Logistic and Linear Regression Report

Paula Garcia

I. Data Sample

The sampled dataset includes metrics from electrical transformers, including the date of testing (year) and measures of water content, total acid (*mgKOH/g*), dissipation factor, color, interfacial tension (*mN/m*), and class. The year of testing was not sampled in the regression models because it would require a time series and was beyond the scope of the assignment.

Five features were considered for the linear regression model, when compared to the response variable (y) of interfacial tension. The five features (water, acid, voltage, dissipation, and color) contained continuous data, allowing proper correlation comparisons with interfacial tension. The top selected features (acid, tension, color) are included in Table 1. The correlation coefficient was used to select the top features because the units of measure differed among the features, making covariance values insufficient for comparisons of relationships.

Feature	Correlation Coefficient (*r*)	Relationship Description
Acid	-0.72	Strong, negative
Dissipation	-0.71	Strong, negative
Color	-0.84	Strong, negative

Table 1: Feature Selection for Linear Regression

Six features were considered for the logistic regression model, when compared to the response variable (y) of class. The six features (water, acid, voltage, dissipation, color, tension) do not include the year data. The top selected features (acid, color, tension) are included in Table 2.

Feature	Correlation Coefficient (*r*)	Relationship Description
Acid	-0.18	Weak, negative
Color	-0.15	Weak, negative
Tension	+0.15	Weak, positive

Table 2: Feature Selection for Logistic Regression

II. Linear Regression

Linear regression is a supervised learning method that models a linear relationship between continuous or ordered independent variable(s) (x) and a dependent variable (y) [3]. The model includes the intercept (b_0) and slope coefficient (b_1) to approximate the estimated changes in the averaged y-values by changes in x.

$$\hat{y}_i = b_0 + b_1 X$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where the coefficient values are found by minimizing the sum of squared residual values [1]. The data can be portioned prior to fitting a linear regression model to facilitate a learning model from the data. Learning models avoid a highly customized model that cannot be used on future datasets. *Simple Linear Regression* models can be seen in Figure 1 for all sampled features. The figures are associated with the correlation values in Table 1.

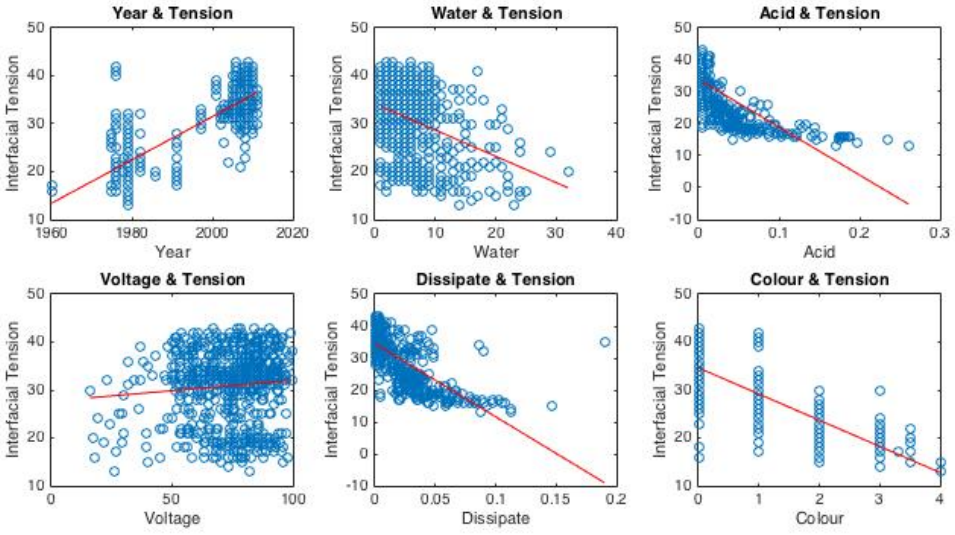

Figure 1: Simple Linear Regression Models for Each Sampled Feature

i. Learning Models

Multivariate linear models were created for all the data sampled and the top three correlated values (acid, dissipation, color), whereby 65% ($x(i)_n$=475) was used for training and 35% ($x(i)_n$=255) was used for testing, where i=variable sampled. An accuracy summary of the multivariate linear model is presented in Figure 2, delineating a comparison between the predicted y-values and actual y-values for the two samples. If the multivariate model had achieved 100% accuracy, all points would lie on the linear model.

