How Does Velocity Affect Fastball Success?

Luke Stageberg

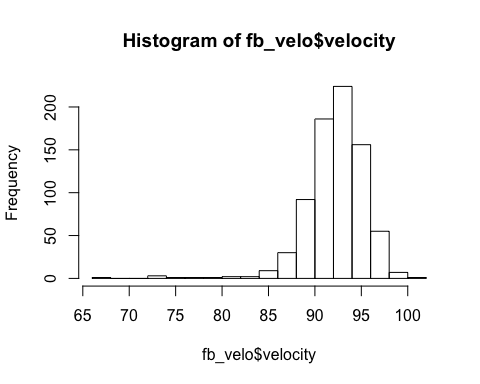
April 16, 2019

Naming the dataset “fb\_velo” and importing it from my data folder in my Fastball\_Velocity folder

fb\_velo = read.csv('../data/Fastball\_Data.csv')

Creating a histogram to show what the average fastball velocity’s were across the MLB in 2018

hist(fb\_velo$velocity, breaks = 20)



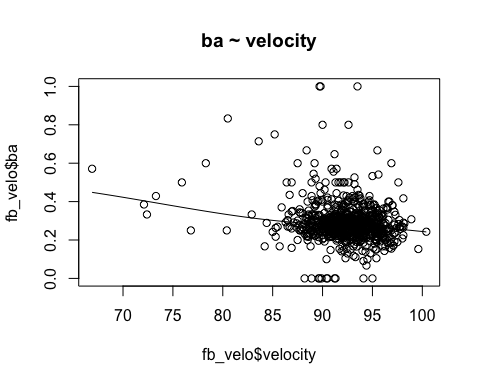
The first model created uses velocity as a predictor variable for batting average.

mod\_velo\_ba = lm(formula = ba ~ velocity, data = fb\_velo)

summary(mod\_velo\_ba)

##   
## Call:  
## lm(formula = ba ~ velocity, data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.31658 -0.05239 -0.01085 0.02828 0.71671   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.870538 0.105564 8.247 7.04e-16 \*\*\*  
## velocity -0.006281 0.001143 -5.495 5.33e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1039 on 769 degrees of freedom  
## Multiple R-squared: 0.03778, Adjusted R-squared: 0.03653   
## F-statistic: 30.19 on 1 and 769 DF, p-value: 5.326e-08

scatter.smooth(x=fb\_velo$velocity, y=fb\_velo$ba, main="ba ~ velocity")



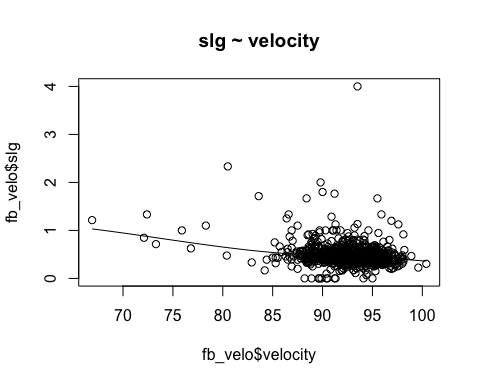
The second model created uses velocity as a predictor variable for sluggling percentage.

mod\_velo\_slg = lm(formula = slg ~ velocity, data = fb\_velo)

summary(mod\_velo\_slg)

##   
## Call:  
## lm(formula = slg ~ velocity, data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5719 -0.1188 -0.0373 0.0641 3.5220   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.134433 0.257532 8.288 5.11e-16 \*\*\*  
## velocity -0.017716 0.002789 -6.353 3.61e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2535 on 769 degrees of freedom  
## Multiple R-squared: 0.04987, Adjusted R-squared: 0.04863   
## F-statistic: 40.36 on 1 and 769 DF, p-value: 3.614e-10

scatter.smooth(x=fb\_velo$velocity, y=fb\_velo$slg, main="slg ~ velocity")



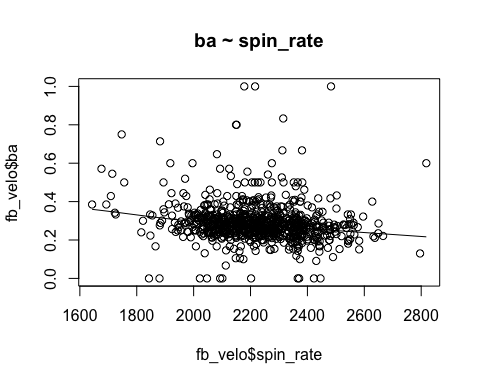
The third model created uses spin rate as a predictor variable for batting average

mod\_spin\_ba = lm(formula = ba ~ spin\_rate, data = fb\_velo)

summary(mod\_spin\_ba)

