IB120/201 - Lab 10

Regression Analysis

Due Date: April 10, 2020

University of California, Berkeley *GSI: Naveed Ziari*

In this lab, we will learn about linear regression. This is a fundamental tool in computational biology research and the foundation of more advanced machine learning techniques.

Background

Regression

Regression is a tool in statistical analysis whereby one seeks to find a relationship between an independent variable (or variables) and a dependent variable. Once such a relationship is established, one can make inferences about a certain statistic when the dependent variable is unavailable.

Linear Regression

The most commonplace form of regression is *linear*. This means that we try to find a first-degree polynomial (AKA a straight line) that fits the data best. Typically, the line of best fit is determined through by minimizing the sum of squared residuals, which are the differences in values between the line and the data point. Using linear algebra, we can extend this notion to include multiple variables

Relationship to Machine Learning

Learning about regression analysis provides a foundation to supervised machine learning. In supervised ML, one trains a model with a *training set* and uses that to make predictions on the *test set*. For instance, let us assume that we have data on housing prices and other various parameters of the homes for sale (e.g. size, number of floors, proximity to certain amenities). We perform regression analysis on the testing set and establish a relationship between the parameters and the price. Now we have a relationship between these attributes and the price, so we can make prudent judgments on how much to put the house up for sale.

Questions

1. Load the dataset `iris`
   (a) Perform a linear regression where you aim to predict the pedal width from its pedal length
   (b) Based off of your results, what would be the predicted pedal width of the iris if its pedal length is 2.3?
   (c) Plot a histogram of the residuals of this analysis and perform a Shapiro-Wilk normality test
2. Load the dataset `trees`
   (a) Perform a linear regression where you aim to predict the volume from both its girth and height
   (b) Try seeing if it is more prudent use only one parameter (single variable regression) instead of both girth and height, and prove it by using the appropriate test
   (c) Find the 95% confidence intervals for the regression coefficients. You may want to use bootstrapping to derive these values.

3. Load the dataset `beavers`

   (a) The last column `beaver1$activ` is a binary response variable. Perform a logistic regression to predict activity from day, time, and temperature.

   (b) Find out if there is a better predictor with less variables than all 3 combined. How can you tell if this is a better predictor?

4. What does the F-statistic tell you?

5. What does the AIC inform on?

6. How is logistic regression an extension of linear regression?

7. **BONUS:** Why is it commonplace to take the sum of **squared** residuals instead of just the residual (not squared)? What advantage(s) does that have?