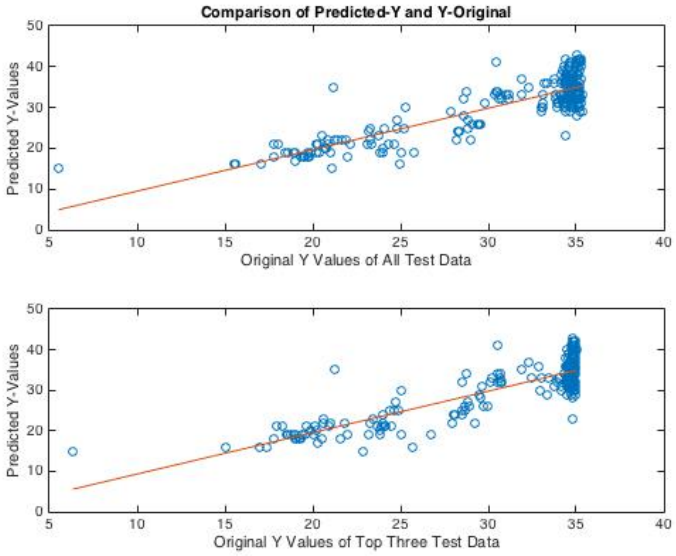

Figure 2: Accuracy of Combined Multivariate Linear Regression Predicted Y vs. Actual Y

The results of the multivariate linear regression on all the data echoed the results of correlation in Table 1, the top three predictors for tension were acid (β =-4.02, p=0.65), dissipation (β =-72.96, p=<.001), and color (β =-4.12, p=<.001). As shown in Figure 2, the three predictors provide a succinct summary of the model (R^2=0.73), without the need for the extra attributes (R^2=0.74). More information is provided in the results section of this report.

K-fold cross validation was used to predict the fit of the linear model with smaller sample sizes [2]. The three sampled features were partitioned into 10 chunks, where each chunk maintained 73 randomly selected samples (n=730). Each chunk was used for both training and testing to generate an average of estimated errors ($RMSE$=3.67, k=10, k_n=73). K-fold cross validation helps to estimate a models performance by minimizing over and under fitting errors. K-fold results provided a 95% confidence interval of the mean-squared errors between [11.81,15.07].

ii. Results

A multivariate linear regression was calculated for ***X = [water, acid, dissipation, color]*** and ***Y=[interfacial_tension].*** The estimated coefficients and statistical metrics are outlined in Table 3. The model yielded a high fit with the observed data (R^2=0.74, F(475,469)=266, p<.001) [2]. The RMSE of this model was 3.7.

Feature	Coefficient Est.	Standard Err	t-Statistic	p-Value
Tension	35.74	0.96	37.2	<.001
Water	-4.02	8.98	-0.45	0.65
Acid	-0.002	0.01	-0.18	0.86
Dissipation	-72.96	15.69	-4.65	<.001
Color	-4.12	0.27	-15.47	<.001

Table 3: Estimated Coefficient Summary Linear Regression

A multivariate linear regression was calculated for ***X = [acid, dissipation, color]*** and ***Y=[interfacial_tension].*** The estimated coefficients and statistical metrics are outlined in Table 4. The model yielded a high fit with the observed data (R^2=0.73, F(475,471)=430, p<.001) [2].

Feature	Coefficient Est.	Standard Err	t-Statistic	p-Value
Tension	35.01	0.21	166.83	0
Acid	-22.85	7.66	-2.98	.003
Dissipation	-57.64	11.13	-5.18	<.001
Color	-4.031	0.23	-17.51	<.001

Table 4: Estimated Coefficient Summary Linear Regression

The k-fold results yielded an MSE of 13.42, with 95% confidence intervals of [*11.60,15.25*]. The original multivariate model's MSE was similar at 13.84. The RMSE of the k-fold was 3.65, providing the average distance from each data point to the fitted line.

Overall, the model with only three predictors is a promising alternative to sampling all the data points, justified by its coefficient of determination and RMSE results. This is especially helpful for the type of data, as sampling all the attributes of an electrical transformer is expensive.

iii. References

[1] J.G, Kerns. "Simple Linear Regression" in *Introduction to Probability and Statistics Using R*, 1st ed. 2011. Ch. 11, pp. 249-275.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, vol. 9, no. 2. 1995.

