Predictive Model

Jennifer Brann

11/4/2018

```
strikeouts <- read.csv("strikeouts.csv")
```

I began by evaluating the dataset and looking at the different variables that would help predict strikeouts. I am looking at the structure of the data and making sure that the data set is ready to begin analyzing.

```
str(strikeouts)
```

```
names(strikeouts)
```

```
sum(is.na(strikeouts))
```

Within the design on my model, I plan on creating a model that had all the variables to see which contributed more to strikeout rate. The variables that didn't significantly contribute to strikeouts at a 0.05 level I removed.

```
trend1 <- glm(K.~ G + IP + ERA + FIP + xFIP + AVG + BB. + Swing. + Contact. + GB. + LD. + FB.,
              strikeouts,
              family = gaussian("identity"))
summary(trend1)
```

```
##
## Call:
## glm(formula = K. ~ G + IP + ERA + FIP + xFIP + AVG + BB. + Swing. +
##     Contact. + GB. + LD. + FB., family = gaussian("identity"),
##     data = strikeouts)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.022457  -0.005979  -0.000502   0.005214   0.038916
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.836e+00  1.053e+00   1.743  0.08239 .
## G           -5.860e-05  6.791e-05  -0.863  0.38896
## IP          -4.411e-05  2.288e-05  -1.928  0.05489 .
## ERA          1.670e-03  8.069e-04   2.069  0.03943 *
## FIP         -4.839e-04  1.012e-03  -0.478  0.63302
## xFIP        -8.463e-02  1.716e-03 -49.326  < 2e-16 ***
## AVG         -3.291e-04  2.895e-02  -0.011  0.99094
## BB.          8.625e-01  3.500e-02  24.643  < 2e-16 ***
## Swing.      -4.512e-02  2.105e-02  -2.144  0.03288 *
## Contact.    -5.468e-02  1.850e-02  -2.956  0.00337 **
## GB.         -1.417e+00  1.054e+00  -1.345  0.17979
## LD.         -1.394e+00  1.053e+00  -1.324  0.18652
## FB.         -9.924e-01  1.053e+00  -0.942  0.34687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.573741e-05)
##
```

```
##     Null deviance: 1.239913  on 295  degrees of freedom
## Residual deviance: 0.024264  on 283  degrees of freedom
## AIC: -1917.1
##
## Number of Fisher Scoring iterations: 2
```

I continued to remove variables with the higher significant levels one by one until I found a model that all the variables have a lower signicant level (p-value) than 0.01. I decided to use the elimination process because as you take more variables out of the model the significant levels and coefficients for each variable change.

```
trend2 <- update(trend1, .~. -G -FIP -AVG -GB. -LD. -FB.)
summary(trend2)
```

```
##
## Call:
## glm(formula = K. ~ IP + ERA + xFIP + BB. + Swing. + Contact.,
##     family = gaussian("identity"), data = strikeouts)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.082107  -0.014952   0.001255   0.016188   0.075008
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.575e-01  4.766e-02  15.896  < 2e-16 ***
## IP          -4.005e-05  5.741e-05  -0.698    0.486
## ERA         -7.107e-04  1.374e-03  -0.517    0.605
## xFIP        -4.447e-02  3.061e-03 -14.529  < 2e-16 ***
## BB.          4.111e-01  9.127e-02   4.504 9.69e-06 ***
## Swing.       7.808e-02  5.721e-02   1.365    0.173
## Contact.    -5.307e-01  4.343e-02 -12.218  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.0007259144)
##
##     Null deviance: 1.23991  on 295  degrees of freedom
## Residual deviance: 0.20979  on 289  degrees of freedom
## AIC: -1290.6
##
## Number of Fisher Scoring iterations: 2
```

```
trend3 <- update(trend2, .~.-ERA)
summary(trend3)
```

```
trend4 <- update(trend3, .~. -IP)
summary(trend4)
```

```
trend5 <- update(trend4, .~. -Swing.)
summary(trend5)
```

```
final <- lm(K.~ xFIP + BB. + Contact., strikeouts)
summary(final)
```

```
##
## Call:
## lm(formula = K. ~ xFIP + BB. + Contact., data = strikeouts)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.072819 -0.016014  0.001516  0.016195  0.079912
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.813172   0.027313  29.772  < 2e-16 ***
## xFIP        -0.044220   0.002841 -15.564  < 2e-16 ***
## BB.          0.351136   0.075148   4.673 4.55e-06 ***
## Contact.    -0.557141   0.040849 -13.639  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02696 on 292 degrees of freedom
## Multiple R-squared:  0.8289, Adjusted R-squared:  0.8271
## F-statistic: 471.4 on 3 and 292 DF,  p-value: < 2.2e-16
```

