hw7

jason grahn

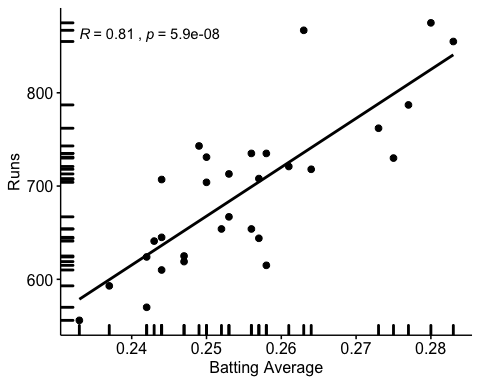
11/23/2018

# The R Part

## 1.

### Choose another traditional variable from mlb11 that you think might be a good predictor of runs. Produce a scatterplot of the two variables and fit a linear model. At a glance, does there seem to be a linear relationship?

#mlb11 %>%   
# select(runs, hits, homeruns, bat\_avg, strikeouts, stolen\_bases, wins) %>%  
# pairs(.)  
  
runs.bat\_avg.scatter <- ggscatter(mlb11,   
 x = "bat\_avg", y = "runs",   
 rug = TRUE,  
 #conf.int = TRUE,   
 add = "reg.line",   
 cor.coef = TRUE, cor.method = "pearson",  
 xlab = "Batting Average", ylab = "Runs")  
  
runs.bat\_avg.scatter



At first flance, we appear to have a strong positive relationship between hits and runs.

## 2.

### How does this relationship compare to the relationship between runs and at\_bats? Use the R2values from the two model summaries to compare. Does your variable seem to predict runs better than at\_bats? How can you tell?

at\_bats.corr <- mlb11 %>%   
 get\_correlation(runs ~ at\_bats) %>%   
 rename(at\_bats.corr = correlation)  
  
bat\_avg.corr <- mlb11 %>%   
 get\_correlation(runs ~ bat\_avg) %>%   
 rename(bat\_avg.corr = correlation)  
  
at\_bats.corr

## # A tibble: 1 x 1  
## at\_bats.corr  
## <dbl>  
## 1 0.611

bat\_avg.corr

## # A tibble: 1 x 1  
## bat\_avg.corr  
## <dbl>  
## 1 0.810

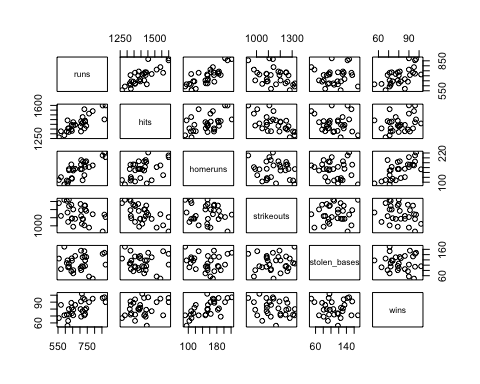
#m0 <- lm(runs ~ at\_bats, data = mlb11)  
#summary(m0) #0.3729  
  
m1 <- lm(runs ~ bat\_avg, data = mlb11)  
#summary(m1) #0.6561

Correlation between bat\_avg and runs is 0.8099859 which is better than the Correlation between at\_bats and runs which is 0.610627. This seems to be a better predictor as the relationship is nearly 20% stronger. The Multiple R-squared for at\_bats is *0.3729* where bat\_avg is *0.6561*; which is an impressive distance. Intuitively, it makes sense given a high batting average would be a key indicator for how often a run is made better than just going to bat.

## 3.

### Now that you can summarize the linear relationship between two variables, investigate the relationships between runs and each of the other five traditional variables. Which variable best predicts runs? Support your conclusion using the graphical and numerical methods we’ve discussed (for the sake of conciseness, only include output for the best variable, not all five).

