Regression Modeling

R Brown Bag Series #4

Gary R Seamans

The MITRE Corporation

March 29, 2018

Agenda

- Overview
- Demonstrations
- Questions

Overview

Statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables.(*Regression Analysis - Wikipedia*) Wikipedia is not an authoritative source, but this is a reasonable operating definition of regression modeling.

Harrell 2001 is, with well over 8k citations, a very authoritative source for regression modeling. A less expensive (read free) alternative is (Harrell 2013) which is a version of the book Dr. Harrell uses to teach regression modeling at Vanderbilt. A copy of the free version is included in the download material.

Overview contd.

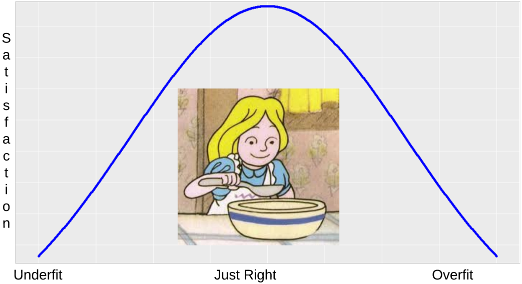
What kinds of regression are there? This article from R-bloggers describes 15 different types of regression. Which one to use?

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Quantile Regression
- Ridge Regression
- Lasso Regression
- Elastic Regression
- Principal Component Regression
- Partial Least Square Retgression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Quasi-Poisson Regression
- Cox Regression

Fitting

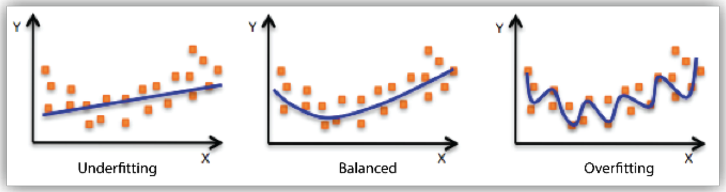
Two of the pitfalls to avoid are:

- Overfitting
- Underfitting



Fitting Continued

Fitting graphic from the R-bloggers post(*15 Types of Regression You Should Know | R-Bloggers*).



Additional Terms

Additional Regression Terminology.

- Outliers - can unduly influence the results
- Multicollinearity - highly correlated independent variables
- Heteroscedasticity - when a dependent variable's value is not equal across values of an independent variable

Galton's Data

In this example we'll look at Galton's data. Sir Francis Galton was an 18th century statistician, among other things, and his data on the relative heights of children and their parents is still relevant today.

There are a number of ways to get the Galton dataset, one is to download the R package **UsingR**. Typing **data()** in an R console will list all of the available datasets. You'll probably be surprised at how many come with the default installation.

To load the Galton data you would use **data(galton)**, more correctly you would create a *promise* that the Galton data will be loaded the first time it is referenced. Typing **summary(galton)** will cause the Galton data to be loaded and display a set of summary statistics.

Credit Card Default

Credit card default is a big problem for banks. This example of multivariate regression shows how you might go about predicting who will default.

Questions

References I

- *15 Types of Regression You Should Know | R-Bloggers*. https://www.r-bloggers.com/15-types-of-regression-you-should-know.
- Harrell, FE (2013). "Regression Modeling Strategies". In: *as implemented in R package 'rms' version* 3.3.
- Harrell, Frank E (2001). "Ordinal Logistic Regression". In: *Regression Modeling Strategies*. Springer, pp. 331–343.
- *Regression Analysis - Wikipedia*. https://en.wikipedia.org/wiki/Regression_analysis.