1 Regression Explained

1.1 Linear vs. Non-Linear Regression Models

One might think that linear models product equations that are straight lines and non-linear models can model curvature. This pre-disposition is misinformed. Both linear and non-linear models can capture curvature. Linear models are restricted to polynomial features. An example of a linear model is illustrated below:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2$$

Non-linear models include non-linear functions like sin, cos, ln, exp, etc. For example:

$$y = \beta_0 \cdot sin(x_1 + \beta_1) + \beta_2 cos(x_2 + \beta_3)$$

1.2 Regression Metrics

We use a set of metrics to attempt to understand the relationships between the independent variables and the dependent variable. These metrics are as follows:

R-Squared (R^2)

The R^2 value, or coefficient of determination, is a measure that evaluates the quality of a linear regression model. The formula for the coefficient of determination can be expressed in terms of residual sum of squares (SS_{res}) and the total sum of squares (SS_{tot}):

$$SS_{res} = \sum_{i=1}^{N} (y_i - \hat{y_i})^2$$

$$SS_{tot} = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

An accurate model has a near 0 SS_{res} and a resulting R^2 value approaching 0. A weak model has a large SS_{res} and an R^2 value near 1.

Pearson Product-Moment Correlation (r-Value)

The r value is the ratio of the covariance of the variable pair to the products of the standard deviations of the variable pair.

$$COV(x, y) = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{N}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N}}, s_y = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \overline{y})^2}{N}}$$

$$r_{x,y} = \frac{COV(x, y)}{s_x \cdot s_y}$$

$$r_{x,y} = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \overline{y})^2}}$$

This metric measures the strength of the linear correlation between two IID variables. r values that are closer to 1 indicate a positive linear relationship, r values close to 0 indicate that a weak or no linear relationship exists, and r values close to -1 indicate that a negative linear relationship exists. Creating a matrix of r values pertaining to the various independent variables is a good way to quickly discover which variables are linearly correlated with the output variable.

Standard Error

The standard error of a regression tells us how far the data points are from the regression line on average. In other words, the standard error delineates how precise the model's predictions are in units of the target variable. This metric is particularly attractive because it is valid for both linear and non-linear regression models. The standard error for a set of predictions and target values is:

$$SE = \hat{\sigma^2} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y_i})^2}$$

We use $N-2$ adjustment instead of N because we estimated two parameters to determine the regression model.

T-Statistic and P-Value

The t-statistic is a statistic that we compute when we are trying to compare the means of two samples. Using the t-statistic we can compute the p-value which is the likelihood that we osberved this data in a world in which the null hypothesis is true. In the context of linear regression we use the t-statistic and the p-value to determine if there is a statistically significant difference between a model with a slope of 0 and the observed data. If the difference is statistically significant then we reject the null hypothesis and determine that the two variables are not linearly independent. To calculate the $N-2$ d.o.f. t-statistic we use the following equation:

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}}$$

When comparing two variables and testing for linearity we assume a linear relationship: $y = \beta_0 + \beta_1 \cdot x$. $\hat{\beta}_0$ and $\hat{\beta}_1$ are our estimates of the coefficients β_0 and β_1 obtained by minimizing the squared error between the predictions and the target values. Conveniently, we obtain the following equations expressed in terms of S_{xy} and S_{xx}:

$$S_{xy} = \sum_{i=1}^{N} x_i \cdot y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}$$

$$S_{xx} = \sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}$$

$$\hat{\beta}_1 = r_{x,y} \cdot \frac{s_x}{s_y} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \cdot \overline{x}$$

From the t-statistic we can compute the p-value and threshold on some pre-determined cutoff. We can also threshold the t-statistic itself to see if we should reject the null hypothesis or not.