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.

2 Dependence Tests

2.1 Numerical-Numerical

T-Test

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. Outside of linear regression the t-statistic is used to determine if the difference between a sample mean and the mean of the population is statistically significant. To compute this statistic we can use the following equation:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\overline{x}$ is the sample mean, μ is the population mean, s is the standard deviation, and n is the number of observations. This statistic has $N-1$ degrees of freedom. 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. In this case, the null hypothesis is that the slope of the linear relationship between the independent and dependent variable is equal to 0. To calculate the $N-2$ degree of freedom 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.

2.2 Categorical-Categorical

Chi-Squared Test

The Pearson's Chi-Squared test is a statistical test with two applications. The first application is called a *goodness of fit* test. This test is used to test if sample data conforms to a distribution from a certain population. I.e., this test determines if your sample data represents the data you would expect to find in the actual population. The null hypothesis for the chi-squared goodness of fit test is that the data comes from the defined distribution. The alternate hypothesis is, of course, that the data did not come from that distribution. The formula for computing the chi-squared value is:

$$\chi_c^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

O_i represents each categorical value and E_i represents the expected categorical value. Use the chi-squared goodness of fit test when the sampling method is simple random sampling, the variable under study is categorical, and the expected value of the number of each sample observations in each level of the variable is at least 5. Note that the degrees of freedom for this test is equal to $N - 1$.

Examples:

1. https://stattrek.com/chi-square-test/goodness-of-fit.aspx
2. https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/pearson-s-chi-square-test-goodness-of-fit

We can also use the chi-squared distribution to perform a chi-square test for independence. We perform this test to compare two categorical variables from a single population to determine if there is significant association between the two variables. The null hypothesis for this test is that the two categorical variables are independent. Thus, the alternate hypothesis is that they are associated with each other. We use this test when the sampling method is simple random sampling, the variables in question are categorical, and the expected frequency count for each cell in the *contingency table* is at least 5. A contingency table is a table summarizing the frequencies of two categorical variables:

	Apples	Oranges	Pears
Males	10	13	2
Females	5	9	5

The first step is to compute the degrees of freedom. The degrees of freedom will later be used in conjunction with the chi-squared value to determine the p-value.

$$DF = (N_r - 1) \cdot (N_c - 1)$$

N_r is the number of levels for the first categorical variable and N_c is the number of levels for the second categorical variable. After computing the degrees of freedom, the expected frequencies for every level of the first and second categorical variables must be computed:

$$E_{r,c} = \frac{(n_r \cdot n_c)}{N}$$

n_r is the total number of r-th level sample observations of the first categorical variable and n_c is the total number of c-th level sample observations of the second categorical variable. N is the sample size. After these values are computed for all r and c, the chi-squared statistic can be computed:

$$\chi^2 = \sum_{r,c} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

The observed frequency at r and c, $E_{r,c}$, is the total number of observations in which both r and c were observed simultaneously. After computing the chi-squared value we can use this value in conjunction with the number of degrees of freedom to determine the p-value. From the p-value we can then decide if we should reject or accept the null hypothesis.

2.3 Numerical-Categorical

ANOVA

Typically, we use a one way ANalysis Of VAriance (ANOVA) to determine if the difference between the means of two or more groups is statistically significant. To do this, we compute the F-statistic. The F-statistic is the ratio of the variance of the between group means to the mean of the within group variances. If the variance between the samples is greater than the variance within the samples then the samples are not drawn from the same population. The null hypothesis for an ANOVA is that all of the population means are equal. The alternative hypothesis is that at least one of the means is different. We can use an ANOVA when the populations from which the samples were obtained are approximately normal, the samples are independent, and the variances of the populations are equal. To conduct an ANOVA, we must first determine the grand mean, x_{GM}, of the samples, the sum of squares total, $SS(T)$, the within group variation, $SS(W)$, and the between group variation, $SS(B)$:

$$\overline{x}_{GM} = \sum_{i=1}^{N} \frac{x_i}{N}$$

The grand mean is the mean of all of the elements from all of the samples mixed into one group.

$$SS(T) = \sum_{i=1}^{N} (x_i - \overline{x}_{GM})^2$$

$$DF(T) = M \cdot N - 1$$

The total variation is the variation of all of the samples in this collective group. The number of degrees of freedom for this calculation is equal to the number of samples, M, times the total number of elements in each sample, N, minus 1. I.e., $DF = M \cdot N - 1$. This is because if we knew the grand mean and we knew all of the data points except for 1 we could figure out the last one.

