Predict 413 Sect. 55:

Predicting the Probability of Blood Donations using Multivariate Linear, Logistic and Poisson Regression

Daren Purnell

Northwestern University

Executive Summary and Introduction

The purpose of this analysis is to predict the probability of an individual donating blood. The data comes from a mobile blood donation vehicle in Taiwan that drives to different college campuses and collects blood from donors during a blood drive. Based off obtained data regarding the frequency and amount of the person's donation, we want to predict the probability of that person donating the next time the vehicle visits their campus.

Donating blood is an accessible way that a person can help save lives. While the specifics about Taiwanese blood donations are difficult to interpret, the American Red Cross estimates that "every two seconds someone in the United States needs blood" (American Red Cross, 2017). Despite the current state of technology, humanity is not able to produce blood on an industrial scale to fulfill the demand. Therefore, understanding the supply chain and logistics of obtaining donations is very important. "Although an estimated 38 percent of the U.S. population is eligible to donate blood at any given time, less than 10 percent of that eligible population actually do each year" (American Red Cross, 2017). If we can better understand the mechanisms that contribute to the decision to donate, policies and programs can be developed to increase our much-needed blood supply.

Exploratory Data Analysis

The mobile blood donation center data consist of six variables; an index "X", the response variable "Made donation in March 2007" and four regressor variables, "Months since last donation", "Number of donations", "Total volume donated CC (cubic centimeters)", and "Months since first donation". The index "X" is a unique identifier for the records and will not be used for the analysis. There are 576 total records and it appears that there are no missing values for any of our variables.

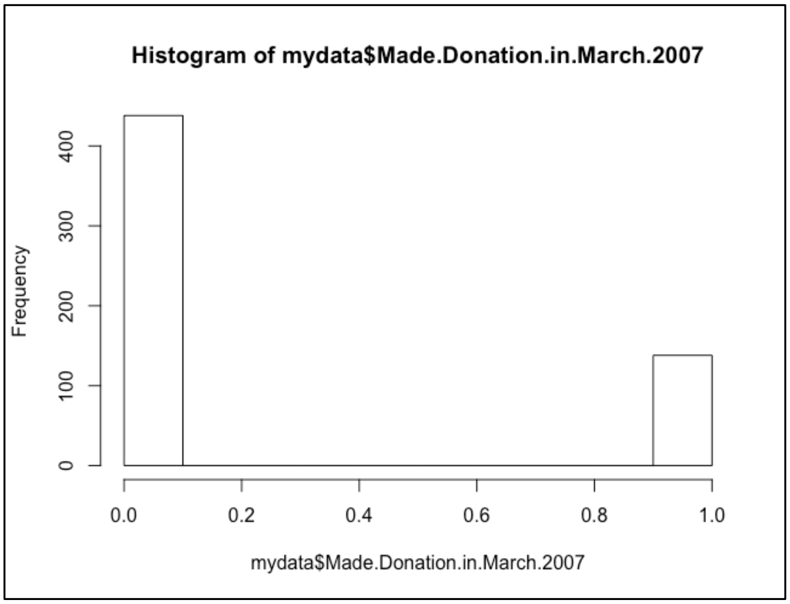
Computing the Pearson Correlation Coefficients of our data set allow us to better understand the relationships between the response and regressor variables. The regressor variable "Months since last donation" has an R value of -0.261 which implies that there is an inverse relationship between the decision to donate and the time since the person's last donation. This makes sense because we can infer that a person that donates less frequently is less likely to donate overall. The regressor variables "Number of donations" and "Total volume donated CC" have the same R value of 0.221. This implies a positive relationship between the two variables and "Made donation in March 2007" where the number and volume of donations increase the probability that an individual will donated again. Further analysis also reveals a 0.622 Pearson Correlation Coefficient value between the two regressor variables which means that they are strongly correlated. This is likely because the volume of each separate blood donation is limited and the only way to significantly increase volume is to increase the frequency of donations. i.e. more donations = more total volume. The regressor variable "Months since first donation" R value is -0.02 which implies that the larger the amount of time since the first donation, the least likely the person is to donate again. In summary, it appears that the amount of time that has passed since the last donation/first donation decreases the probability the person will donate and the number and total volume of donations increases the probability that the person will donate.

