Residual Report

Eric Polverari, RJ Leon, Matthew DiLeonardo & Cristian Hendricks

4/28/2021

```
model91 <- lm (woba ~ I(isolated_power^2) + sprint_speed + b_hit_line_drive + isolated_power
               + popups_percent + straightaway_percent + in_zone_swing, data = statcast3)
summary(model91)
```

```
## 
## Call:
## lm(formula = woba ~ I(isolated_power^2) + sprint_speed + b_hit_line_drive + 
##     isolated_power + popups_percent + straightaway_percent + 
##     in_zone_swing, data = statcast3)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.049580 -0.010630  0.000658  0.011144  0.067937 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          1.731e-01  1.717e-02  10.083  < 2e-16 ***
## I(isolated_power^2)  4.126e-01  1.663e-01   2.481  0.01334 *  
## sprint_speed         1.470e-03  4.701e-04   3.127  0.00184 ** 
## b_hit_line_drive     8.210e-04  4.989e-05  16.457  < 2e-16 ***
## isolated_power       3.534e-01  6.589e-02   5.364 1.11e-07 ***
## popups_percent      -2.431e-03  2.620e-04  -9.280  < 2e-16 ***
## straightaway_percent  6.051e-04  2.111e-04   2.867  0.00427 ** 
## in_zone_swing       -2.662e-05  8.105e-06  -3.285  0.00107 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.01715 on 700 degrees of freedom
## Multiple R-squared:  0.7482, Adjusted R-squared:  0.7457 
## F-statistic: 297.1 on 7 and 700 DF,  p-value: < 2.2e-16
```

Best Model

The “best” model our group has found so far includes the variables sprint speed (the velocity a player generates when running from home base to first base in the STATCAST tracking data), line drive (a powerfully hit ball that travels in the air and relatively close to and parallel with the ground.), isolated power (measures the raw power of a hitter by taking only extra-base hits – and the type of extra-base hit – into account.), popups percent (total number of weakly hit balls divided by total number of batting attempts), straightaway percent (percent of the time a batter hits the ball towards centerfield), in zone swing (number of times a batter swings at a pitch located in the strike zone), as well as a quadratic term for isolated power.

From our tests, we can see that each variable reports a significant p-value, the residual standard error shows that 95% of the values will fall between 2 standard errors, the adjusted R-squared shows a relatively high value of 0.7457 and we get a significant p-value < 2.2e-16 for the model.

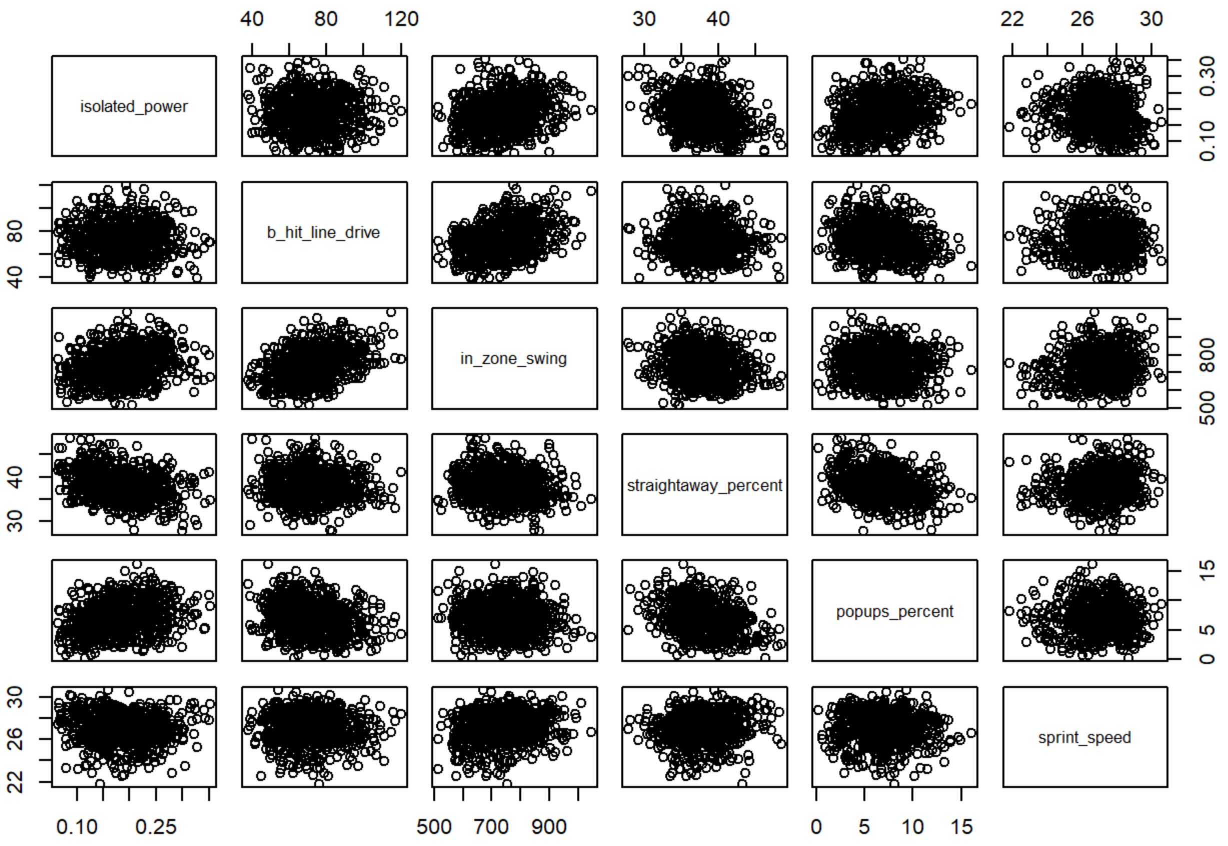
Does out “best” model safisfy the assumptions of regression inference?

