The Classification of a Binary Target Variable Using a Binomial Generalized Linear Model

TJ Smith

Section 1 - Introduction

College is a staple in the world where young people pursue their passions and become well rounded adults. Older generations often donate to colleges so that the younger generations have sufficient opportunities to pursue their dreams. Like how students pick their colleges, there are a lot of factors that donors consider when choosing a college to donate to. With so many different things to consider, it can be difficult for the average donor to weigh their options.

The data set at hand provides numerous variables that relate to colleges. An analysis of these variables can indicate which variables are important in predicting the target variable. As prescribed in the problem, a binomial model will be used to analyze the data.

The plan for the remainder of the report is as follows. Section 2 will discuss the nature of the data and provide some summary findings about the importance of it. The details of how we selected our model, and the interpretation of that model will be found in section 3. The selected model will be applied in section 4 to exhibit how well it compares to the data itself. The report ends with section 5 which contains concluding remarks.

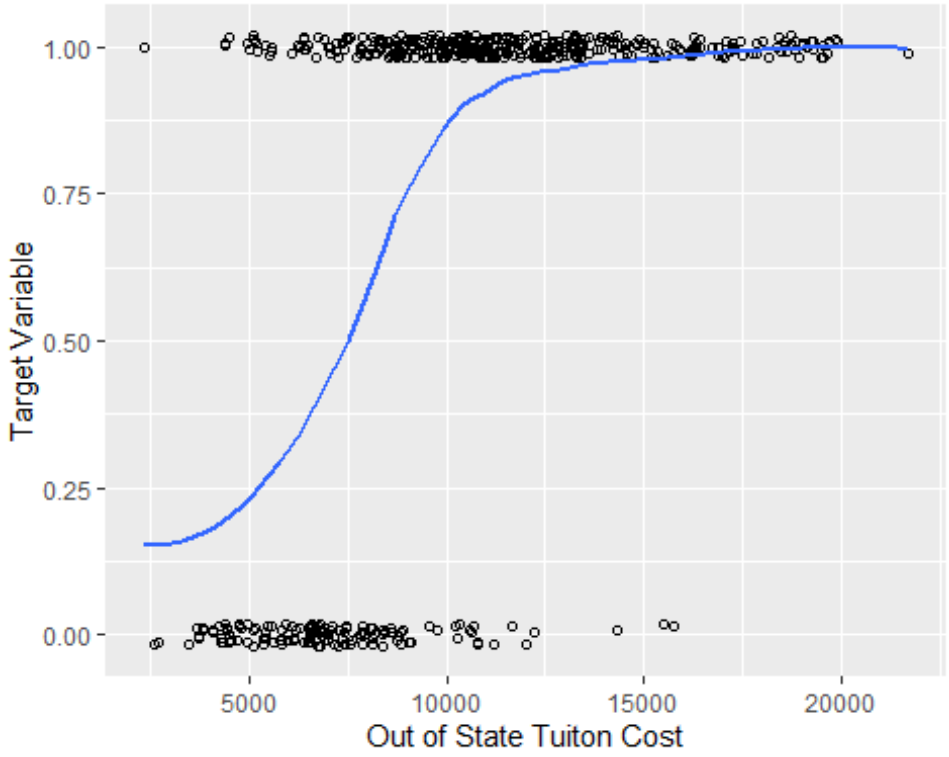
Section 2 - Data Characteristics

The data at hand is cross-sectional and observational, as the data collection time is undisclosed. It contains information from 500 different unnamed universities across the United States. We are given 17 factors (independent variables) that may sway ones decision when choosing a college to donate to. These different variables span a wide variety of topics such as financial (tuition,personal cost), academic (applicants accepted, percent of faculty with phd's), and other miscellaneous things such as graduation rate, student-faculty ratio, and percentage of alumni who donate. From these factors, the colleges are given a binary value of 0 or 1. It is our job to discover which variables play a significant role in determining whether the college is valued as a 0 or 1.

