Project 1: Multivariate Linear Regression

CS425

Samuel Steinberg

September 25th, 2019

Description:

This paper analyzes multivariate linear regression in machine learning, which is a method by which to measure the degree that one or more independent variables contribute to the dependent variable and thus needing multiple coefficients to be determined. The problem in this case was to determine miles per gallon (MPG) of automobiles depending on their other attributes. The database being tested was a data file called *auto-mpg.data* which consists of 398 automobile attribute data entries. Also given for the project was *auto-mpg.names*, which includes the source of the dataset along with the characteristics of the attributes (which came in the form of both continuous and discrete multi-valued). The data was analyzed with both standardization and without it.

Data Pre-Processing:

Data was given in the form of the automobile attributes:

1. MPG: Continuous value, multivariate regression performed to predict this value based off the other attributes
2. Cylinders: Multi-valued discrete value, number of cylinders automobile has
3. Displacement: Continuous value, amount of displacement
4. Horsepower: Continuous value, amount of horsepower
5. Weight: Continuous value, weight of vehicle
6. Acceleration: Continuous value, acceleration of vehicle
7. Model Year: multi-valued discrete, production year of automobile
8. Origin: multi-valued discrete value
9. Name of Car: String, brand and name

The first step taken was to clean the data; the name of the car was removed due to not being significant or numeric, and six missing horsepower values were replaced with the mean value (mean imputation). Next, data was separated into test, validation, and learning sets. Of the 398 data entries, 50 were used for testing and validation, and the remaining were used for learning. Then, the data was placed into a matrix. The MPG was removed from the matrix (since the other attributes needed to predict this value) and placed into a separate single-columned matrix and was replaced with a column of 1's to get the Y-intercept.

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & & \\ 1 & x_1^N & x_2^N & \cdots & x_d^N \end{bmatrix}$$

Figure 1: Input Variable Matrix: Each x represents an input variable (each column is one set of variables)

In Figure 1 above, the input variables are the attributes for the car (besides MPG and name, of course). This matrix will be used to determine the variable coefficients.

Solution Description:

After data pre-processing is complete, it is possible to begin computation. Along with the *X*-matrix in Figure 1, the MPG matrix (denoted as *r*) is created and assumed to be a linear function (weighted sum) of several input variables:

$$\boldsymbol{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

Figure 2: r-matrix that acts as a weighted sum of several input variables

Thus, a weight matrix *w* (Figure 3) is needed in the calculation of coefficients.

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d, \end{bmatrix}$$

Figure 3: Weight matrix for variables

Normal equations can be derived with respect the parameters (Figure 4), re-written as in Figure 5, finally allowing the parameters to be solved for (Figure 6) by solving for the coefficients:

$$\begin{aligned}
\sum_t r^t &= N w_0 + w_1 \sum_t x_1^t + w_2 \sum_t x_2^t + \cdots + w_d \sum_t x_d^t \\
\sum_t x_1^t r^t &= w_0 \sum_t x_1^t + w_1 \sum_t (x_1^t)^2 + w_2 \sum_t x_1^t x_2^t + \cdots + w_d \sum_t x_1^t x_d^t \\
\sum_t x_2^t r^t &= w_0 \sum_t x_2^t + w_1 \sum_t x_1^t x_2^t + w_2 \sum_t (x_2^t)^2 + \cdots + w_d \sum_t x_2^t x_d^t \\
&\vdots \\
\sum_t x_d^t r^t &= w_0 \sum_t x_d^t + w_1 \sum_t x_d^t x_1^t + w_2 \sum_t x_d^t x_2^t + \cdots + w_d \sum_t (x_d^t)^2
\end{aligned}$$

Figure 4: Derivation of normal equations

$$\mathbf{X}^T\mathbf{X}\boldsymbol{w} = \mathbf{X}^T\boldsymbol{r}$$

Figure 5: Normal equations re-written with matrices X, r, and w simplified

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{r}$$

Figure 6: Equation solving for parameters

Thus, the weight matrix can now be solved for with Figure 6. The complete multivariate linear model is shown in Figure 7 below (ε represents error):

$$r^t = g(\boldsymbol{x}^t | w_0, w_1, \ldots, w_d) + \epsilon = w_0 + w_1 x_1^t + w_2 x_2^t + \cdots + w_d x_d^t + \epsilon$$

Figure 7: Note the error included in the equation. This project calculated weights both with and without standardization.

Testing was performed with and without feature scaling and standardization. Standard deviation was found with the formula in Figure 8:

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Figure 8: Standard Deviation formula implemented

In the formula above, x represents each individual attribute value and $\hat{x}$ represents the mean for the attribute population. The product of the subtraction is summed and divided by the size of the attribute set, denoted n. This is used for standardization. Additionally, the z-score was calculated for each individual attribute after finding standard deviation to help interpret the data around the mean. The formula used is displayed below in Figure 9:

$$Z = \frac{x - \mu}{\sigma}$$

Figure 9: Z-Score implementation

Here, z-score Z is solved for by taking the individual attribute value x and subtracting the mean of attribute population μ, and then dividing by the standard deviation for the attribute population σ.

Analysis and Discussion:

For the purpose of this report the weight matrix attribute values will be displayed in the following format:

[[Intercept]

[Cylinders]

[Displacement]

[Horsepower]

[Weight]

[Acceleration]

[Year]

[Origin]]

After running the program without standardization, the resulting weight matrix *w* was produced:

[[-1.80586717e+01]

[-4.18254893e-01]

[1.88870416e-02]

[-1.13851962e-02]

[-6.71865820e-03]

[1.02620868e-01]

[7.56755050e-01]

[1.41751561e+00]]

Figure 9: Weight matrix without standardization

The values produced without standardization are not easy to distinguish, with most weights being smaller decimals and not having a huge effect. Here, Origin had the largest positive effect on the MPG. These values are difficult to make conclusions on since the range of all features are not normalized and therefore are not proportional. The standard deviation for the Origin population was only 0.7168, while the standard deviation for weight was ~717, which is expected due to the large differences between different types of automobiles.