To judge if our model satisfies the assumptions of regression inference, we'll begin with checking for any regression pitfalls. After deleting variables from the dataset that came back not estimable, we checked for evidence of multi-collinearity by running pairwise with correlation and scatterplot matrices.

```
## read in data from excel

library(readxl)

## new variable
statcast_cor_new <- read_excel("statcast cor new.xls")



## correlation matrices

cor(statcast_cor_new)
```

```
##                       isolated_power b_hit_line_drive in_zone_swing
## isolated_power           1.000000000      0.003935993    0.25403417
## b_hit_line_drive         0.003935993      1.000000000    0.38512368
## in_zone_swing            0.254034173      0.385123682    1.00000000
## straightaway_percent    -0.304988743     -0.055893876   -0.13430126
## popups_percent           0.249750707     -0.207752347   -0.04051902
## sprint_speed            -0.078337657     -0.019320147    0.17148716
##                       straightaway_percent popups_percent sprint_speed
## isolated_power                 -0.30498874     0.24975071  -0.07833766
## b_hit_line_drive               -0.05589388    -0.20775235  -0.01932015
## in_zone_swing                  -0.13430126    -0.04051902   0.17148716
## straightaway_percent            1.00000000    -0.32195598   0.11533353
## popups_percent                 -0.32195598     1.00000000  -0.01997025
## sprint_speed                    0.11533353    -0.01997025   1.00000000
```

```
## scatterplot matrices

plot(statcast_cor_new)
```



Looking at the pairwise with correlation matrices, we can see that no two variables share any potential collinearity.

From our plot matrices, we get a lot of blobby looking plots, which indicates there’s not much, if any, correlation between any two variables.

Next, we’ll calculate VIF scores to furthur check for multi-collinearity,

```
# regression to get first order terms that need to be checked for collinearity
model90 <- lm (woba ~  sprint_speed + b_hit_line_drive + isolated_power
               + popups_percent + straightaway_percent + in_zone_swing, data = statcast3)
summary(model90)
```

```
##
## Call:
## lm(formula = woba ~ sprint_speed + b_hit_line_drive + isolated_power +
##     popups_percent + straightaway_percent + in_zone_swing, data = statcast3)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.048964 -0.010949  0.000682  0.010872  0.066739
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.554e-01  1.568e-02   9.915  < 2e-16 ***
## sprint_speed          1.637e-03  4.670e-04   3.505 0.000485 ***
## b_hit_line_drive      8.229e-04  5.006e-05  16.437  < 2e-16 ***
## isolated_power        5.138e-01  1.284e-02  40.009  < 2e-16 ***
## popups_percent       -2.486e-03  2.620e-04  -9.488  < 2e-16 ***
## straightaway_percent  6.178e-04  2.118e-04   2.917 0.003647 **
## in_zone_swing        -2.842e-05  8.102e-06  -3.508 0.000481 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01721 on 701 degrees of freedom
## Multiple R-squared:  0.746,  Adjusted R-squared:  0.7438
## F-statistic: 343.1 on 6 and 701 DF,  p-value: < 2.2e-16
```

```
## library

library(car)

## vif

vif(model90)
```