The graph below provides an example between the variable Outstate and the target outcome. It is quite clear that if there is a higher out-of-state tuition cost, the target variable is highly likely to be 1. From this, one could conclude that out-of-state tuition cost is a significant variable in predicting the target outcome.

```
pdc <- ggplot(data = co,
              mapping = aes(x = Outstate,
                            y = Target))
pdc <- pdc + geom_jitter(height = 0.02, pch = 1)
pdc <- pdc + geom_smooth(se=FALSE)
```

```
pdc <- pdc + labs(x = "Out of State Tuiton Cost",
                  y = "Target Variable")
pdc

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The target variable in the data set is categorized as a 0 or 1. As seen below, the mean of the target variable is 0.728, indicating that the classification of the target variable is more commonly a 1 as opposed to 0.

```
summary(co$Target)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   0.728   1.000   1.000
```

Section 3 - Model Selection/Interpretation

The previous section provided preliminary plots and summary statistics for the data we have at hand. This section will discuss the process of narrowing down the data to fit a binomial generalized linear model. Following the final model and its interpretation, the section will exhibit some graphics that drove the selection of this specific model.

As we were tasked with creating a binomial model, there was no need for analysis to find the correct family of functions to use.

The final model was fitted using a combination of log-likelihood tests and Wald tests, which weeded out variables that proved insignificant in predicting our target variable. It was

determined that the significant coefficients were Outstate, PhD, perc.alumni, Apps, and Expend which are described below.

Outstate -> Out-of-state tuition PhD -> Percent of faculty with PhD degrees perc.alumni -> Percent of alumni who donate Apps -> Number of applications received Expend -> Instructional expenditure per student

A summary of the final model is shown below:

```
adjmodel5 <- glm(Target ~ Outstate+PhD+perc.alumni+Apps+Expend,data=co,family
=binomial)
summary(adjmodel5)

##
## Call:
## glm(formula = Target ~ Outstate + PhD + perc.alumni + Apps +
##     Expend, family = binomial, data = co)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4627  -0.0553   0.0533   0.1794   2.9399
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.298e+00  1.191e+00  -1.090 0.275769
## Outstate     7.242e-04  1.063e-04   6.813 9.55e-12 ***
## PhD         -8.352e-02  1.924e-02  -4.342 1.41e-05 ***
## perc.alumni  8.646e-02  2.348e-02   3.682 0.000231 ***
## Apps        -5.358e-04  7.846e-05  -6.829 8.52e-12 ***
## Expend       2.704e-04  1.151e-04   2.351 0.018743 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 585.24  on 499  degrees of freedom
## Residual deviance: 176.87  on 494  degrees of freedom
## AIC: 188.87
##
## Number of Fisher Scoring iterations: 7
```

The equation for the final model:

$$logit(mu) = -1.297 + 0.00072 * Outstate + 0.0864 * perc.alumni + 0.00027 * Expend - 0.0835 * phd - 0.00053 * Apps$$

Unlike a normal linear regression, the interpretation of the binomial generalized linear model can be quite difficult to understand. A variable can be interpreted as follows:

If out-of-state tuition is increased by $2,000 then there is a 1.448 (2000*0.0007242341) unit change in the log odds. Again, it is difficult to interpret what a 1.448 increase in the log odds is, so we can convert this into an odds ratio, by exponentiating the coefficient. exp(1.448) = 4.25. An odds ratio of 4.25 tells us that we are 4.25 times more likely to predict target = 1 than target = 0.

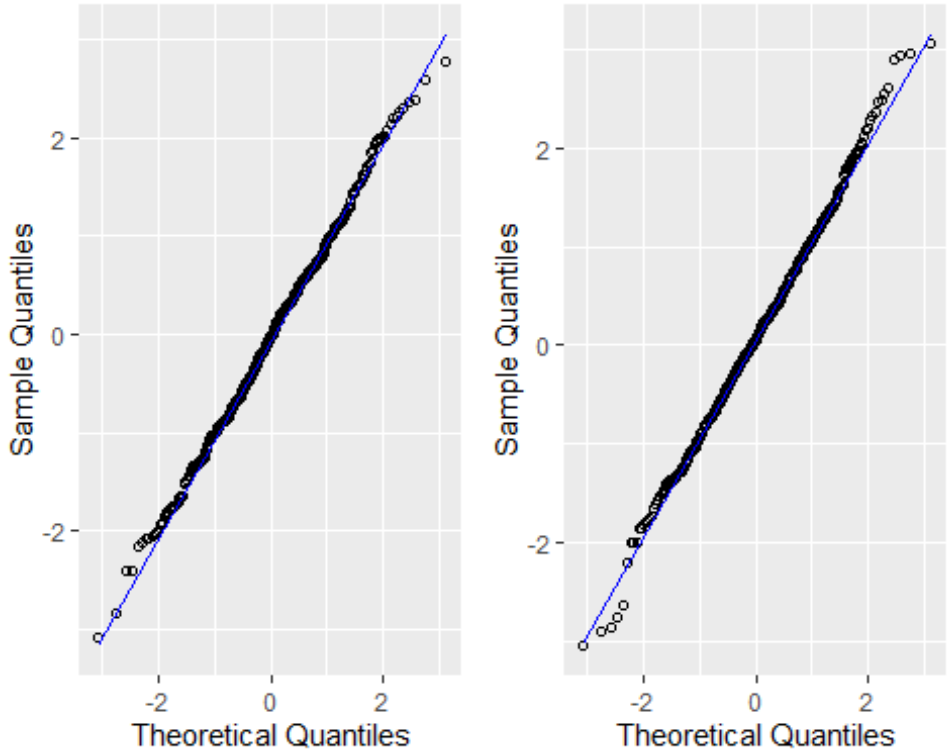
To put everything together, in context, if out-of-state tuition is increased by $2,000, holding all other variables constant, we are 4.25 times more likely to predict the correct target variable.

An out-of-model vs quantile residual plot was used to ensure the model did not exclude any significant variables. If the plot exhibits a pattern that is drastically different from a line at y=0, then the variable may add value to the model. The most extreme case of deviation in the data of the room and board variable is shown below. However, there was not sufficient evidence that we needed it in the model, meaning there are no variables in the data set that would add any additional predicting value.

