BAS 320 - Assignment 6 - Multiple Regression Part 1

Jack Guptill

## Analysis of Wins for College Football Teams

I’m using a multiple regression model to predict the number of wins for a college football teams by knowing and understanding the amount of redzone scores, redzone points, pass yards per game allowed, the average pass yards attempted, and the percentage of redzone attempts.

Overall, I am interested in how well you could predict wins from an offensive and defensive statistic perspective and how narrow of a prediction interval of wins you could get. I am also interested in what predictor variable on its own out of the ones I have chosen have the most weight when it comes to teams winning games.

The data I am using comes from the BAS320 Fall dataset and the data frame itself contains 970 rows with 145 variables. I have reduced the number of columns to a subset containing 6 columns for my analysis.

N = lm(Win~. , data=FOOTBALL) #using a subset of the college football df  
  
summary(N)

##   
## Call:  
## lm(formula = Win ~ ., data = FOOTBALL)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.9514 -1.1667 -0.0957 1.1041 5.1801   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.859397 2.033534 0.423 0.67267   
## Redzone.Scores 0.178709 0.056083 3.187 0.00149 \*\*   
## Redzone.Points 0.164496 2.447079 0.067 0.94642   
## Pass.Yards.Per.Game.Allowed -0.022133 0.001584 -13.969 < 2e-16 \*\*\*  
## Pass.Yards.Attempt 0.464028 0.059783 7.762 2.13e-14 \*\*\*  
## Redzone.Attempts 0.002973 0.046787 0.064 0.94935   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.729 on 964 degrees of freedom  
## Multiple R-squared: 0.7023, Adjusted R-squared: 0.7008   
## F-statistic: 454.8 on 5 and 964 DF, p-value: < 2.2e-16   
  
TO.PREDICT <- data.frame(Redzone.Scores = c(38.88), Redzone.Points = c(.8375),Pass.Yards.Per.Game.Allowed=c(228.4), Pass.Yards.Attempt = c(7.359),Redzone.Attempts=(46.45))  
  
predict(N,newdata=TO.PREDICT,interval="prediction") #Get 95% prediction intervals

## fit lwr upr  
## 1 6.443013 3.048636 9.837389

The regression equation is:

Wins = .859 + .179\*Redzone.Scores + .164\*Redzone.Points -.022\*Pass.Yards.Per.Game.Allowed + .464\*Pass.Yards.Attempt + .003\*Redzone.Attempts

There is a 95% chance that the true amount of wins for a team that contain the mean values for every predictor used in the model is between 3.09 and 9.84 where the most likely amount of wins is approximately 6 games (rounding down to a whole number).

Two otherwise identical teams that differ by 1 redzone scores, are expected to differ in wins by .1787 games, where the team with more redzone scores is expected to have more wins. Redzone scores, in terms of the model, are statistically significant as the p value is less than .05 (.001). Knowing that it is statistically significant, we know that this predictor variable is bringing in new information to the model and is reducing the sum of squared errors.

In contrast, two otherwise identical teams that differ by 1 pass yard per game allowed, are expected to differ by .02 wins, where the team with more pass yards per game allowed are expected to have a lower number of wins. This predictor variable is also statistically significant, meaning that it brings forth new additional information to the model that it wouldn’t necessarily have if it wasn’t included.

Having a partial f test greater than .05 suggests that the predictor variables used in the test do not bring in any new additional information to the model on their own. This could potentially be due to multicollinearity. We can drop both the redzone attempts and redzone points predictor variables as the partial f test was greater than .05 at .9977.

The relationship between wins and each of the predictor variables needs to be linear, homoscedastic (equal spread), and the residuals need to be approximately normal. When evaluating the regression model each predictor shows linear data in an eclipse shape with equal spread along with the qq plot displaying normality of the residuals. With the specific data we are using for the model we are assuming that if there was enough duplicate values that the overall model would have a p-value greater than .05. We can make this assumption given that the each individual plot for each predictor value and the overall model seem to have linear data.