##   
## Call:  
## lm(formula = ba ~ spin\_rate, data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.32858 -0.05143 -0.01143 0.02646 0.73491   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.114e-01 5.041e-02 10.146 < 2e-16 \*\*\*  
## spin\_rate -9.922e-05 2.261e-05 -4.388 1.3e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1046 on 769 degrees of freedom  
## Multiple R-squared: 0.02443, Adjusted R-squared: 0.02316   
## F-statistic: 19.26 on 1 and 769 DF, p-value: 1.302e-05

scatter.smooth(x=fb\_velo$spin\_rate, y=fb\_velo$ba, main="ba ~ spin\_rate")



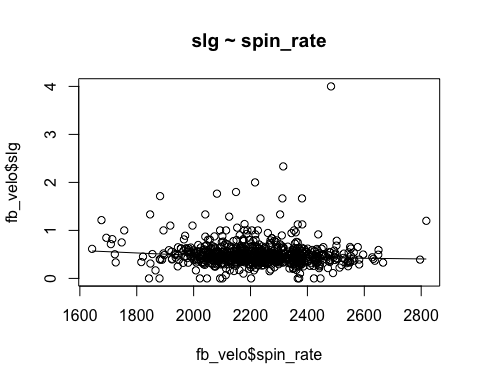
The fourth model created uses spin rate as a predictor variable for sluggling percentage.

mod\_spin\_slg = lm(formula = slg ~ spin\_rate, data = fb\_velo)

summary(mod\_spin\_slg)

##   
## Call:  
## lm(formula = slg ~ spin\_rate, data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5408 -0.1192 -0.0406 0.0610 3.5290   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.419e-01 1.250e-01 5.936 4.42e-09 \*\*\*  
## spin\_rate -1.091e-04 5.606e-05 -1.946 0.052 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2594 on 769 degrees of freedom  
## Multiple R-squared: 0.0049, Adjusted R-squared: 0.003606   
## F-statistic: 3.787 on 1 and 769 DF, p-value: 0.05202

scatter.smooth(x=fb\_velo$spin\_rate, y=fb\_velo$slg, main="slg ~ spin\_rate")



The next step to produce another pair of models was to create a new varible that is the percentage of a swing and miss (dividing whiffs by swings). Any “na” in the data is replaced with “0”

fb\_velo$swing\_and\_miss\_pct = with(fb\_velo, whiffs / swings)  
fb\_velo[is.na(fb\_velo)] <- 0

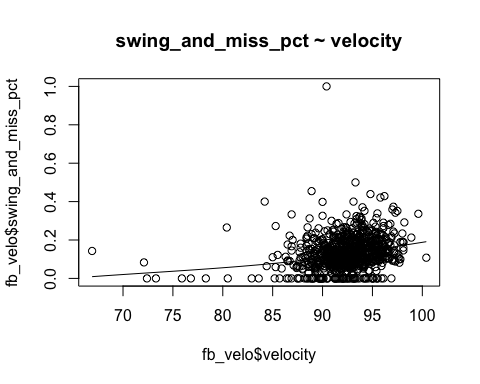
The fifth model created uses velocity as a predictor variable for swing and miss percentage.

mod\_velo\_swing\_and\_miss = lm(formula = swing\_and\_miss\_pct ~ velocity, data = fb\_velo)

summary(mod\_velo\_swing\_and\_miss)

##   
## Call:  
## lm(formula = swing\_and\_miss\_pct ~ velocity, data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.16230 -0.05512 -0.00920 0.04267 0.87870   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.4488132 0.0889291 -5.047 5.61e-07 \*\*\*  
## velocity 0.0063066 0.0009629 6.549 1.06e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08753 on 769 degrees of freedom  
## Multiple R-squared: 0.05283, Adjusted R-squared: 0.0516   
## F-statistic: 42.89 on 1 and 769 DF, p-value: 1.057e-10

scatter.smooth(x=fb\_velo$velocity, y=fb\_velo$swing\_and\_miss\_pct, main="swing\_and\_miss\_pct ~ velocity")