Trend 5 is my final model. The model is K% = 0.8132 -0.0442* xFIP + 0.3511 *BB.* - *0.5571*Contact.

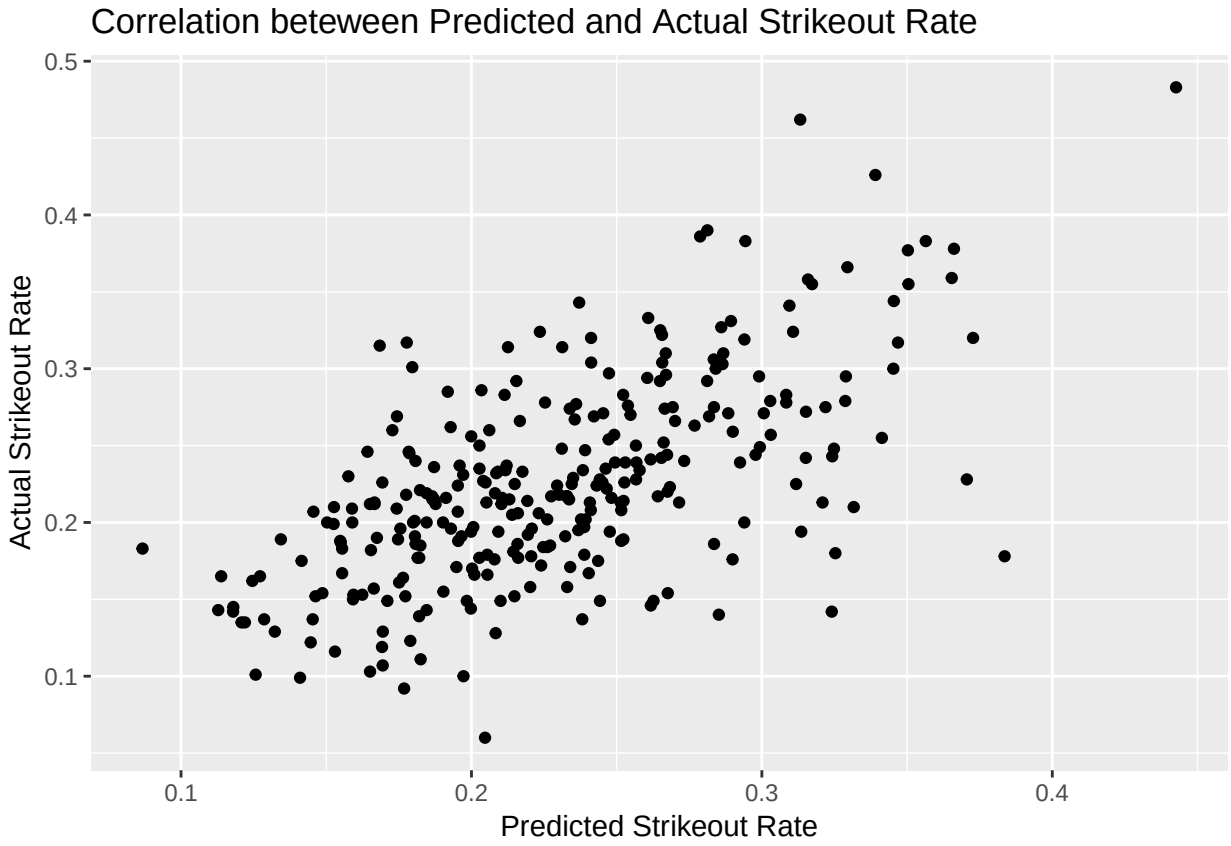
Key numbers to notice is that R^2 is equal to 0.829 (R equals 0.91), RSE equals 0.027, and all the p-values are lower than 0.01. The high R^2/ R value (correlation value) indicates that the variables chosen strongly correlates with strikeout rate. The low RSE (residual standard error) indicates that the model fits very well with the strikeout rate data. For reference, if RSE equals 0 then the model would fit the data perfectly. All the very low p-values mean that these three variables are significant when determining strikeout rate. All of these key numbers help support that these three variables create a strong model for predicting the strikeout rate for the second half of the season.