[3] A. Mooman, "Introduction to Linear Regression and Correlation Analysis" Dept. Computing and Information Sciences., New York, March. 2016.

III. Logistic Regression

Logistic regression is commonly used when the y-variable of the data only takes on two values (dichotomous) or results in a 0,1 (binary). Similarly to linear regression, logistic regression attempts to model the relationship between the x and y variable(s) to predict future outcomes. Although the purpose is similar to linear regression, logistic regression maintains limits in its fitted model whereas the linear model does not. Applying a linear model to a logistic model would result in a model that exceeds the bounds of the possible outcomes for the y-variable.

The multivariate equation for more than one x-value is:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

where the model is created through the log odds of a linear function of the x-variables. The odds are calculated by the categories assigned to the y-values.
When the logistic regression has more than two classes, the likelihood function that is fit to the data includes the possibilities for each x variable [1]:

$$\Pr\left(Y = c|\vec{X} = x\right) = \frac{e^{\beta_0^{(c)} + x\cdot\beta^{(c)}}}{\sum_c e^{\beta_0^{(c)} + x\cdot\beta^{(c)}}}$$

The first coefficient calculated is the intercept; the remaining signify the slope values or relative odds of being in one category over the other.

i. Learning Models

The multivariate and multinomial models for logistic regression used a 65% training set ($x(i)_n$=475) and 35% training set ($x(i)_n$=255). As was done in the linear regression model, the numbers sampled during each training session was randomly chosen.

Two learning models were created based on the partitioning of the training and testing data sets. The multinomial logistic model was created for all the data sets X= ***[water, acid, dissipation, color, tension]*** and *y=[class]*. The classes in *y* included three options (N,B,M,G) which were converted to numbers (1,2,3,4) for the purposes of fitting the model.

ii. Results

The coefficient estimations by the multinomial model is summarized in Table 5. The main predictors for each associated class were water (p=0.03, p=0.01, p=0.009), dissipation (p=0.01, p=<.001, p=0.006) and tension (p=0.003, p=<.001, p=0.24). The model showed a strong relationship between the observed and expected points, with an estimated dispersion parameter of (2169, n=730,6)=0.46, with Φ=1.

Feature	Coefficient (N)	Coefficient (B)	Coefficient (M)
Class	26.88	-16.37	8.80
Water	-0.23	-0.23	-0.20
Acid	-18.90	-26.51	-36.05
Dissipation	-0.01	-0.04	-0.03
Color	-0.01	-0.04	-0.03
Tension	-46.93	-102.64	-53.86

Table 5: Estimated Coefficient Summary by Class, where Class (G)=0, as Reference Class

Although water showed a high *p-value*, acid was chosen over water because of its higher correlation when compared singularly between the class. The resulting three samples were ***X=[acid, color, tension]*** and ***y=[class]***. A multinomial logistic regression was created for this sampled data, yielding an estimated dispersion parameter of (2178, n=730,3)=0.95, with Φ=1.

Feature	Coefficient (N)	Coefficient (B)	Coefficient (M)
Class	11.77	-34.42	-7.01
Acid	-26.34	-43.49	-44.61
Color	-0.45	-0.91	-0.66
Tension	-0.30	1.73	0.66

Table 6: Estimated Coefficient Summary by Class where Class G=0, as Reference Class

When comparing the residual sum of squares (RSS) per class, the model with the most features (variables shown in Table 5), had lower values than the model with the top three features (variables shown in Table 6). Lower residual sum of square values show that the model has a lower difference between the data and the generated model. The results, as shown in Table 7, did not vary widely.

Model	RSS (N)	RSS (B)	RSS (M)	RSS (G)
All features	14.31	5.78	17.57	4.43
Top 3 features	15.89	6.42	18.34	5.64

Table 7: Estimated Coefficient Summary by Class where Class G=0, as Reference Class

Since the difference in RSS between the two models is low, it would be suggested to use the model trained on the top 3 features for classification. The model with the top 3 features is descriptive enough to yield each class with a minimal loss in fit to the data when compared to the model with all the features.

An improvement to this model would be a four-part binomial logistic regression model, where each class would be tested separately with the outcome of 0 or 1 to indicate fit. A binomial logistic regression model would provide additional means of analysis for more accurately predicting the probability of fit.

iii. References

[1] "Logistic Regression". *Chapter 12*. 2012. [online] Available: http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf. [Accessed: 15-March-2016].