STA 3180 Statistical Modelling: Model Selection

1. Explain the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

Solution: The Akaike Information Criterion (AIC) is a measure of the relative quality of a statistical model for a given set of data. It is based on the maximization of the likelihood function and takes into account the number of parameters in the model. The AIC penalizes models with more parameters, so it is useful for model selection. The Bayesian Information Criterion (BIC) is also a measure of the relative quality of a statistical model for a given set of data. It is based on the maximization of the posterior probability of the model and takes into account the number of parameters in the model. The BIC penalizes models with more parameters, so it is also useful for model selection. The main difference between the two is that the AIC is based on the maximization of the likelihood function while the BIC is based on the maximization of the posterior probability of the model.

2. Describe the steps involved in the stepwise regression procedure.

Solution: Stepwise regression is a method of model selection that involves adding or removing variables from a model in a stepwise fashion. The steps involved in the stepwise regression procedure are as follows:

1. Start with an empty model.
2. Add the variable that has the highest correlation with the response variable.
3. Calculate the adjusted R-squared value for the model.
4. If the adjusted R-squared value increases, keep the variable in the model.
5. If the adjusted R-squared value decreases, remove the variable from the model.
6. Repeat steps 2-5 until no further improvement in the adjusted R-squared value is observed.

3. Explain the concept of cross-validation and its importance in model selection.

Solution: Cross-validation is a technique used to evaluate the performance of a model on unseen data. It involves splitting the dataset into two parts: a training set and a test set. The model is trained on the training set and then evaluated on the test set. This allows us to assess the model's performance on unseen data and helps us to select the best model for our data. Cross-validation is important in model selection because it allows us to assess the performance of different models on unseen data and select the best one.