#build a correlation plot for remaining variables.  
#this helps visualizize relationships  
pairs(~runs + hits + homeruns + strikeouts + stolen\_bases + wins, data = mlb11)



#find best correlation between variables  
m2 <- lm(runs ~ hits, data = mlb11)  
#m3 <- lm(runs ~ homeruns, data = mlb11)  
#m4 <- lm(runs ~ strikeouts, data = mlb11)  
#m5 <- lm(runs ~ stolen\_bases, data = mlb11)  
#m6 <- lm(runs ~ wins, data = mlb11)  
  
summary(m2) #0.6419

##   
## Call:  
## lm(formula = runs ~ hits, data = mlb11)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -103.718 -27.179 -5.233 19.322 140.693   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -375.5600 151.1806 -2.484 0.0192 \*   
## hits 0.7589 0.1071 7.085 1.04e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 50.23 on 28 degrees of freedom  
## Multiple R-squared: 0.6419, Adjusted R-squared: 0.6292   
## F-statistic: 50.2 on 1 and 28 DF, p-value: 1.043e-07

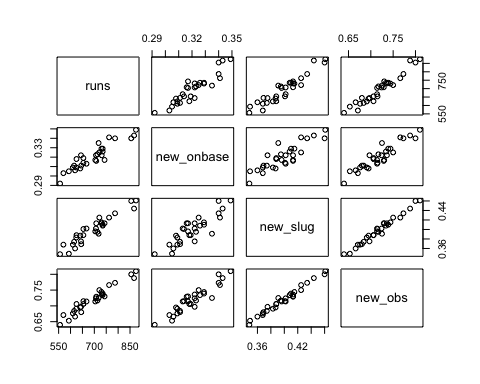
#summary(m3) #0.6266  
#summary(m4) #0.1694  
#summary(m5) #0.002914  
#summary(m6) #0.361

The highest correlation to runs from the remaining variables is with hits at approximately 0.8012.

## 4.

### Now examine the three newer variables. These are the statistics used by the author of Moneyball to predict a team’s success. In general, are they more or less effective at predicting runs that the old variables? Explain using appropriate graphical and numerical evidence.

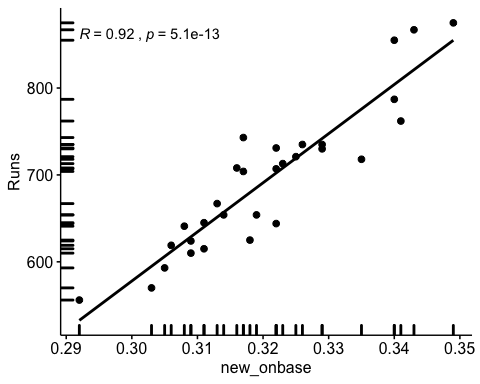
#build a correlation plot for remaining variables.  
#this helps visualizize relationships  
pairs(~runs + new\_onbase + new\_slug + new\_obs, data = mlb11)



#find best correlation between variables  
m7 <- lm(runs ~ new\_onbase, data = mlb11)  
summary(m7) #0.8491

##   
## Call:  
## lm(formula = runs ~ new\_onbase, data = mlb11)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.270 -18.335 3.249 19.520 69.002   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1118.4 144.5 -7.741 1.97e-08 \*\*\*  
## new\_onbase 5654.3 450.5 12.552 5.12e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.61 on 28 degrees of freedom  
## Multiple R-squared: 0.8491, Adjusted R-squared: 0.8437   
## F-statistic: 157.6 on 1 and 28 DF, p-value: 5.116e-13

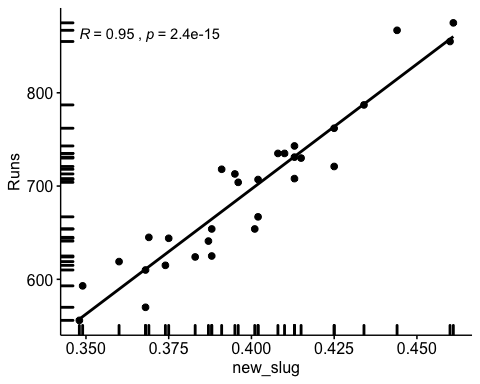
#mlb11 %>% get\_correlation(runs ~ new\_onbase) #0.9214691  
runs.new\_onbase.scatter <- ggscatter(mlb11,   
 x = "new\_onbase", y = "runs",   
 rug = TRUE,  
 #conf.int = TRUE,   
 add = "reg.line",   
 cor.coef = TRUE, cor.method = "pearson",  
 xlab = "new\_onbase", ylab = "Runs")  
runs.new\_onbase.scatter



