2 + X_4$	13490.59
$Y \sim X_1 + X_3 + X_4$	13476.56
$Y \sim X_2 + X_3 + X_4$	13474.19
$Y \sim X_1 + X_2 + X_3 + X_4$	13410.86

Reference excel sheet for calculation

	Subset	Forward	Backward	Criteria Value
M1	Y~X1	Y~X1	Y~X1+X2+X3+X4	AIC
M2	Y~X2+X4	Y~X2+X4	Y~X1+X2+X3+X4	AIC
M3	Y~X1+x2+X3	Y~X1+x2+X3	Y~X1+X2+X3+X4	AIC
M4	Y~X1+X2+X3+X4	Y~X1+X2+X3+X4	Y~X1+X2+X3+X4	AIC
Final Model	Y~X1+X2+X3+X4	Y~X1+X2+X3+X4	Y~X1+X2+X3+X4	AIC

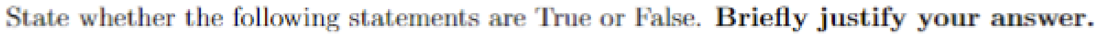


b. Your colleague performs ridge regression in R. Let Y denote the response and X denote all the predictors. She estimates the optimal tuning parameter λ by 10-fold cross-validation. You take a look at her code:

```
> library(glmnet)
> X = model.matrix(Y~.,data=data)[,-1]
> Y = as.vector(data[,1])
> train = sample(1:nrow(X), nrow(X)/2)
> test=(-train)
> Y.test = Y[test]
> grid = 10^seq(10,-2,length=100)
> cv.out = cv.glmnet(X,Y,alpha = 0, lambda = grid)
> cv.out$lambda.min
   [1] 0.5
> ridge.train = glmnet(X[train,],Y[train],alpha=0,lambda=grid)
> ridge.pred = predict(ridge.train,s=0.5,newx=X[test,])
> mean((ridge.pred-Y.test)^2)
   485.1199
```

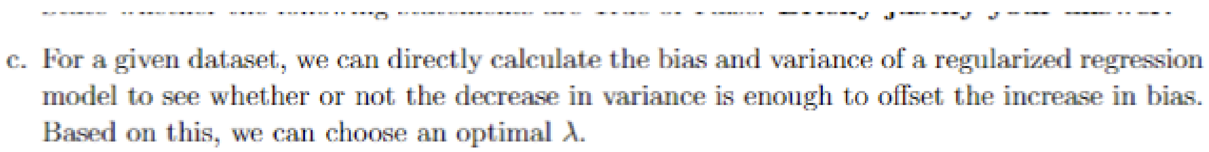
Your colleague concludes that the optimal $\lambda = 0.5$ and the test MSE of ridge regression with $\lambda = 0.5$ is 485.1199. Do you agree? If so, explain why. If not, provide another way to estimate the test MSE.

Note: There are no syntax errors in the code: `alpha = 0` performs ridge regression and the default for `cv.glmnet` is 10-fold CV.



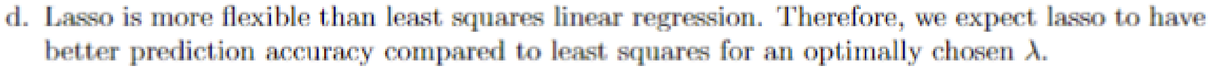
State whether the following statements are True or False. **Briefly justify your answer.**

Yes, she is correct. Her method of finding the optimal lambda is correct as she generates 100 values and uses glmnet to get a pool of lambda values. Afterward, she filters the data to find the smallest lambda value. And lastly the use predict function to generate the test MSE.



c. For a given dataset, we can directly calculate the bias and variance of a regularized regression model to see whether or not the decrease in variance is enough to offset the increase in bias. Based on this, we can choose an optimal λ.

False because you do not directly calculate it. The only way for you to find the optimal lambda is to perform glmnet function over a pool of values then choose the optimal lambda from that set.



d. Lasso is more flexible than least squares linear regression. Therefore, we expect lasso to have better prediction accuracy compared to least squares for an optimally chosen λ.

True, lasso and ridge are more accurate in prediction accuracy because they can optimized the lambda values.

