STA 3180 Statistical Modelling: Resampling

Lecture Notes on Resampling for STA 3180 Statistical Modelling

Resampling is a statistical technique used to assess the accuracy of a model or estimate by using the same data multiple times. It is a powerful tool for evaluating the performance of a model or estimator and for making decisions about the reliability of results. Resampling can be used to assess the accuracy of a model or estimate, to compare different models or estimators, and to determine the optimal parameters for a model or estimator.

Key Concepts:

• Resampling is a statistical technique used to assess the accuracy of a model or estimate by using the same data multiple times.

• Resampling can be used to assess the accuracy of a model or estimate, to compare different models or estimators, and to determine the optimal parameters for a model or estimator.

• Resampling techniques include bootstrapping, cross-validation, and jackknifing.

Definitions:

• Bootstrapping: Bootstrapping is a resampling technique where a sample is drawn with replacement from the original dataset. This allows for the estimation of the variability of a statistic from the same data.

• Cross-validation: Cross-validation is a resampling technique where a dataset is divided into two parts, a training set and a test set. The model is trained on the training set and then tested on the test set. This allows for the evaluation of the model's performance on unseen data.

• Jackknifing: Jackknifing is a resampling technique where a sample is drawn without replacement from the original dataset. This allows for the estimation of the bias of a statistic from the same data.

Rules:

• When using resampling techniques, it is important to ensure that the data is randomly sampled and that the samples are independent.

• When using bootstrapping, it is important to ensure that the sample size is large enough to accurately estimate the variability of the statistic.

• When using cross-validation, it is important to ensure that the training set and test set are of similar size and that the model is evaluated on unseen data.

• When using jackknifing, it is important to ensure that the sample size is large enough to accurately estimate the bias of the statistic.

Examples:

• Bootstrapping: Suppose we have a dataset of 100 observations and we want to estimate the variability of the mean. We can draw a sample of 100 observations with replacement from the original dataset and calculate the mean. We can repeat this process 1000 times and calculate the mean of the 1000 means. This will give us an estimate of the variability of the mean.

• Cross-validation: Suppose we have a dataset of 100 observations and we want to evaluate the performance of a linear regression model. We can divide the dataset into a training set of 80 observations and a test set of 20 observations. We can then fit the linear regression model to the training set and evaluate its performance on the test set.

• Jackknifing: Suppose we have a dataset of 100 observations and we want to estimate the bias of the median. We can draw a sample of 100 observations without replacement from the original dataset and calculate the median. We can repeat this process 1000 times and calculate the mean of the 1000 medians. This will give us an estimate of the bias of the median.