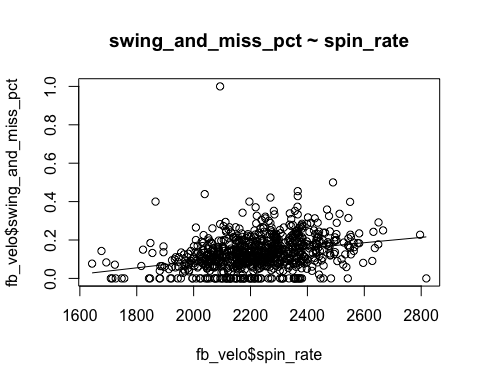
The sixth model created uses spin rate as a predictor variable for swing and miss percentage.

mod\_spin\_swing\_and\_miss = lm(formula = swing\_and\_miss\_pct ~ spin\_rate, data = fb\_velo)

summary(mod\_spin\_swing\_and\_miss)

##   
## Call:  
## lm(formula = swing\_and\_miss\_pct ~ spin\_rate, data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.22513 -0.05061 -0.00761 0.04293 0.88686   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.102e-01 4.151e-02 -5.062 5.19e-07 \*\*\*  
## spin\_rate 1.545e-04 1.862e-05 8.295 4.84e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08616 on 769 degrees of freedom  
## Multiple R-squared: 0.08213, Adjusted R-squared: 0.08094   
## F-statistic: 68.81 on 1 and 769 DF, p-value: 4.835e-16

scatter.smooth(x=fb\_velo$spin\_rate, y=fb\_velo$swing\_and\_miss\_pct, main="swing\_and\_miss\_pct ~ spin\_rate")



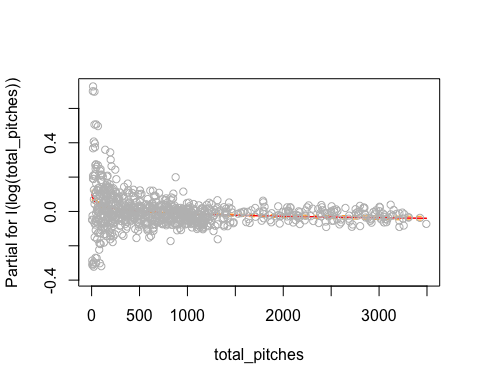
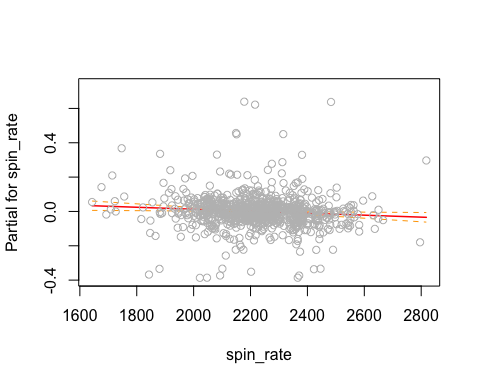
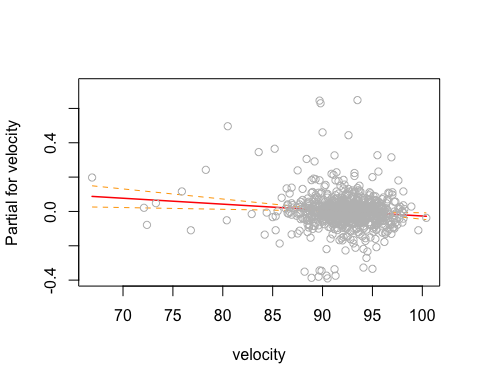
The seventh model created uses velocity, spin rate, and total pitches as a predictor variable for batting average.

mod\_velo\_spin\_ba = lm(formula = ba ~ velocity + spin\_rate + I(log(total\_pitches)), data = fb\_velo)

summary(mod\_velo\_spin\_ba)

##   
## Call:  
## lm(formula = ba ~ velocity + spin\_rate + I(log(total\_pitches)),   
## data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.39841 -0.04228 0.00004 0.03505 0.65259   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.703e-01 1.025e-01 8.491 < 2e-16 \*\*\*  
## velocity -3.449e-03 1.215e-03 -2.838 0.00466 \*\*   
## spin\_rate -5.758e-05 2.345e-05 -2.456 0.01428 \*   
## I(log(total\_pitches)) -2.118e-02 2.959e-03 -7.157 1.93e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1003 on 767 degrees of freedom  
## Multiple R-squared: 0.1054, Adjusted R-squared: 0.1019   
## F-statistic: 30.14 on 3 and 767 DF, p-value: < 2.2e-16

termplot(mod\_velo\_spin\_ba,partial.resid = T, se = T)