Now to predict the strikeouts for the second half. The variable "Predict" are my predictions for the strikeout rate for the second half the season.

```
Predict <- predict(trend5, strikeouts, type = "response")
```

Here I'm creating a plot that graphs my predictions verus the actual strikeout rate for the second half. The plot visually shows how closely related my predictions were to the real strikeout rate for the second half.

```
ggplot(strikeouts) +
  geom_point(aes(x= Predict, y=X2ndHalfK.)) +
  labs(x= "Predicted Strikeout Rate",
       y= "Actual Strikeout Rate",
       title = "Correlation beteween Predicted and Actual Strikeout Rate")
```



Next, I created a linear regression between the actual rate and my predicted rate. The results from the regression will show how well my model did at predicting the strikeout rate.

```
accuracy <- lm(X2ndHalfK. ~ Predict, strikeouts)
summary(accuracy)
```

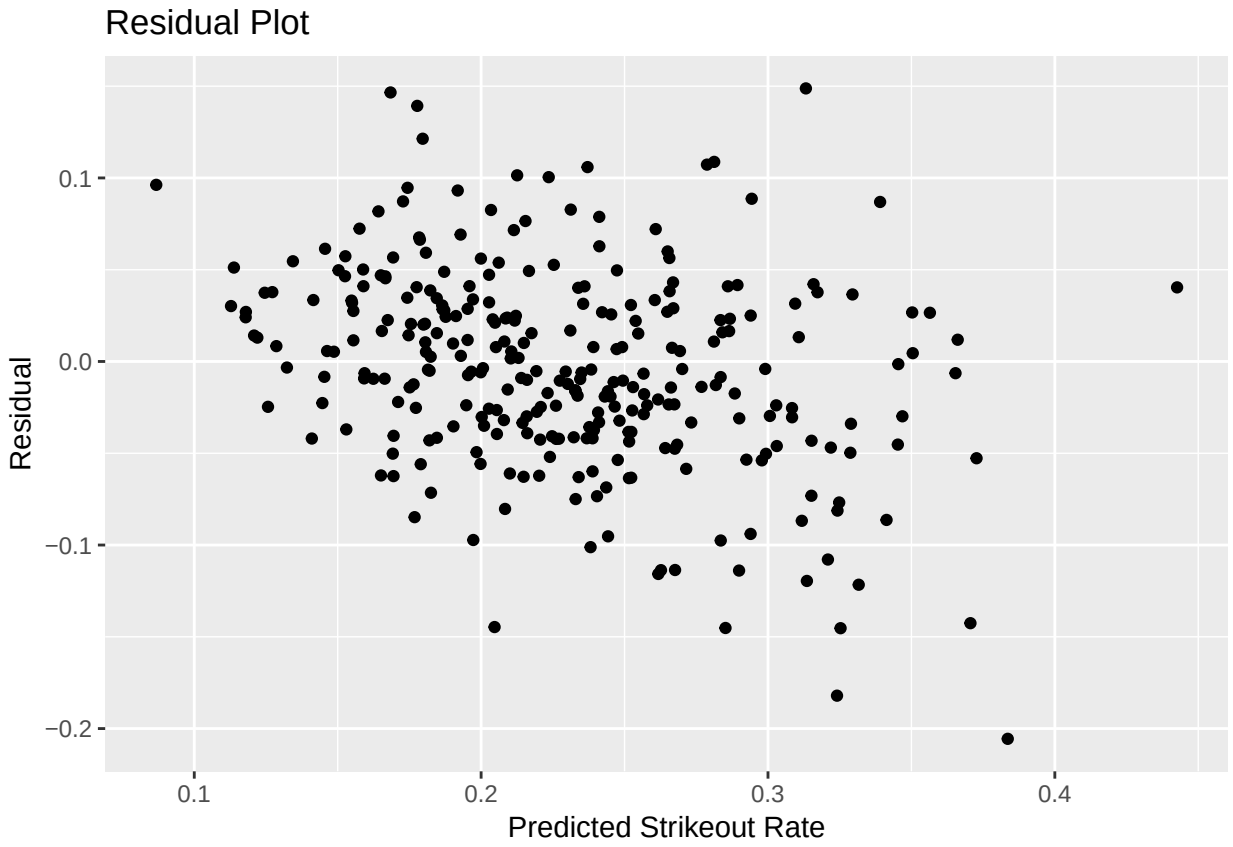
```
##
## Call:
## lm(formula = X2ndHalfK. ~ Predict, data = strikeouts)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.155882 -0.030154 -0.002369  0.031538  0.177541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06446    0.01207   5.342 1.85e-07 ***
## Predict      0.70239    0.05098  13.779  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05168 on 294 degrees of freedom
## Multiple R-squared:  0.3924, Adjusted R-squared:  0.3903
## F-statistic: 189.9 on 1 and 294 DF,  p-value: < 2.2e-16
```

The R^2 value is 0.3924 which means that my model explains 39.24% of the varaiation in the data. The R-value is 0.626. The R-value demonstrates the strength of the correlation between the predict and actual rate. For reference, the closer R-value is to 1 the better the model fits the actual strikeout rate for the second