```
pom <- ggplot(data = co,
              mapping = aes(x = Room.Board,
                            y = adjmodel5.rQ1))
pom <- pom + geom_point() + geom_smooth(se = TRUE)
pom <- pom + labs(x = "Cost of Room and Board",
                  y = "Quantile Residuals",
                  title = "Model")

qom <- ggplot(data = co,
              mapping = aes(x = Room.Board,
                            y =adjmodel5.rQ2))
qom <- qom + geom_point() + geom_smooth(se = TRUE)
qom <- qom + labs(x = "Cost of Room and Board",
                  y = "Quantile Residuals",
                  title = "Model")
grid.arrange(pom, qom, nrow = 1)

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

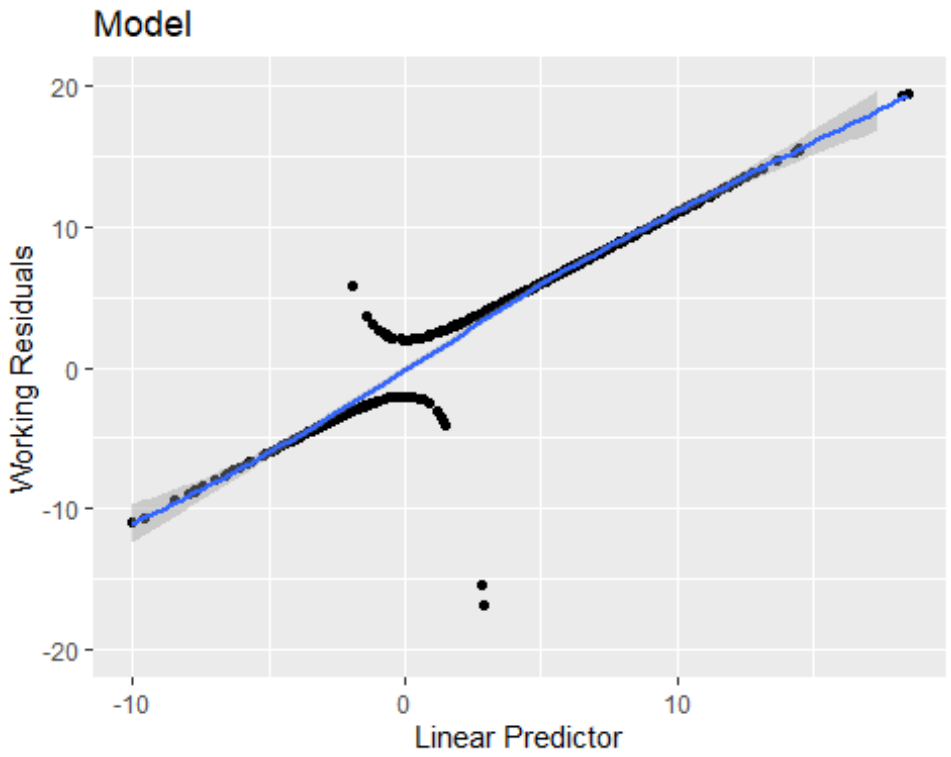
Although it was specified to use a binomial generalized linear model, it is important to ensure that the data at hand fit a binomial model. The sample quantile vs theoretical quantile plot is used to do this. If the sample quantiles fit a linear pattern, then the model chosen is correct. The graph below exhibits that the data fit the linear model, so the binomial family is the correct choice.

```
pt <- ggplot(data = co,
             mapping = aes(sample = adjmodel5.rQ1))
pt <- pt + geom_qq(pch = 1) + geom_qq_line(color = "blue")
pt <- pt + labs(x = "Theoretical Quantiles",
                y = "Sample Quantiles")

qt <- ggplot(data = co,
             mapping = aes(sample = adjmodel5.rQ2))
qt <- qt + geom_qq(pch = 1) + geom_qq_line(color = "blue")
qt <- qt + labs(x = "Theoretical Quantiles",
                y = "Sample Quantiles")
grid.arrange(pt, qt, nrow = 1)
```