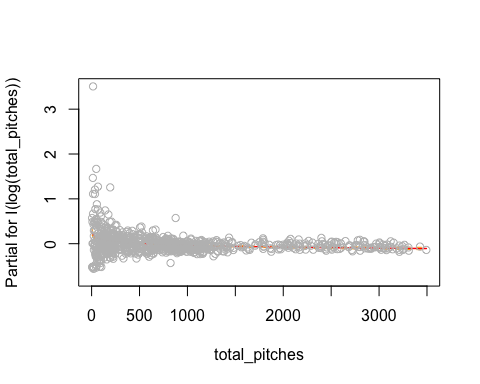
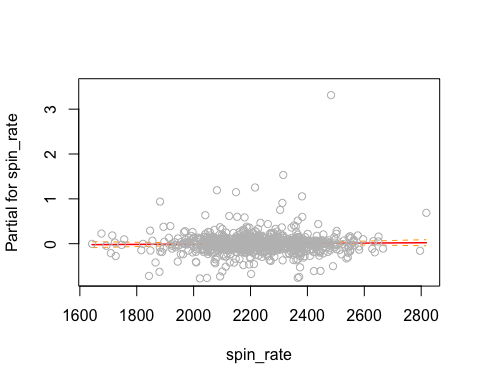
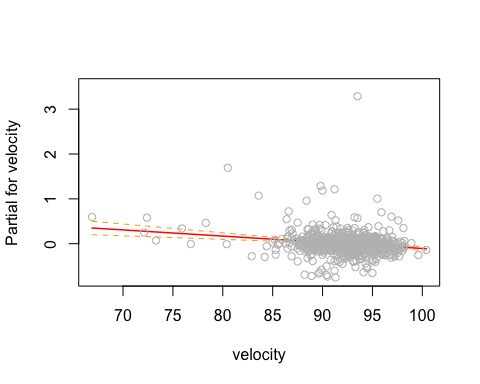
The eighth model created uses velocity, spin rate, and total pitches as a predictor variable for slugging percentage.

mod\_velo\_spin\_slg = lm(formula = slg ~ velocity + spin\_rate + I(log(total\_pitches)), data = fb\_velo)

summary(mod\_velo\_spin\_slg)

##   
## Call:  
## lm(formula = slg ~ velocity + spin\_rate + I(log(total\_pitches)),   
## data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7665 -0.0982 -0.0034 0.0812 3.3032   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.050e+00 2.492e-01 8.225 8.31e-16 \*\*\*  
## velocity -1.379e-02 2.955e-03 -4.667 3.60e-06 \*\*\*  
## spin\_rate 3.685e-05 5.701e-05 0.646 0.518   
## I(log(total\_pitches)) -5.724e-02 7.196e-03 -7.955 6.42e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2439 on 767 degrees of freedom  
## Multiple R-squared: 0.1225, Adjusted R-squared: 0.1191   
## F-statistic: 35.7 on 3 and 767 DF, p-value: < 2.2e-16

termplot(mod\_velo\_spin\_slg,partial.resid = T, se = T)



The ninth model created uses velocity, spin rate, and total pitches as a predictor variable for swing and miss percentage.

mod\_velo\_spin\_swing\_and\_miss = lm(formula = swing\_and\_miss\_pct ~ velocity + spin\_rate + I(log(total\_pitches)), data = fb\_velo)

summary(mod\_velo\_spin\_swing\_and\_miss)

##   
## Call:  
## lm(formula = swing\_and\_miss\_pct ~ velocity + spin\_rate + I(log(total\_pitches)),   
## data = fb\_velo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.19798 -0.05239 -0.01186 0.03870 0.94346   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.837e-01 8.513e-02 -5.681 1.90e-08 \*\*\*  
## velocity 2.664e-03 1.009e-03 2.639 0.00849 \*\*   
## spin\_rate 1.229e-04 1.947e-05 6.310 4.70e-10 \*\*\*  
## I(log(total\_pitches)) 1.558e-02 2.458e-03 6.339 3.94e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08331 on 767 degrees of freedom  
## Multiple R-squared: 0.144, Adjusted R-squared: 0.1407   
## F-statistic: 43.02 on 3 and 767 DF, p-value: < 2.2e-16

termplot(mod\_velo\_spin\_swing\_and\_miss,partial.resid = T, se = T)