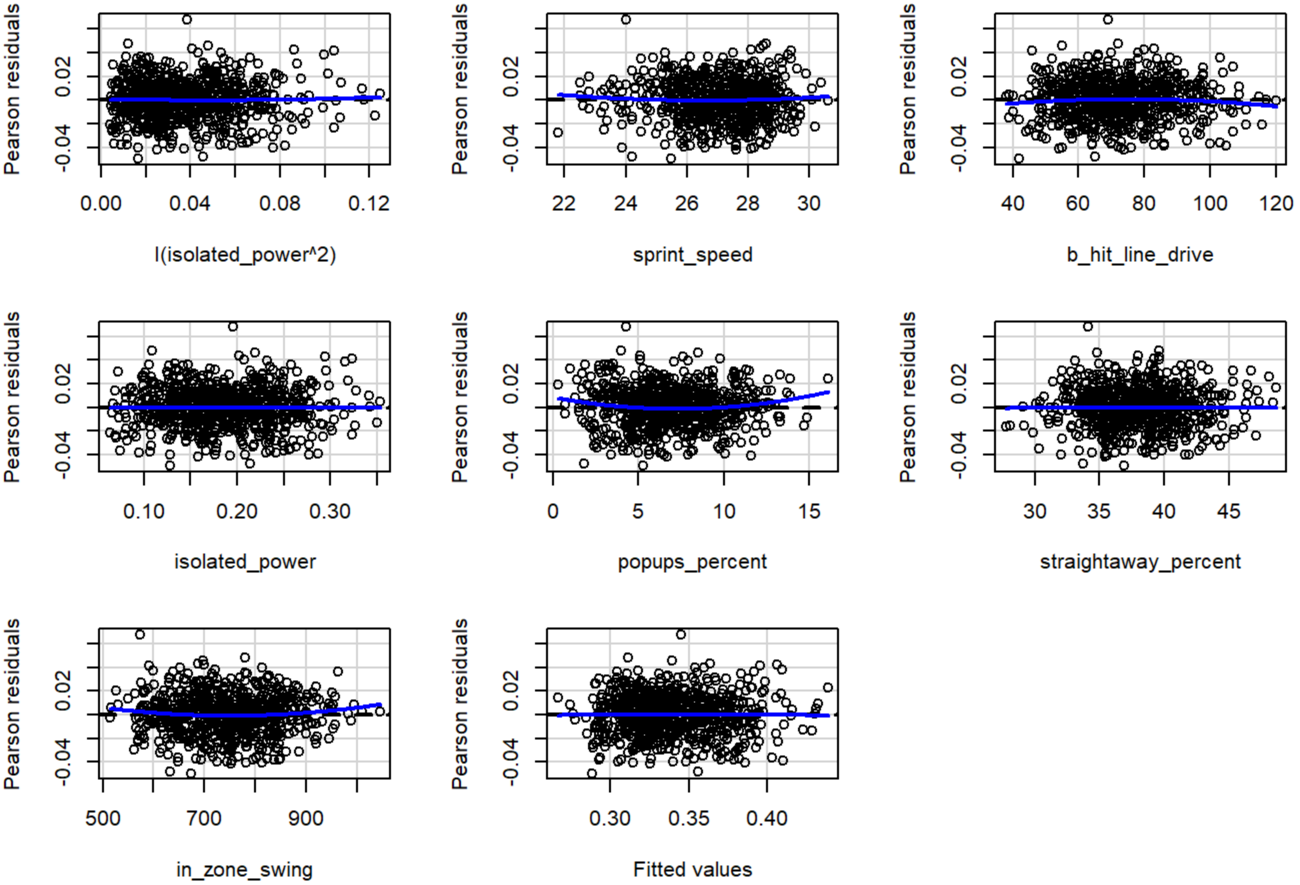
```
##         sprint_speed     b_hit_line_drive       isolated_power
##             1.073726             1.253170             1.227783
##       popups_percent straightaway_percent        in_zone_swing
##             1.221841             1.223828             1.345639
```

The above VIF scores are all significantly less than our arbitrary cutoff point of 10.

After looking at matrices and VIF scores to check for multi-collinearity, we have found no evidence to suggest multi-collinearity exist in our model. Thus, we’ll proceed to check the residuals for our “best” model.