half. An R-value of 0.626 demonstrates that my model correlates with the actual strikeout rate. Another number to notice is the RSE is equal to 0.052. Again, the low RSE demonstrates that predictions help accurately predict strikeout rate for the second half.

I included a residual plot to visually show that the residuals don't have a pattern. A residual is a difference between the actual strikeout rate and the expected (predicted) strikeout rate. The residual value is how much my prediction is off from the actual rate. No pattern in the residual plot means that the linear regression is appropaite for the data.

```
Residual <- strikeouts$X2ndHalfK. - Predict
ggplot(strikeouts, aes(x= Predict, y=Residual)) +
  geom_point() +
  labs(x= "Predicted Strikeout Rate",
       y= "Residual",
       title = "Residual Plot")
```



Conclusion:

Based on the evaluation, my final predictive model is K% = 0.8132 -0.0442*xFIP + 0.3511* BB. - 0.5571*Contact.

When deciding on which model to choose, I only chose the factors that had a signifacnt value (p-value) lower than 0.01. Based on the multiple linear regression model I chose, I found that the correlation value is 0.91 which demonstrates that these factors are strongly correlated with strikeout percentage. Along with the high R value, each variable has p-values lower than 0.001 meaning that these variables are significant in correlating the model with strikeout rate.

After creating my predictions for the second half, I created a linear regression between my predictions and the actual strikeout rate. With this regression, I got an R value of 0.626 and a low residual standard error value. Both of these results indicts that there is an association between my predictions for the second half and real strikeout rate.