Response Variable Pearson Correlation Coefficients

Response Variable	R Value
Months since last donation	-0.261
Number of donations	0.221
Total volume donated CC	0.221
Months since first donation	-0.02

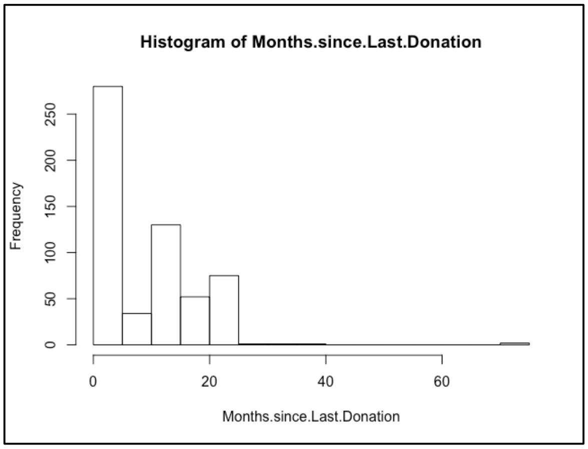
Made donation in March 2007

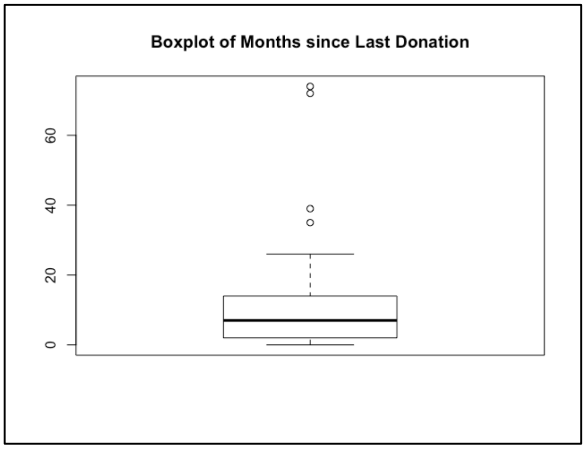
The response variable "Made donation in March 2007" represents whether an individual donated blood to the Taiwanese mobile blood donation center during the March 2007 blood drive. A value of "0" means that the person did not donate blood and a "1" means that a person did donate blood. The categorical variable has 438 null "0" values and 138 affirmative "1" values. There are no missing values for the variable "Made donation in March 2007".



Months since last donation

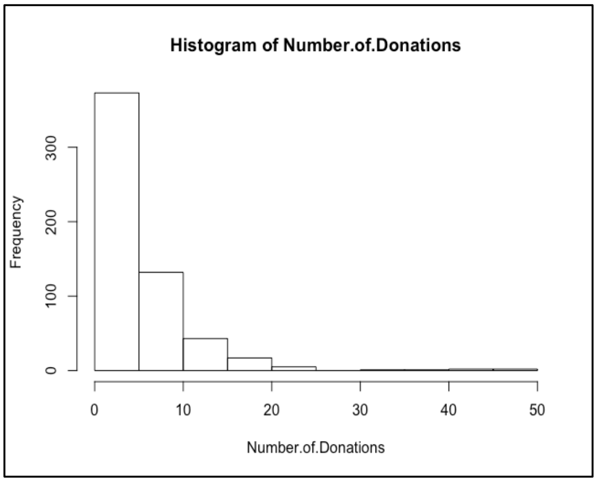
The regressor variable "Months since last donation" represents the number of months that have passed since the person's last blood donation. This continuous variable has a mean of 9.439 and median of 7 months. The histogram plot shows a heavily right-skewed distribution that indicates that most of the individuals have either not donated before or that March 2007 was their first donation. A box plot of the regressor shows four outliers above the upper (Q3 + 3*(Q3-Q1)) bounds that may influence our regression attempts.

