Stats369 Test 2020
Hasnain Cheena
190411106
hche737

Question 1

a) Cross-validation is preferred to a train-test split when the dataset is small, as there is not much data and so separating off a test set is wasteful.
In comparison, a train-test split is preferable in the case when there is enough data and as it is slightly easier to code up in R than cross-validation.
b) Mean squared prediction error (MSPE) is a method of measuring the accuracy of a predictive model. It is found by computing predictions and calculating the average of the sum of squared residuals. The true MSPE of a predictive model is hard to estimate as the dataset would need to cover every possible combination effects which just doesn't happen in reality.
c) Apparent error is the residual variance from a fitted model. It is the residual sum of squares (RSS) divided by the number of observations (n) minus the number of predictors (p).
It is not a useful estimate for MSPE. This is because it is a biased estimate as the same data that was used to fit the model has been used to evaluate its performance.

Question 2

The columns outputted from the code will be:

Month	Inc	Trans	Disc

- The month column will contain the number of the month e.g. 12
- The Inc column will contain the total sum of item sales for each month in 2019.
- The Trans column will contain the number of sales transactions for each month in 2019.
- The Disc column will contain the average discount for each month in 2019.

Altogether there will be 12 rows; 1 for each month in 2019.

Question 3

a) If the number of predictors (p) is much greater than the number of observations (n) then variable selection is needed. Two methods that can address this issue are **Lasso regression and using an AIC-type penalty** (e.g. Penalised RSS).
The **advantages of using Lasso regression** over an AIC-type penalty is that Lasso regression relatively much faster to fit than AIC-type penalties and thus performing cross-validation to find the best model is much faster. Furthermore, Lasso regression performs both sparsity and shrinkage versus AIC-type penalties only do variable selection.
In comparison, **the advantages of using AIC-type penalty** over Lasso regression is that the coefficients remain unbiased meaning the magnitude of the coefficients has not be altered. They are either in or out of the model.
b) If the relationship between the response and predictors is non-linear then either **Harmonic models or splines** (typically cubic splines) can be used.

The **advantages of using a Harmonic model** compared to a spline is that Harmonic models can be extrapolated into the future for prediction very easily and they are not affected by extreme values/outliers.
In comparison, **the advantages of using splines** compared to harmonic models is that splines account for the year-on-year variation and can be fit with multiple predictor variables.

c) In a situation where the variance of the error term is non-constant, we can apply **bootstrapping or use a sandwich estimator**.
The **advantages of using a sandwich estimator** over bootstrapping is that it works well if you have a small datasets, if the model is a bit complex and is less computationally intensive than bootstrapping.
The **advantages of bootstrapping** over using a sandwich estimator is that the sandwich estimator is not as easy to understand because of the mathematics behind it and bootstrapping is much easier to code up in R than a sandwich estimator.

Question 4

a) Yes, I would expect the estimated error to be biased. This is because dataset used to find the 100 best variables is the same as the one used in the cross-validation process.
b) No, I would not expect the estimated error to be biased. This is because the dataset used to select the 100 variables is separate to the one used in the cross-validation process.
c) Yes, I would expect the estimated error to be biased. This is because the dataset used to find the best 100 variables (even if they are preselected) is the same one used in the cross-validation process.

Question 5

The ggplot commands that could have been used to make this plot are:

1. Ggplot(aes(x=GROSS_VEHICLE_MASS, y=CC_RATING))
    a. Map GROSS_VEHICLE_MASS to the x axis and CC_RATING to the y-axis
2. geom_point(alpha=1/20)
    a. Add points on the graph with some transparency (example given is alpha=1/20)
3. geom_smooth(method=lm, se=FALSE)
    a. Add a linear regression line with no error bands onto the scatterplot
4. facet_wrap(~flag2005)
    a. Subset the data into pre-2005 and post-2005 and then plot the two groups separately. Note that flag2005 is an example of a variable that can be used to do this.
5. labs(title='Vehicle size by vehicle mass in cars registered before and after 2005')
    a. Add the title specified to the graph

Question 6

Done on R markdown file.