Name: Chee Kitt Win
Student Number: 45589140
STAT3006 Assignment 1

Part a)

Breiman raises a concept he calls the Rashomon Effect or the multiplicity of good models. Explain what this means and its implications for both statistics and machine learning, ideally with examples (real or imagined). Additionally, if n is the sample size and p is the number of measured variables, discuss how this effect might change as n and/or p are increased.

The Rashomon Effect as described by Professor Breiman comes up in the context of model selection when there are a large number of different models[1] of the same class which all perform similarly well. In his paper, this is discussed primarily in the context of feature selection where we have a potentially large number of explanatory variables and want to choose say the *k* most important of these explanatory variables. This is typically done by selecting the subset of *k* explanatory variables which, when fitted to the data, minimizes some measure of error. Let us call this subset which has the lowest error, *s*. The problem occurs when a large number of the candidate subsets have an error so close to *s* that it is difficult to justify any one model being the clear winner.

My view is that the implications of this effect for statistics and machine learning depends on what the goal of our analysis is. If our goal is prediction, then the Rashomon Effect might not be an issue as long as we can get accurate predictions. Say for instance that we want to predict housing prices in a particular neighbourhood based on a number of explanatory variables. We fit a linear regression model based on data of a large sample of the houses in the neighbourhood. In this case, even if the Rashomon Effect appears, we do not care as long as we are able to accurately predict housing prices.

However, we have to be careful if we wish to interpret the results of our model. We cannot just look at our fitted multiple linear regression model and immediately conclude that the coefficients with the largest absolute values are indicative of the "most important explanatory variables". Perhaps some of the explanatory variables were number of rooms, number of beds and house size which are strongly correlated. Then we have multicollinearity, and the coefficients of the explanatory variables will have high standard errors. This high standard error in coefficients means that if we resampled the data from say a different set of houses from the same neighbourhood and fit the model again, the coefficients might change drastically, leading to what Professor Breiman calls an "unstable model".

With regards to the number of explanatory variables, *p*, using the same feature selection example in the first paragraph, the number of candidate subsets is *p* choose *k.* The Rashomon Effect is thus amplified when we increase the number of explanatory variables, *p*. On the other hand, an increased sample size, *n*, reduces this effect. This is because as we increase the sample size, *n*, the sampling variance is reduced. This means if we resampled our data from the same population, our model would not change that much. Interestingly, machine learning models such as neural networks are typically trained on large datasets with an incredibly large number of explanatory variables. The Rashomon Effect is in full force here, but is not an issue because neural networks have been designed purely to predict.

[1] Here I refer to model selection in a loose sense and include feature selection/choosing important explanatory variables as part of model selection.

Part b)

Breiman was a proponent of what he called "algorithmic culture", or what Efron (2020) calls "prediction models". On p201 of "The Two Cultures", Breiman made a statement about his perception as he returned to university: "Predictive accuracy on test sets is the criterion for how good the model is." Explain when you believe this is reasonable and when it is not, with examples.

I think it is indisputable that if we were truly able to measure the predictive power of a model, it would be the best metric to assess a model. After all, a model that can achieve 100% predictive accuracy on a response given a set of explanatory variables has surely perfectly captured the relationships between them. Unfortunately, measuring predictive accuracy on test sets is not the same as measuring the predictive accuracy on the whole population. Test sets are after all just a sample of the population. However, what this means is that if the test set is indeed a good representative of the population, then predictive accuracy on it is an excellent criterion for how good the model is, not only at prediction, but also at capturing the relationships between inputs and outputs.

Of course, I disagree that it is *the definitive criterion* for how "good" a model is as it does not measure how interpretable the model is, or how good it is a confirming a theory. However, even when it comes to prediction, sometimes test sets are unhelpful. I want to focus on the "data modelling culture" vs "algorithmic modelling culture" aspect of this question. My view on this aligns with that on Professor Cox in that I think our decision on which of the above approaches is better suited for a particular problem should by informed by what we know about the underlying process generating the data.