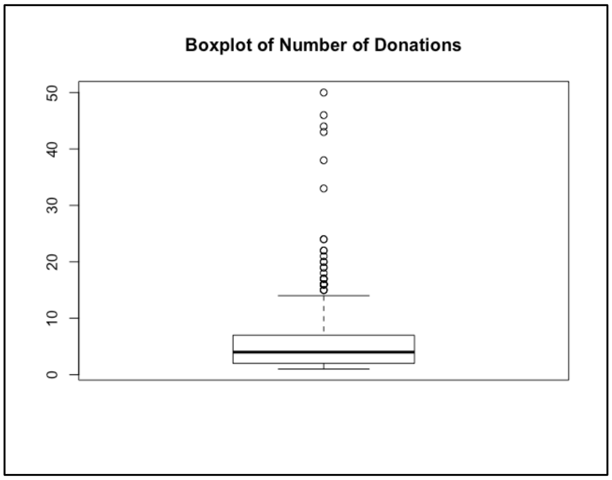




Number of donations

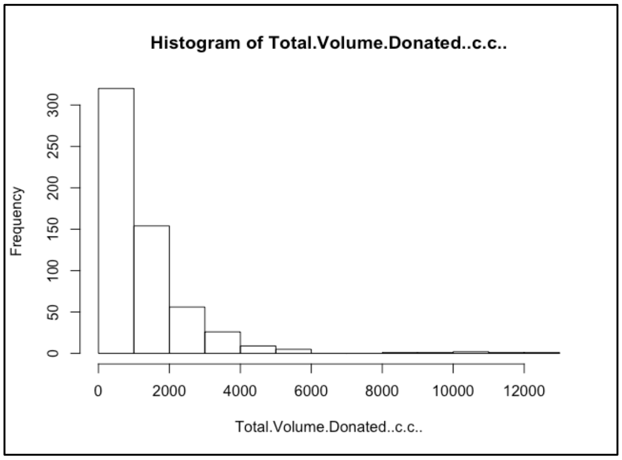
The regressor variable “Number of donations” represents the number of total blood donations that a donor has made. This continuous variable has a mean of 5.472 and median of 4 donations. The histogram plot shows a heavily right-skewed distribution that indicates that most of the individuals probably have not donated before. A box plot of the regressor shows several outliers above the upper (Q3 + 3*(Q3-Q1)) bounds that may influence our regression attempts.

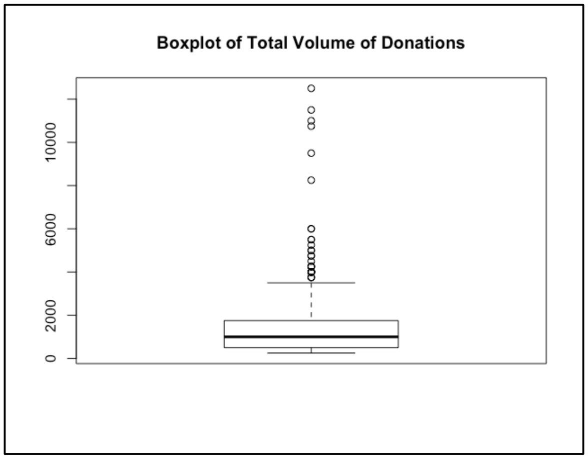




Total volume donated CC (cubic centimeters)

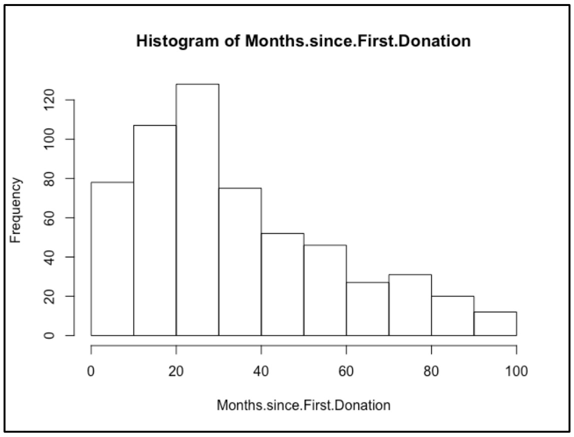
The regressor variable “Total volume donated” represents the total volume of blood donations that a donor has made in cubic centimeters. This continuous variable has a mean of 1357 and median of 1000 cubic centimeters. The histogram plot shows a heavily right-skewed distribution that indicates that most of the individuals have donated between 0 to 1000 cubic centimeters of blood. A box plot of the regressor shows several outliers above the upper (Q3 + 3*(Q3-Q1)) bounds that may influence our regression attempts.