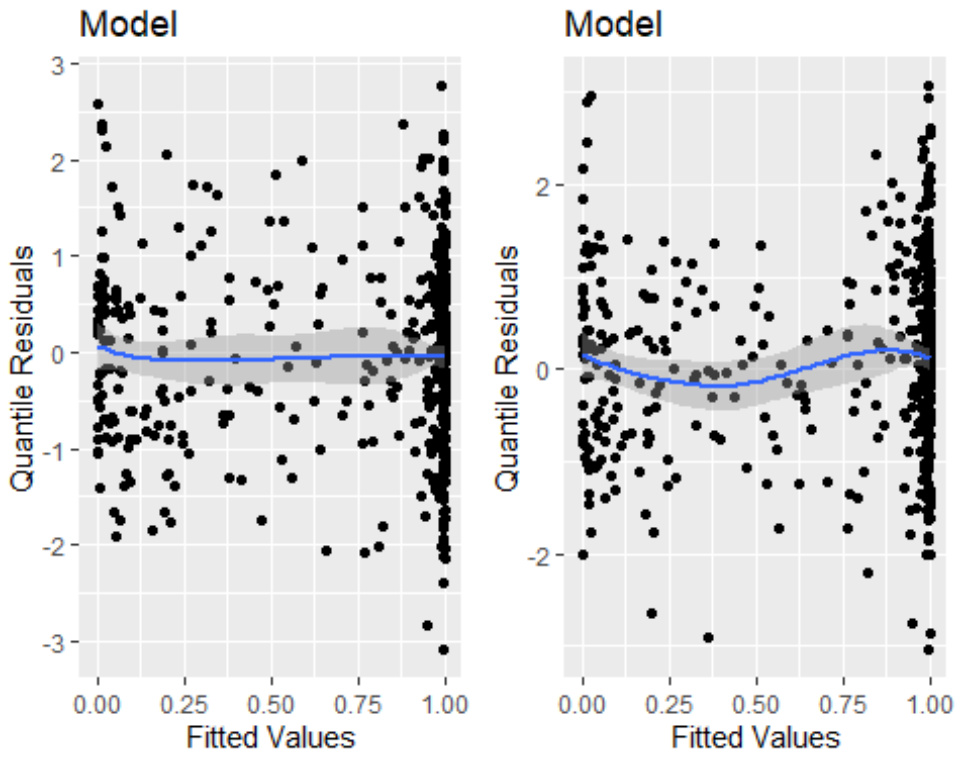


These final two plots ensure that the link function is accurate and that the overall model is a good predictor. The linear predicting graph exhibits a clear straight line indicating that the link function of the model is a great fit. The fitted values vs quantile residuals plot demonstrates a line across y=0, indicating that the overall binomial generalized linear model is a great fit.

```
p <- ggplot(data = co,
            mapping = aes(x = adjmodel5.eta,
                          y = adjmodel5.wr))
p <- p + geom_point() + geom_smooth(se = TRUE)+ylim(-20,20)
p <- p + labs(x = "Linear Predictor",
              y = "Working Residuals",
              title = "Model")
p

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 4 rows containing non-finite values (stat_smooth).

## Warning: Removed 4 rows containing missing values (geom_point).
```




```
p <- ggplot(data = co,
            mapping = aes(x = adjmodel5.mu,
                          y = adjmodel5.rQ1))
p <- p + geom_point() + geom_smooth(se = TRUE)
p <- p + labs(x = "Fitted Values",
              y = "Quantile Residuals",
              title = "Model")

q <- ggplot(data = co,
            mapping = aes(x = adjmodel5.mu,
                          y = adjmodel5.rQ2))
q <- q + geom_point() + geom_smooth(se = TRUE)
q <- q + labs(x = "Fitted Values",
              y = "Quantile Residuals",
              title = "Model")
grid.arrange(p,q, nrow = 1)

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Now that the model has been evaluated and determined that it is a good fit for the data, we need to determine how good the model actually is. One way to do this is by analyzing the model's accuracy, sensitivity, and specificity using a confusion matrix, which is shown below.

```
(confmat<- xtabs( ~ Target + adjmodel5.Prediction,
       data = co))

##       adjmodel5.Prediction
## Target   0   1
##      0 119  17
##      1  18 346
```

The accuracy of our procedure is calculated by dividing the number of correct predictions by the total number of predictions. The model is calculated to be 93% accurate at predicting the correct target variable.

