1 Estimate coefficient and Variance of Coefficient in Linear Regression

1.1 Best Estimate of Coefficient

Simple derviation for the case of linear regression:

y=a+bx+u is the true relationship, also known as population regression line. Here, u is random error.

Then,

$y_i = \hat{a} + \hat{b}x_i$, where $\hat{a}, \hat{b}$ are the estimated coefficients.

and $\bar{y} = \hat{a} + \hat{b}\bar{x}$, where $\bar{y}, \bar{x}$ are the mean of y and x, respectively.

$y_i - \bar{y} = \hat{b}(x_i - \bar{x})$

or, $(x_i - \bar{x})(y_i - \bar{y}) = \hat{b}(x_i - \bar{x})^2$ (I don't know why I should multiply the equation by $(x_i - \bar{x})$, but the standard formula has it this way).

Since we will be dealing with a dataset with many events, it is more meaningful to sum over all data points.

$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \hat{b} \sum_i (x_i - \bar{x})^2$

or, $\hat{b} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$ is the best estimate of the coefficient b.

1.2 Variance of Coefficient

$\text{var}(\hat{b}) = \text{var}(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2})$

Variance of constants = 0. So, we will only keep the non-constant terms. Expand y_i and $\bar{y}$ in terms of x_i and $\bar{x}$.

$\text{Var}(\hat{b}) = (\frac{1}{\sum_i (x_i - \bar{x})^2})^2 \, Var(\sum_i (x_i - \bar{x}) u_i)$ Online websites say that only u is random variable, and we know that $Var(kX) = k^2 \, Var(X)$, so, we get:

$\text{Var}(\hat{b}) = \frac{\sum_i (x_i - \bar{x})^2}{(\sum_i (x_i - \bar{x}))^2} \, Var(u_i)$

$= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$

2 Estimate coefficient and Variance using Matrix Approach

Root and other software use matrices to estimate best values for coefficients and their corresponding variances.

See the following link for linear regression:

https : //goo.gl/yTDs7Z

Basically, $\hat{b} = \frac{X^T Y}{(X^T X)}$

and $\text{Var}(\hat{b}) = \frac{\sigma^2}{X^T X}$

Where X is $(n \, X \, 1)$ matrix with n datapoints, Y is $1 \, X \, n$ matrix with n labels. See Minuit manual to understand how root uses Hessian and covariance matrices to determine errors in the estimation of the coefficients.

3 Cross-Validation

In cross-validation, you divide the training dataset into n subsets (in sklearn, you specify the value of n via the parameter 'cv'). Then, you fit the machine learning classifier on $(n-1)$ subsets, and evaluate the fit on the nth subset. There are n possible ways of selecting $(n-1)$ datasets, so this validation is done n times, each time with a different set of $(n-1)$ datasets. The evaluation is done by computing MSE (mean square error) between predicted and actual labels of the nth dataset.

For each validation, you get an MSE. So, at the end of cross validation, you get n MSE values. You can compute the mean and standard deviation of these MSEs in sklearn. The smaller the mean and std deviation, the better is the model fit to your data.

4 Bootstrapping

Bootstrapping example: assume we are interested in the average (or mean) height of people worldwide. We cannot measure all the people in the global population, so instead we sample only a tiny part of it, and measure that. Assume the sample is of size N; that is, we measure the heights of N individuals. From that single sample, only one estimate of the mean can be obtained. In order to reason about the population, we need some sense of the variability of the mean that we have computed. The simplest bootstrap method involves taking the original data set of N heights, and, using a computer, sampling from it to form a new sample (called a 'resample' or bootstrap sample) that is also of size N. The bootstrap sample is taken from the original by using sampling with replacement (e.g. we might 'resample' 5 times from [1,2,3,4,5] and get [2,5,4,4,1]), so, assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original "real" sample. This process is repeated a large number of times (typically 1,000 or 10,000 times), and for each of these bootstrap samples we compute its mean (each of these are called bootstrap estimates). We now have a histogram of bootstrap means. This provides an estimate of the shape of the distribution of the mean from which we can answer questions about how much the mean varies. (The method here, described for the mean, can be applied to almost any other statistic or estimator.)

In Bootstrapping, the main assumption is that the sample of N data points is an accurate representation of the population. One must understand that Bootstrapping does not give any new information other than what is contained in the sample itself. The power of this method is that resampling over that sample on a large enough scale reveals the sampling distribution of the statistic at issue, for example the mean.