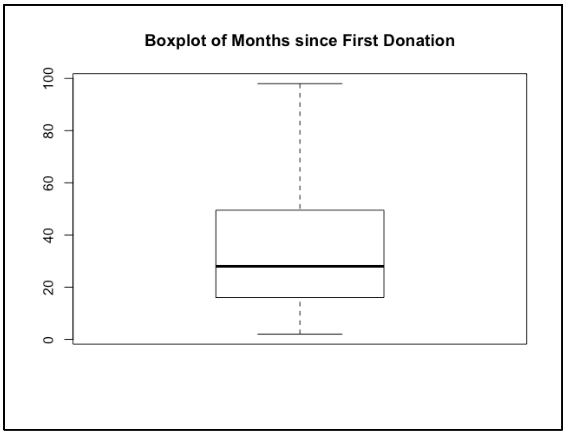




Months since first donation

The regressor variable “Months since first donation” represents the number of months that have passed prior to the individual’s March 2007 blood donation. This continuous variable has a mean of 34.05 and median of 28 months. The histogram plot shows a heavily right-skewed distribution that indicates that most of the individuals have waited around 34 months to donate again. A box plot of the regressor shows no outliers above the upper or lower bounds that may influence our regression attempts.

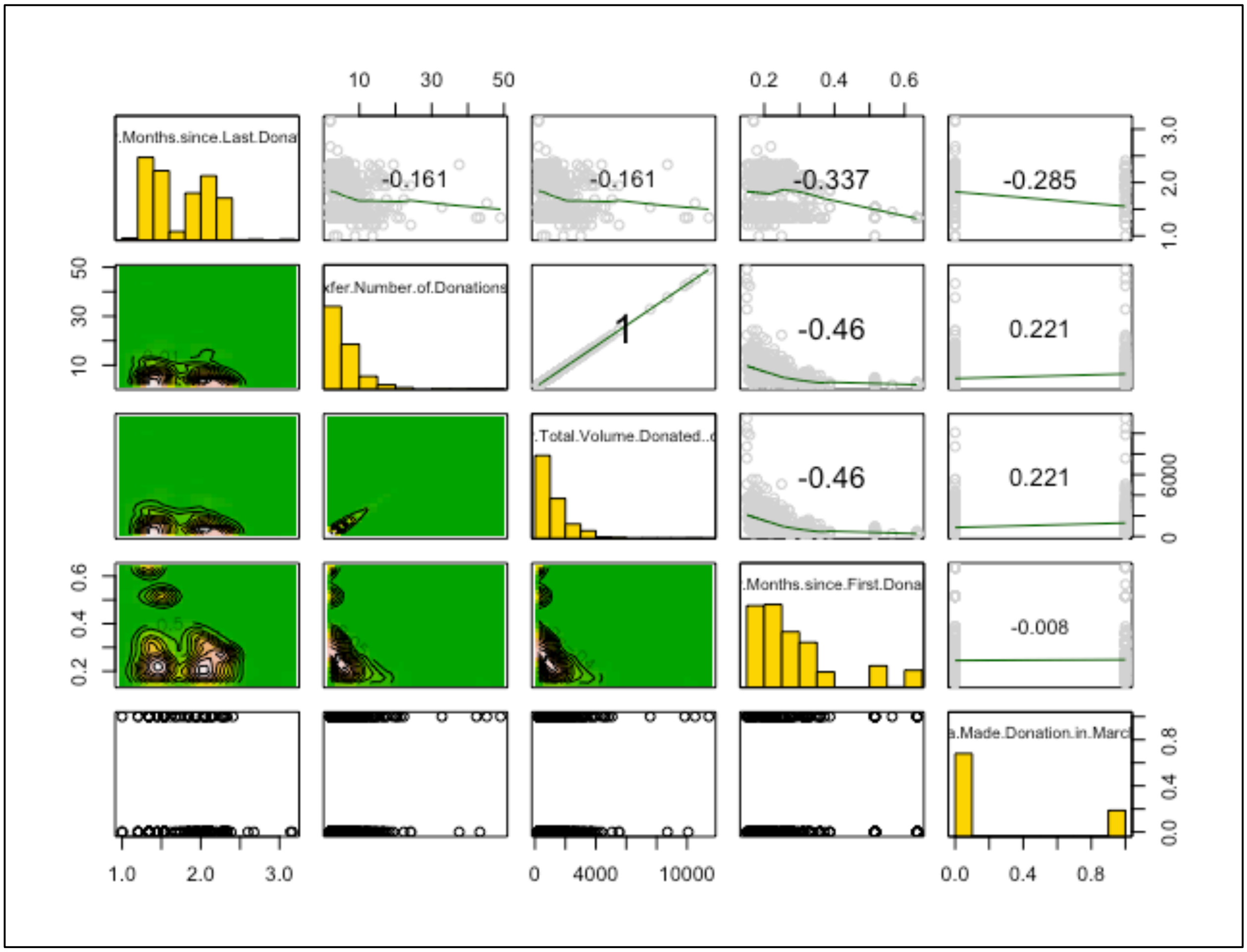




Data Preparation

Data preparation efforts consisted of variable transformation for ordinary least squares regression. While not necessary for generalized linear models, to include logistic regression, Box Cox variable transformation produced an improved predictive model over the baseline Linear, Logistic, and Poisson regression models. Logistic, additive, and multiplicative transformation were attempted, however the Box Cox transformation proved to be superior. Adhering to the assumptions of linear regression (i.e. normal and independent distribution, linear relationship between variables, no or limited autocorrelation/multicollinearity, homoscedasticity) produced a more accurate model. No outliers were removed from the data to their effects upon the selected Linear, Logistic, and Poisson Regression models.

The matrix below provides a summary of the Box Cox transformation upon the regressor variables. In the bottom, right diagonal of the matrix we have 2D kernel density estimates and in the upper, left diagonal their bivariate scatterplots and Pearson Correlation Coefficient values.



Model Development & Selection