Take for example a problem such as analyzing the spread of the Covid-19 virus in Queensland. The winner in this case is clearly a data modelling approach. We understand the mechanism behind how the virus spreads, so we are able to create a model which captures the underlying relationships between the relevant variables. Crucially, although our model might perform worse than a machine learning model on a test set, we would find that it makes better predictions. Test set accuracy as a model validation criterion, fails because the test set is not at all representative of the population.

On the other hand, if, based on our knowledge of the data generating process and our sample, we are confident that our sample is a good representative of the population and is big enough, then we can go ahead and use a test set as the primary criteria for model performance.
Consider the earlier example on predicting housing prices. Here, we have a large sample size, and more importantly, based on what we know about houses, we can be confident that our sample of houses will likely be representative of the population of houses. Therefore, in this situation an algorithmic model would be suitable, and a test set would be a good criterion for model performance.

My stance here is that it is necessary to be informed by what we know of the domain in question before deciding on what our approach should be, and that we cannot choose a model just by looking purely at the data.

Part c)

Breiman describes an example project studying ozone levels. He made the following statement. "Then the problem was to construct a function f(x) such that for any future day and future predictor variables x for that day, f(x) is an accurate predictor of the next day's ozone level y." Explain and justify your views on whether this was the full scope of this problem. Also describe a data analysis problem in which prediction is not the primary goal and describe mathematically what the problem is in this case.

The main aim of the ozone project was to predict ozone levels in order to accurately issue warnings 12 hours in advance. The project ultimately failed due to the false alarm rate of the predictor being too high. Given the available data, perhaps the scope of the problem could have been widened to include probabilistic models rather than a purely predictive one. Instead of a pure predictor based on regression, perhaps a prior distribution model could have been constructed based on historical data, and then updated in real time based on the predictor variables measured in the past few days. By doing this, it would have been possible to make announcements every night regarding the likelihood of the next day's ozone levels being in a particular range rather than having to choose between issuing a warning or not issuing a warning the night before. I also note that Professor Breiman states that the first 5 years of data were used as the training set and the last 2 years as the test set. This assumes that the data for all 7 years were all representative of each other. But if somewhere down the line, due to advances in technology or just climate changes, the data started looking different, then the trained model would no longer be useful. Bayesian models on the other hand, can always be updated with a new posterior given the most recent data.

An example of a data analysis problem in which prediction is not the primary goal is a study by Caiazza, Andrew, and Alberto (2010) to investigate whether the targets in cross-border bank mergers and acquisitions are materially different from those banks targeted in domestic merger and acquisition deals. They propose that the null hypothesis that the banks are not materially different, and the alternative hypothesis is that they are. They then construct a binomial model for whether or not a bank is a merger and acquisition target. A multinomial model distinguishes between (i) targets in domestic operations, (ii) targets in cross-border operations and (iii) non-targets. After that, they justify their model design choices based on their domain knowledge, and after fitting the data to their model, give a descriptive write-up of insights gleaned by their model. The final conclusion is also descriptive. Regardless of the quality of experimental design, the approach used is completely different. The model used is a very simple one, but very interpretable. The goal of the problem is descriptive and seeks to gain insights based on the assumption that the model holds true. The model is validated not based on test sets or residuals but by domain knowledge arguments. Despite that, this is still a data analysis problem, and my takeaway here is that models can be used for a wide variety of purposes, and for a model to be "good" requires the consideration of both statistical and domain specific aspects.

REFERENCES

Breiman, L.(2001). Statistical Modeling: The Two Cultures. *Statistical Science,16*(3),199-231.

Caiazza, S., Andrew C., and Alberto F. P.(2010). What Do Foreigners Want? Evidence From Targets In Bank Cross-Border M&As. *Economics And Statistics Discussion Paper,058*(11),1-31.