Residuals

```
# residual plots
residualPlots(model91)
```

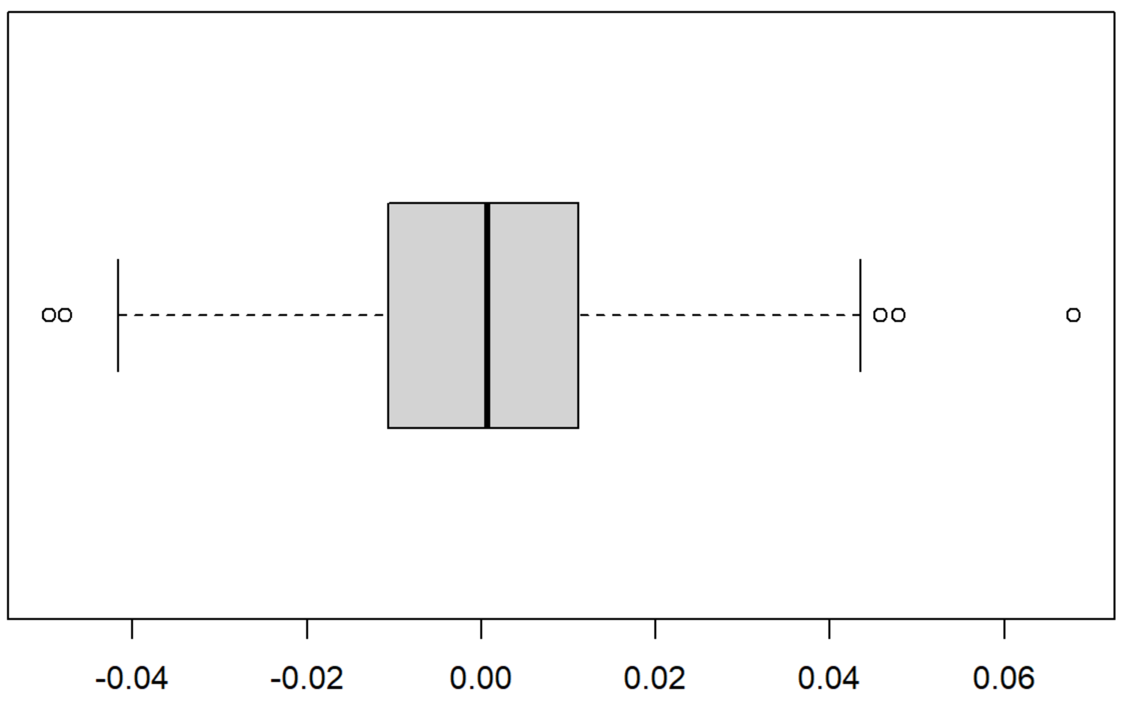


```
##                      Test stat Pr(>|Test stat|)
## I(isolated_power^2)     1.1672         0.243544
## sprint_speed            0.9592         0.337812
## b_hit_line_drive       -1.2855         0.199044
## isolated_power         -2.0315         0.042581 *
## popups_percent          2.7736         0.005691 **
## straightaway_percent   -0.1869         0.851771
## in_zone_swing           1.8624         0.062966 .
## Tukey test             -0.2032         0.838981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

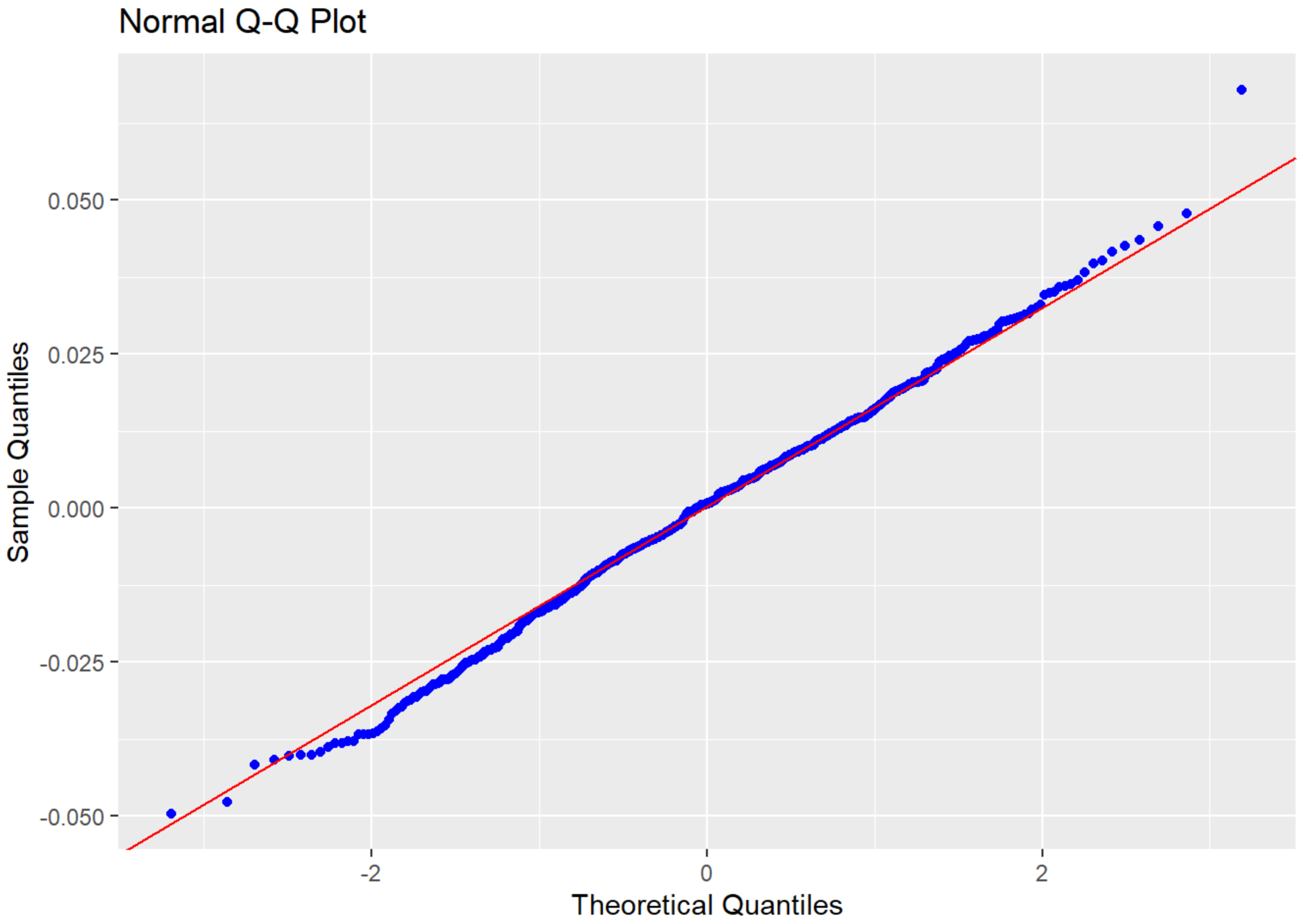
The residual plots show a pretty even distribution of values above and below the line, which is what we were hoping to see. Looking at the fitted values plotted against the residuals, we see the points roughly fit into a rectangular shape, indicating our model meets the homoscedasticity requirement.

Now we want to check a boxplot to see if there's any reason to suggest non-linearity and a Q-Q plot to see if there's linearity.