In the process of our analysis we developed three classes of models using the R program: a Multivariate Linear Regression model, Logistic Regression model, and Poisson Regression model. Typically, Logistic and Poisson Regression are special case distributions used for specific scenarios. Logistic Regression is used when modeling dichotomous outcomes, such as in this situation where a "0" value indicates a null donation and a "1" confirms a blood donation. Poisson Regression is used when we are attempting to predict the count of data such as the number of donations someone has made to the mobile blood donation center.

Some examples, from peer-reviewed journals of how the types of models we developed were used in similar situations are:

- In the peer-reviewed journal Diabetes Care article *Insulin Sensitivity, Vascular Function, and Iron Stores in Voluntary Blood Donors* (Zheng, et al. 2007) Linear Regression is used to determine whether iron-dependent changes in glucose metabolism may contribute to improved vascular function in blood donors.

- The American Journal of Epidemiology's *Logistic Regression in Survival Analysis* (Abbott 1985) article discusses the use of Logistic Regression to examine the relationship between risk factors and various disease events.
- The Journal of Marriage and Family's *Logistic Regression: Description, Examples, and Comparisons* (Morgan and Teachman 1988) article highlights the use of Logistic Models to study topics such as marriage formation & dissolution, contraceptive use, poverty, premarital pregnancy, and spouse abuse.
- The Journal of Biopharmaceutical Statistics article *Poisson Regression Analysis in Clinical Research* (Kianifard and Gallo 2007) discusses the use of Poisson Regression in analyzing data from clinical trials and epidemiological studies. In the article, Poisson Regression is illustrated on a data set for the clinical trial for the treatment of bladder cancer.
- The Corvinus Journal of Sociology and Social Policy's *The Use of Poisson Regression in Sociological Study of Suicide* (Hegedus 2014) discusses the use of Poisson Regression in studies where the dependent variable describes the number of occurrences of some rare event such as suicide.

