Midterm 2

**Question 1: Concept Review**

1. I think that my colleague’s approach to estimate the optimal λ for ridge regression using 10-fold cross validation is valid. However, I think the issue in her approach is the way the train and test sets are created. She uses the first half of the data as the training set and the second half as the test set. This approach could lead to biased results since there is not a random split of the data. I think to obtain a more reliable estimate of the test MSE, it is better to use a random split of the data, ensuring that both sets are representative of the entire dataset.
2. (a) Presence of a perfect correlation can lead to multicollinearity issues in least squares. When

two predictors are perfectly correlated, it becomes very difficult to estimate the individual coefficients for these predictors. As a result, the model can be overfit and have poor performance.

(b) Ridge regression is designed to address multicollinearity issues by adding a penalty term. The presence of perfect correlation is less likely to break down ridge regression as it can still provide coefficient estimates, although they may be biased towards one of the correlated predictors. Ridge regression can stabilize coefficient estimates in the presence of multicollinearity.

(c) If there are perfectly correlated predictors, the covariance matrix can become noninvertible. This singularity can make it impossible to calculate the inverse of the matrix, which might break down LDA.

(d) QDA relies on estimating class specific covariance matrices. If there is perfect correlation between predictors, it can cause the matrix to become noninvertible and could break down QDA.

1. Regularized models introduce bias by shrinking the coefficient estimates towards zero. This bias reduces the variance of the estimates. While linear regression tends to have higher variance when the number of predictors is large or when multicollinearity is present. When the true underlying relationship between predictors and the response is complex or has noise, linear regression can overfit the data, leading to a high variance in predictions. Regularized models, by reducing variance, can provide more stable predictions.

Lass and elastic net have a feature selection property. They can drive some coefficients to exactly zero, eliminating less relevant predictors from the model. Linear regression on the other hand, includes all predictors and if there are irrelevant predictors, it can introduce more noise into predictions.

Overall, regularized models can achieve a lower test MSE than linear regression in some situations because they trade off a controlled amount of bias for a significant reduction in variance, resulting in more stable, and better-regularized predictions. They are effective when dealing with high dimensional data or situations where feature selection and reduced overfitting are important.

1. ઠ0(x1) = x1(3.4 / 4) – (3.4 / 2\*4) + log(0.32)

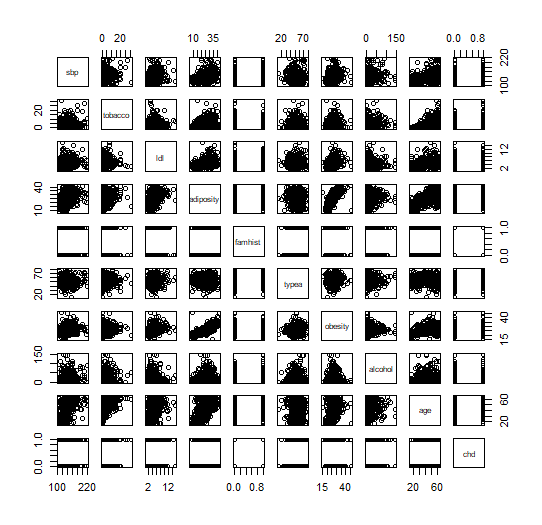
ઠ1(x1) = x1(5.1 / 4) – (5.1 / 2\* 4) + log(0.68)

If ઠ0(x1) > ઠ1(x1), then the test observation will be assigned to Y=0

If ઠ0(x1) ≤ ઠ1(x1), then the test observation will be assigned to Y=1

**Question 2: Coding**

1. Predicted probability = 0.4215756
2. Standard error = 0.02686035
3. 95% confidence interval is (0.3812336, 0.4678004)
4. First, I changed famhist so that it is not longer a categorical “present” or “absent”. I created a plot to see the correlation between the predictors.



Because there are many predictors with 0’s and 1’s to represent the categorical variables, this plot is kind of hard to interpret the relationship between tobacco and famhist.

I created a model with chd as the response and age+famhist as the response. The results show that famhist is a statistically significant predictor of chd.

A screen shot of a computer error

Description automatically generated

Following this, I created a model with chd as a the response and age+tobacco+famhist. The results show that tobacco is a statistically significant predictor of chd.

A screenshot of a computer error

Description automatically generated

Before creating the model with the interaction term, I made a chi-square table to see if tobacco and famhist had a significant association with each other.

A black background with white text

Description automatically generated

The results indicate that famhist and tobacco are not correlated as the p-value 0.3491 is not less than a significance level of 0.05.

Next, created a model with chd as the response and age+tobacco+famhist+tobacco\*famhist. The results show that tobacco and tobacco\*famhist are not statistically significant predictors of chd.

A screenshot of a computer screen

Description automatically generated

We can also look at the BIC values of the two models. The model without tobacco has a BIC of 525.0648. The model with age, famhist and tobacco has a BIC of 519. 9277. The model with tobacco and the interaction term has a BIC of 525.6417. The model without the interaction term has a better BIC and should be chosen over the others.

For further testing, another method we could use to check the significance of tobacco and an interaction term tobacco\*famhist on chd predictions would be leave one out cross validation.

Overall, the final model chd = age + famhist + tobacco is the best model.