```
## boxplot

boxplot(model91$residuals, horizontal=TRUE)
```



```
## Library to get olsrr
library(olsrr)

## Q-Q Plot
ols_plot_resid_qq(model91)
```

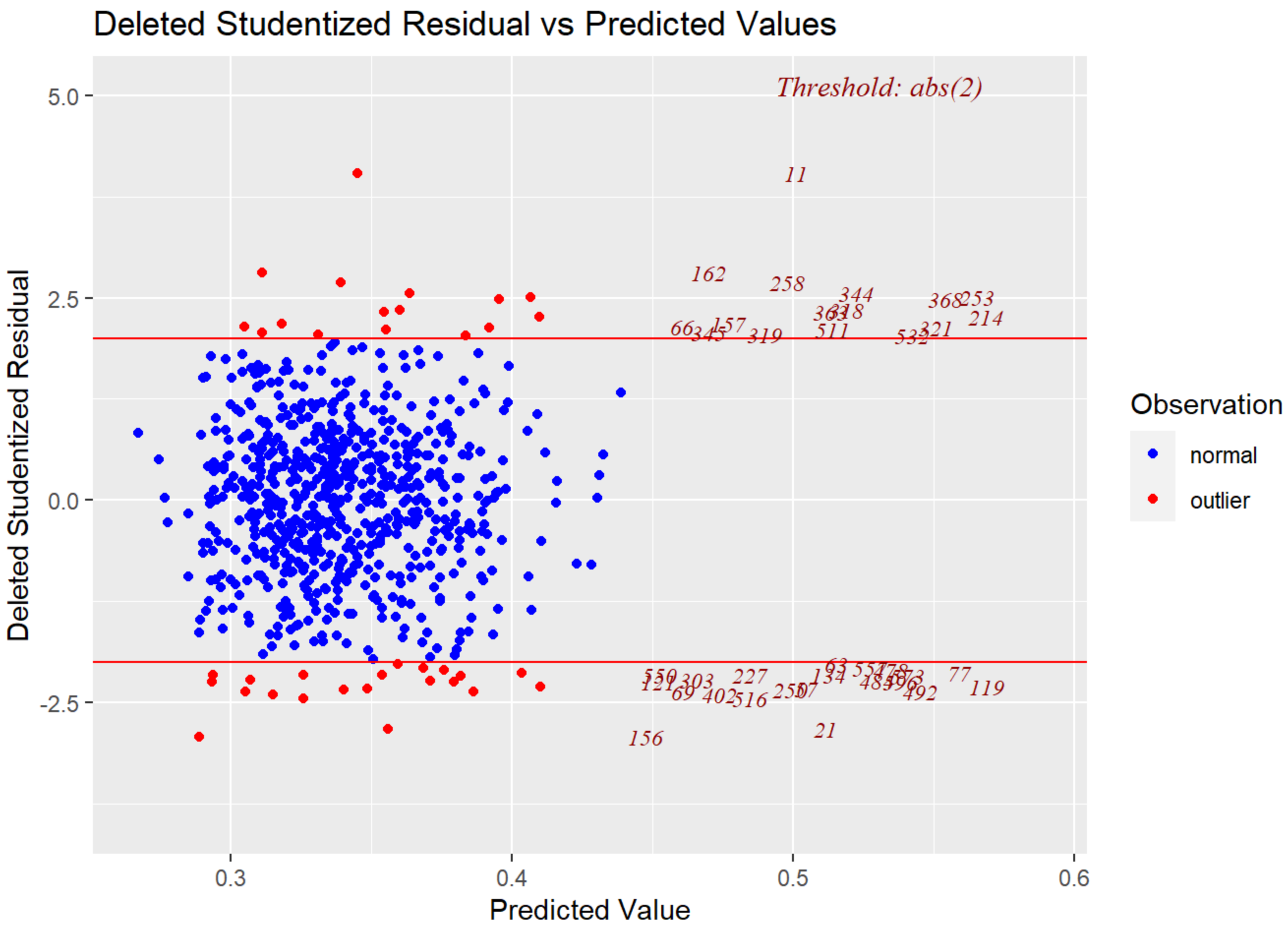


The boxplot doesn't show any signs of nonlinearity and the Q-Q Plot is showing a strong linear dataset. Both these graphics suggest that our model reasonably meets the normal residuals requirement.

Outliers, High Leverage & Influential points

Next, we'll investigate any outliers, points with high leverage and influential points.