$$\frac{119 + 346}{119 + 18 + 17 + 346} = \frac{465}{500} = 0.93$$

The sensitivity of the procedure determines how often the model correctly predicts when the target variable is 1. The model used is calculated to have a sensitivity of 95.05%, meaning when the target variable is 1, the model predicts a 1 95.05% of the time.

$$\frac{346}{18 + 346} = \frac{346}{364} = 0.9505$$

The specificity of the procedure determines how often the model correctly predicts when the target variable is 0. The model is calculated to have a specificity of 87.5%, meaning that when the target variable is 0, the model predicts a 0 87.5% of the time.

$$\frac{119}{119+17} = \frac{119}{136} = 0.875$$

Section 4 - Score Function

The following function can be used to apply the binomial model created above, to a new data set containing the same variables. The function requires the user to pass in the data set, and the function will return the same data set, along with the predicted target value (0 or 1) based on the created binomial model.

```
score <- function(dta) {
  Outstate <- co$Outstate
  phd <- co$PhD
  perc.alumni <- co$perc.alumni
  Apps <- co$Apps
  Expend <- co$Expend
  eta <- -1.2978349723 + 0.0007242341*Outstate + 0.0864645073*perc.alumni + 0
.0002704450*Expend - 0.0835232929*phd - 0.0005358043*Apps
  mu <- 1/(1+exp(-eta))
  co$Prediction <- ifelse(mu>0.5 ,1,0)
  return(co)
}

tmps <- score(co)
```

Section 5 - Conclusion

Although many variables can affect the decision to donate to a college, it is possible to establish a model that determines the target variable well. The Binomial model that was fit to the data uses Outstate, PhD, perc.alumni, Apps, and Expend to predict the outcome variable.

The analysis of this data is based on the classification by educational donors of the target variable. This model may not translate to other educational donors, as the variables used in this model may not be as significant in determining the target variable for others. Nonetheless, the techniques used in this report are applicable to other college data sets assembled by the same educational donors.