Question 2: Bootstrap + Logistic Regression (20 points)

We will use a heart disease data set for this problem. To see information on the data and how to load it into R, check the `heart.R` file. You can download the data from Canvas or read the data from the website directly.

a. Fit a logistic regression model on the **entire** dataset with `chd` as the response and `age` as the only predictor. Report only the maximum likelihood estimates relevant to your model. Do not just copy/paste the summary output.

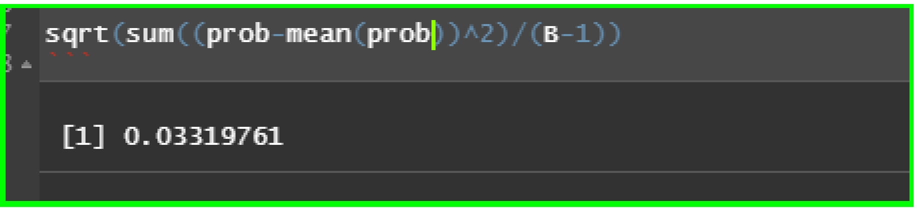
b. What is $\hat{P}(Y = 1|\text{age} = 50)$? In other words, what is the predicted probability that a 50-year-old individual has `chd`? Report that probability.

c. Bootstrap the standard error of $\hat{P}(Y = 1|\text{age} = 50)$. Report the standard error.

d. Based on your model in part (a), at what age does an individual have a roughly 50% of experiencing coronary heart disease (chd =1)?

a. b0= -3.521710, b1 = 0.064108

b. 0.4215753 aka 42% chance that a 50 year old has chd.



```
sqrt(sum((prob-mean(prob))^2)/(B-1))
```

```
[1] 0.03319761
```

c.

d. 55 years old

Question 3: Data application (10 points)

We will continue working with the heart disease dataset. To ensure we all get the same answers, split the dataset so that the first 100 observations are the test set and the remaining observations are the training set. The goal is to use logistic regression to carry out classification of `chd`. All other variables **except** family history will be used as our predictors. To save time, you may directly copy/paste raw R output.

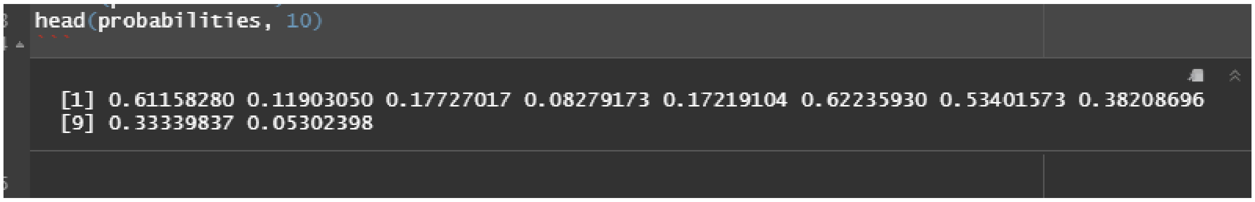
a. Fit a logistic regression model on the training set only. Report the first 10 predicted probabilities from your training set.

b. Implement logistic regression on the test set. What threshold will give us the smallest overall misclassification rate? Use that threshold to obtain the confusion matrix and report the overall misclassification rate.

c. What threshold would give us the smallest false positive rate (classifying someone with chd when they do not have it)? Use that threshold to obtain the confusion matrix and report the overall misclassification rate. How does this compare to your answer in part (b)?

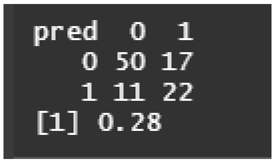
d. Give one compelling reason to justify why least square/regularized regression models are not appropriate for this setting.

a.



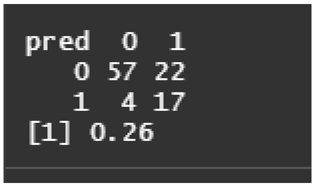
```
head(probabilities, 10)

 [1] 0.61158280 0.11903050 0.17727017 0.08279173 0.17219104 0.62235930 0.53401573 0.38208696
 [9] 0.33339837 0.05302398
```

b. Misclassification rate: 0.28



```
pred  0  1
   0 50 17
   1 11 22
[1] 0.28
```

c. .4 will produce the smallest false positive rate



```
pred  0  1
   0 57 22
   1  4 17
[1] 0.26
```

d. Least square regression requires an assumption that the data is linear. In this case, the data is not linear, thus we cannot use LSRM.