```
## outliers
ols_plot_resid_stud_fit(model91)$outliers
```

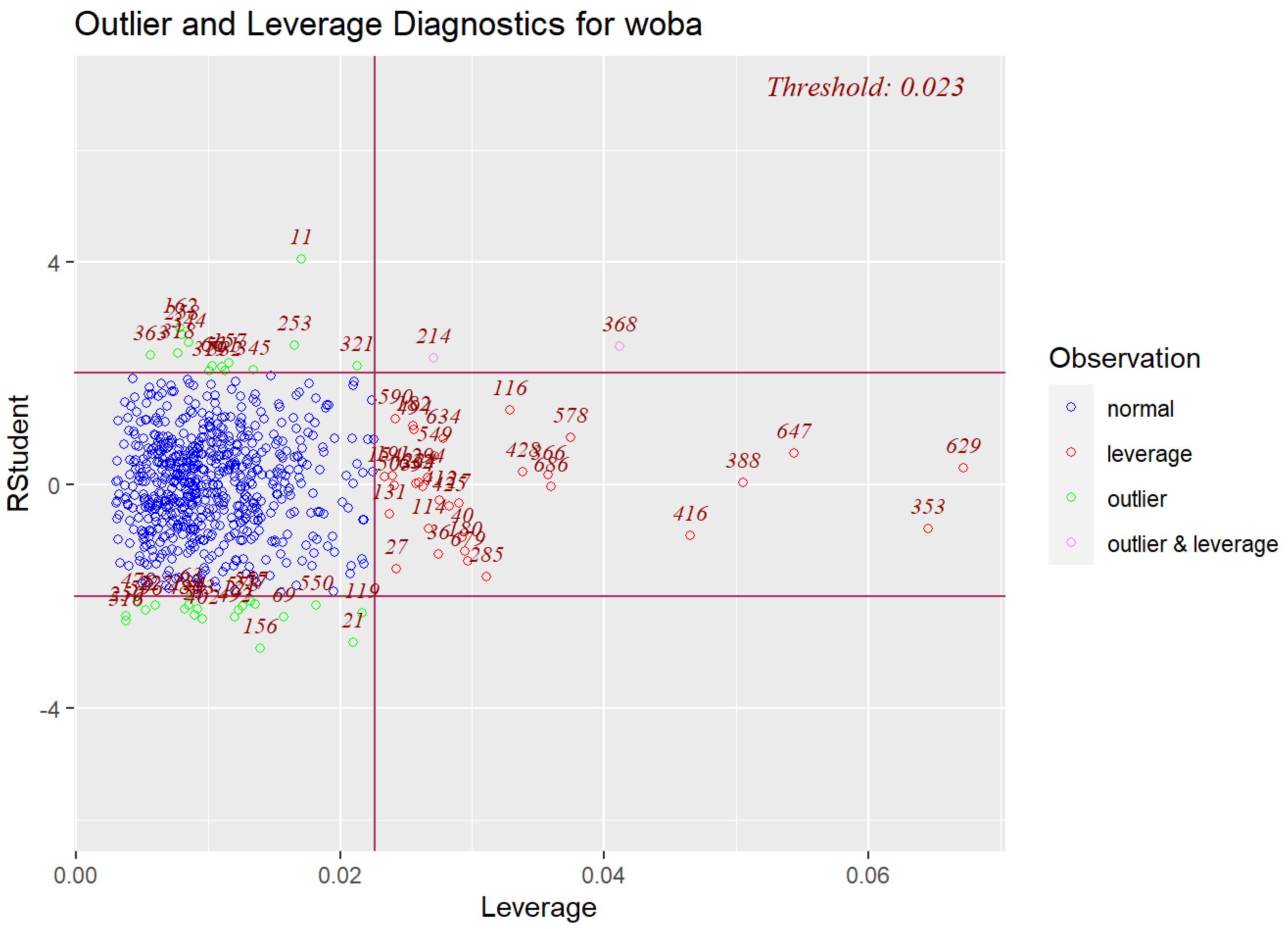


```
## NULL
```

First, by looking at the studentized deleted residual plot, we'll notice some outliers, most notably point 11. However, the data points that are deemed outliers don't stray too far away from the rest of the points, so we feel confident that it won't significantly affect our model and we can include it in our statistical analysis.