Due to these discrepancies, standardization was also tested and produced the following weight matrix *w*:

[[23.51457286]

[-0.64725951]

[1.72538104]

[-0.340442]

[-4.82346954]

[0.21987236]

[2.40401249]

[1.0160753]]

Figure 10: Weight matrix with standardization

With standardization the results become clearer (and did not change sign). Weight, Cylinders, and Horsepower end up producing negative effects while Displacement, Acceleration, Year, and Origin have positive effects. Weight ended up having the largest effect on MPG, which is not surprising due to needing more energy and decreasing fuel efficiency. Year and Displacement have the largest positive effects, which is also unsurprising due automobile manufacturers improving their designs year-to-year and more displacement having a positive effect on fuel efficiency. Of the other parameters, Cylinders and Horsepower only slightly negatively effected the MPG. These were two parameters of interest due to being swayed by a number of factors out of their control (such as weight). Lastly, Acceleration and Origin had slight-medium positive effects. Acceleration was another value that has the potential to be sidelined by other values, such as weight and horsepower and is not very representative of MPG. Origin of parts or car can play a larger role, where parts made in a country where better materials are used will generally have better performance, break down less, and be more efficient.

For better interpretation of individual data points, a z-score calculation was implemented. Also known as a standard score, it seemed important to include in this project to show how individual data points stray from the mean, and how this affects the overall weight and significance the independent attribute has on the dependent attribute. See Figure 11 on the page below.

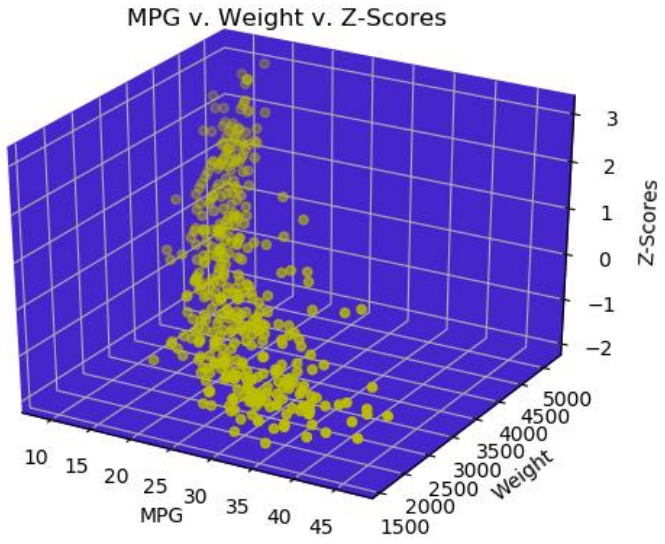

Figure 11: MPG vs. Weight set against the z-score for Weight

Here, it can be seen that the z-scores for the Weight attribute that many of the lighter vehicles have good MPG values and have better z-scores than the heavier, less fuel-efficient vehicles that have high z-scores. This is a good representation of how individual variables can have an effect on the weight of the entire variable population. For a sharper view of standardization effects, see the z-scores for Year below:

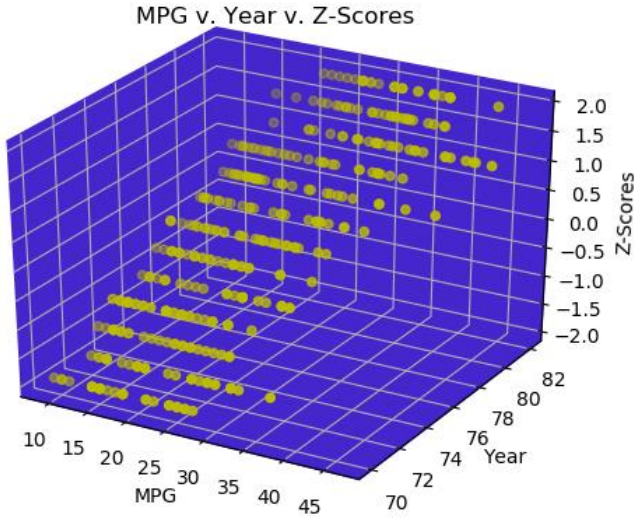

Figure 12: MPG vs. Year against the z-score for Year

As seen in Figure 12, as the variable for Year increases and (assumingly) car manufacturers make more efficient vehicles, the positive effect on MPG will grow. This can not only be seen in the relationship along the XY-axis, but also against the z-axis with higher z-scores meaning a higher standard score above the mean. This relationship signals a positive relationship on MPG.

Overall, a standardized weight matrix led to more predictable and less ambiguous coefficients. Though the non-standardized matrix had the correct signs (signaling the correct relationship), the values were not sufficiently distinguishable. In other words, the multivariate linear regression model is concluded to be much stronger when standardized. This project found that the Weight parameter ultimately had the greatest negative and overall effect on MPG, while Year and Displacement had the greatest positive effects. The remaining parameters: Cylinders, Horsepower, Acceleration, and Origin had varying effects on the model, but not as significant as those previously stated (though Origin was still significant). This makes sense from both a mathematical and practical viewpoint, with some parameters being dominant to others (such as weight vs. horsepower).