$$SS(W) = \sum_{i,c} (x_{i,c} - \overline{x}_c)^2$$

$$DF(W) = M \cdot (N - 1)$$

The within group variation is the variation that exists within each sample. For each sample group, we have $N - 1$ degrees of freedom because if we know the means of each group and $N - 1$ of the elements in each group, we could determine the final element. If we have M groups, we have $M \cdot (N - 1)$ degrees of freedom. The within group variation is calculated for each sample, c, and aggregated.

$$SS(B) = \sum_{c} N_c \cdot (\overline{x}_c - \overline{x}_{GM})^2$$

$$DF(B) = M - 1$$

The sum of squares between is the total variation due to the differences between the means. To compute this, we sum over all of the squared differences between the means of each sample and the grand mean N_c times for each sample group, c. For this calculation, assuming we know the mean of means, we can calculate one of the sample means if we have the other two sample means. Thus, the number of degrees of freedom in this calculation is equal to $M - 1$. Given the above metrics, an interesting relationship occurs:

$$SS(T) = SS(W) + SS(B)$$

Thus, the total variation in the data can be described as the sum of the variation within each of these groups plus the sum of the variation between the groups. This is useful because it shows that we can split up the total variation of the samples into the variations between the groups and the variation within the groups. The degrees of freedom of the calculations of the variations also add up:

$$DF(W) + DF(B) = M \cdot (N - 1) + (M - 1) = M \cdot N - M + M - 1 = M \cdot N - 1 = DF(T)$$

Using the variations defined above, we can compute the F-statistic:

$$F = \frac{\frac{SS(B)}{DF(B)}}{\frac{SS(W)}{DF(W)}}$$

If the numerator is greater than the denominator then the variation present in the data is due primarily to the variation between the groups. Thus, a large F-statistic indicates that the considered samples do not have the same mean. Conversely, an F-statistic less than 1 shows that the variation of the data is explained more so by the variation within each sample. Thus, it is more likely that the samples have the same mean. Using the F-statistic we can compute the p-value. From the p-value and our threshold value (typically 0.05), we can determine if the observed results are statistically significant. Note that the F table is dependent on both the degrees of freedom of the numerator ($DF(B)$) and the denominator ($DF(W)$).
When comparing a numerical and categorical variable, we split the numerical variable up into the various categories delineated by the unique values of the categorical variable. Then, to determine if the numerical and categorical variable are dependent, we conduct an ANOVA.

Explanations:

1. https://people.richland.edu/james/lecture/m113/anova.html
2. https://www.khanacademy.org/math/statistics-probability/analysis-of-variance-anova-library/

3 Feature Selection

3.1 SciKit Learn

SciKit Learn allows us to quickly select features based on a wide variety of univariate statistical tests. We can remove all but the k highest scoring features or all but the features in a pre-defined user-specified percentage of features subject to a statistical test (e.g. χ^2).

4 Model Validation

4.1 Performance Metrics

4.2 Multi-Class Classification

Using the softmax function we can generate probabilistic predictions. We can then choose to penalize these predictions in any way we see fit. Naively, we can consider the accuracy calculated by taking the sum of all of the correct predictions and dividing by the total number of predictions. Unfortunately, this metric does not make distinctions between classes. A correct answer for 'Class A' is weighted the same as a correct answer for 'Class B' and 'Class C'. But what if there are 100 'Class A' targets and only 5 'Class B' and 'Class C' targets? A model that never predicted 'Class C' or 'Class B' would do well.

4.2.1 Micro Multi-Class Accuracy

To compute this accuracy we calculate the total number accurate predictions for each class and then divide by the total number of false predictions for each class.

4.2.2 Micro vs. Macro Metrics

If we have a dataset with balanced data then the accuracy metric that we use isn't entirely relevant. Since the target variable has a similar number of elements for each category biasing towards the categories with more elements doesn't have much of an impact.

If our dataset has a significant imbalance in the target variable then we have to decide whether we want a model that is biased towards the most populated targets or whether we want a model that is biased to the least populated targets (and as a result, is not biased towards the most populated targets). This choice is situational and is dependent not on the data, but on the context surrounding the data.

If we wish to bias our model to more populated targets we choose a micro metric. Conversely, if we wish to bias our model to less populated targets we choose a macro metric.

4.2.3