Next, we'll look at the resid-leverage plot.

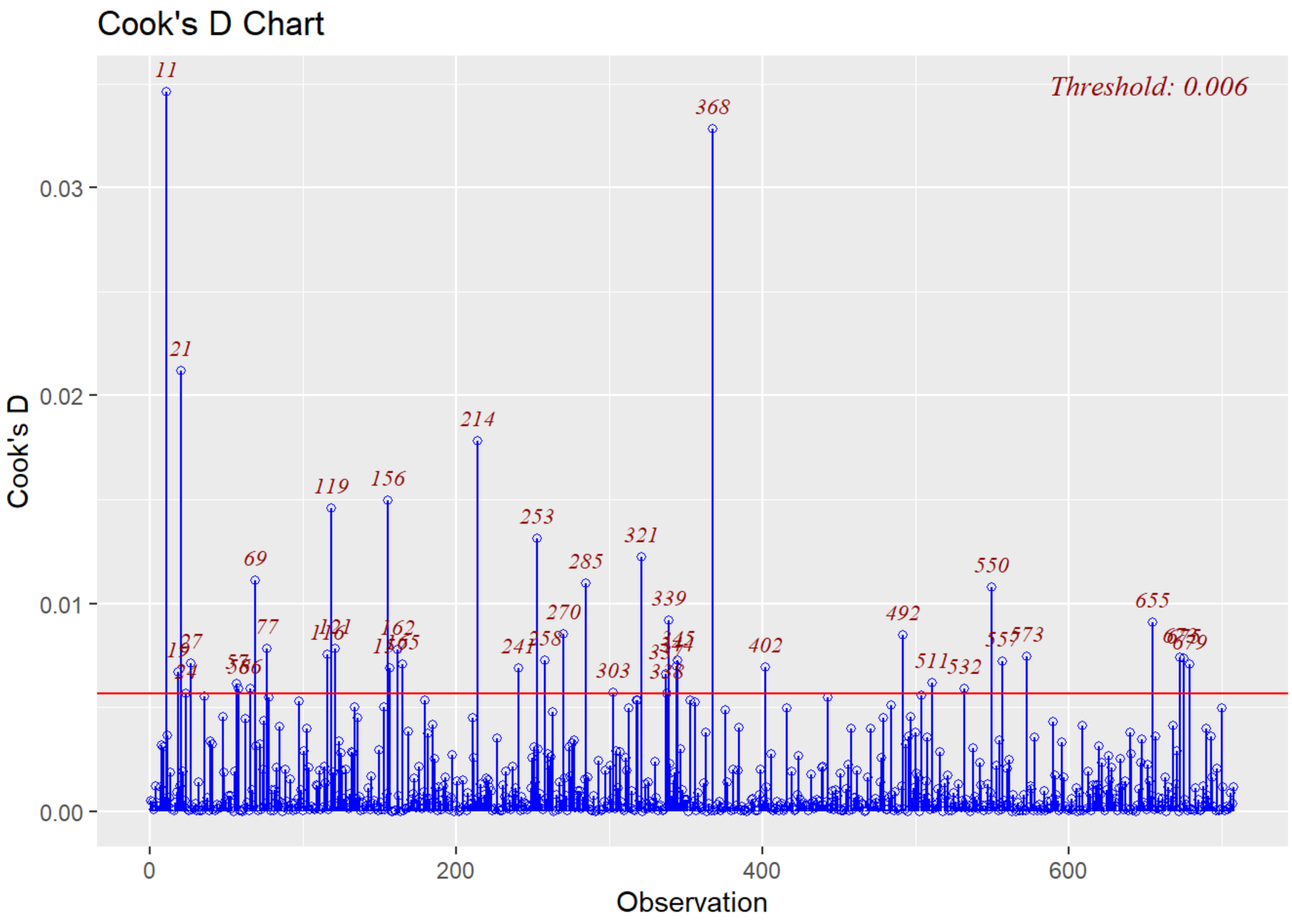
```
## resid-leverage plot
ols_plot_resid_lev(model91)
```



Our focus is in the upper right corner, where we see two points of interest (214 and 368) that are both an outlier & leverage point. However, these points aren't significantly far away from the rest of the dataset and we feel confident that they won't affect our model.

Lastly, we'll look at a Cooks D plot.