Following variable transformation, we progressed to splitting the designated training data into thirds with 2/3rd composing the revise training set and 1/3rd creating a new validation data set. This validated data was later used to develop an accuracy metric that compared each model's predictions against the validated data's true results. Counter to our initial assumption regarding the use of Logistic Regression for this situation, our Multivariate Linear Regression model with Box Cox transformation outperformed our Logistic and Poisson Regression models on both the validation and test data sets. Despite our understanding on the use of Logistic Regression models to predict the outcomes of dichotomous variables, we selected the Linear Regression model due to its overall performance on both sets of data.

A summary of the accuracy outcomes is below:

Model Type	**Validated Data Accuracy Results**	**Test Data (Log Loss) Results from DRIVENDATA**
Linear Regression w/ Box Cox Transformation	79%	0.4681
Logistic Regression w/ Box Cox Transformation	77%	0.4714
Poisson Regression w/ Box Cox Transformation	78%	0.4692

The selected Linear Regression model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Where:

In Model	In Data	Beta	Value
Y	Made donation in March 2007	β_0	25.50547
X1	Months since last donation	β_1	-0.21913
X2	Number of donations	β_2	-25.54871
X3	*Total volume donated CC*	β_3	0.10748
X4	Months since first donation	β_4	0.34749
$\boldsymbol{\varepsilon}$	Error Term		

With Fit Statistics:

AIC	AICc	BIC	AdjR2
-678.0440191	-677.8212075	-654.3401638	0.1001339

Model Notes:

- Number of donations has a negative coefficient which is counter to intuition. I would hypothesize that a person that made frequent donations would be more likely to donate blood during the next blood drive.

- Months since first donation has positive coefficient which is counter to intuition. I would hypothesize that the longer a person has waited since their first donation, the least likely they are to donate during the next blood drive.

- I worry that my model may be over trained due to the negative coefficients on the two regressor variables that run counter to my intuition. While the Multivariate Linear Regression model performs well on the validation and competition data set, I think it may be limited when applied to larger set of data. Additionally, the two negative coefficients make the model difficult to explain to others because it runs counter to assumptions about the impact of the regressor variables.

Summary

During this analysis, we preformed Multivariate Linear Regression, Logistic Regression, and Poisson Regression to predict the probability of a person donating during a blood drive. Exploratory Data Analysis was conducted to understand the distributions and relationships between our regressor and response variables. We then implemented regressor variable Box Cox transformations to adhere to the assumptions of Linear Regression and improve the performance of our models. Finally, an iterative trail-and-error process was used to develop our three classes of models based of an accuracy metric that compared the results of each model to validation data. Despite our understanding on the use of Logistic Regression models to predict the outcomes of dichotomous variables, we selected the Linear Regression model due to its performance on the validation and test data. If we had more time, we would like to further explore the performance of the Linear Regression model over the other two models, attempt additional variable transformations to further optimize our model, and implement a binning process to remove and replace outliers with a measure of central tendency.

Works Cited

Abbott, R. D. (1985). Logistic Regression in Survival Analysis. *American Journal of Epidemiology*, 465-471.

American Red Cross. (2017, 07 29). *Blood Facts and Statistics* . Retrieved 07 29, 2017, from Red Cross Blood: http://www.redcrossblood.org/learn-about-blood/blood-facts-and-statistics

Hegedus, F. M.-l. (2014). The Use of Poisson Regression in the Sociological Study of Suicide. *Corvinus Journal of Sociology and Social Policy*, 97-114.

Kianifard, F., & Gallo, P. P. (2007). Poisson Regression Analysis in Clinical Research. *Journal of Biopharmaceutical Statistics*, 115-129.

Morgan, S. P., & Teachman, J. D. (1988). Logistic Regression: Description, Examples, and Comparisons. *Journal of Marriage and the Family*, 929-936.

Zheng, H., Patel, M., Cabel, R., Young, L., & Katz, S. (2007). Insulin Sensitivity, Vascular Function, and Iron Stores in Voluntary Blood Donors. *Diabetes Care*, 2685-2699.