```
## Cooks D plot
ols_plot_cooksd_chart(model91)
```



Two data points stick out like a sore thumb (11 & 368). While these values appear to be quite different from the rest, they aren't having a heavy enough influence to remove from the dataset.

To further investigate any problematic outliers or leverage points, we were able find the players that the values belonged to. Point 11 is for Minnesota Twins Outfielder Byron Buxton for the 2017 season. Considered one of the fastest players in baseball, Buxton's sprint score may explain why he has a higher wOBA than expected, as the formula in wOBA doesn't really take into account speed. Point 368 belongs to Catcher for the, at the time, Toronto Blue Jays, Russel Martin, for the year 2015. Martin has a reputation for having a good eye for the ball, so perhaps his ability to avoid popups and reluctance to swing too much makes him a player that is better than the traditional wOBA statistic suggests.

Based on researching the values and critically thinking about how these player's values may not fit our dataset, we determined that there could be a reasonable explanation as to why the points are noticeably considered "off" from the rest of the data and since these values don't seem to be unexplainable or significantly different from the rest of the data, we decided to not remove them or any other data points.

After doing a thorough investigation of residuals, we are confident that our model satisfies the assumptions of regression inference. Therefore, we will consider this our "best" model.

Conclusion

In order to process our model refining, we needed skills that we learned from lower-division classes, such as Intro to Philosophy: Critical Thinking. Without sharp critical thinking skills, we could have easily gone down the wrong path in regards to which variables to choose and may not have been able to avoid regression pitfalls, which would be easy to do without a refined skillset of logical reasoning. It was also crucial for all of us to have had basic coursework in Statistics, mainly from Math 165 and 265, which gave us a great foundation to help us with this process. As residuals and outliers are covered in both classes, we felt like we had a really good sense of how to interpret outliers or influential plots, especially visually, as we spent a good deal of time examining different plots, graphs and statistical tests in our beginner Statistics courses. Without that foundation, it would've been easy for us to blindly follow the residual test that showed us our outliers and remove them from the data, but since we've sharpened our tools of Statistical learning, we were able to properly asses whether the data points were actually problematic or not. To creatively build our “best” model, it was imperative for us to start from the ground up before applying more complex rules and techniques to our model. It took a combination of years of experience and important classes that shaped our ability to find our “best” model.