m8 <- lm(runs ~ new\_slug, data = mlb11)  
summary(m8) #0.8969

##   
## Call:  
## lm(formula = runs ~ new\_slug, data = mlb11)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.41 -18.66 -0.91 16.29 52.29   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -375.80 68.71 -5.47 7.70e-06 \*\*\*  
## new\_slug 2681.33 171.83 15.61 2.42e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.96 on 28 degrees of freedom  
## Multiple R-squared: 0.8969, Adjusted R-squared: 0.8932   
## F-statistic: 243.5 on 1 and 28 DF, p-value: 2.42e-15

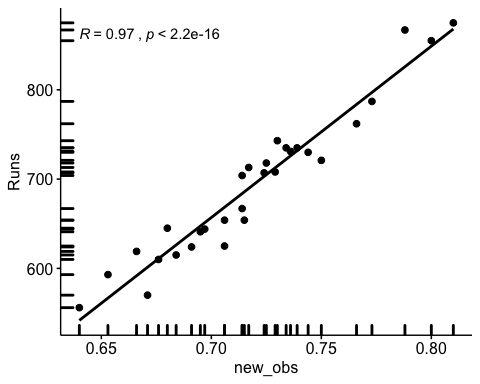
#mlb11 %>% get\_correlation(runs ~ new\_slug) #0.9470324  
runs.new\_slug.scatter <- ggscatter(mlb11,   
 x = "new\_slug", y = "runs",   
 rug = TRUE,  
 #conf.int = TRUE,   
 add = "reg.line",   
 cor.coef = TRUE, cor.method = "pearson",  
 xlab = "new\_slug", ylab = "Runs")  
runs.new\_slug.scatter



m9 <- lm(runs ~ new\_obs, data = mlb11)  
summary(m9) #0.9349

##   
## Call:  
## lm(formula = runs ~ new\_obs, data = mlb11)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -43.456 -13.690 1.165 13.935 41.156   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -686.61 68.93 -9.962 1.05e-10 \*\*\*  
## new\_obs 1919.36 95.70 20.057 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.41 on 28 degrees of freedom  
## Multiple R-squared: 0.9349, Adjusted R-squared: 0.9326   
## F-statistic: 402.3 on 1 and 28 DF, p-value: < 2.2e-16

#mlb11 %>% get\_correlation(runs ~ new\_obs) #0.9669163  
runs.new\_obs.scatter <- ggscatter(mlb11,   
 x = "new\_obs", y = "runs",   
 rug = TRUE,  
 #conf.int = TRUE,   
 add = "reg.line",   
 cor.coef = TRUE, cor.method = "pearson",  
 xlab = "new\_obs", ylab = "Runs")  
runs.new\_obs.scatter



These three new variables show much stronger positive relationships with runs. The Multiple R-Squared figures for these three new variables are *0.8491* for new\_onbase, *0.8969* for new\_slug, and *0.9349* for new\_obs - all quite higher than our traditional measurements.

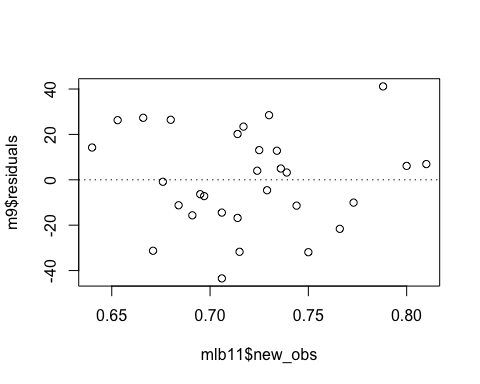
Of all ten variables we’ve analyzed, which seems to be the best predictor of runs? Using the limited (or not so limited) information you know about these baseball statistics, does your result make sense?

Of all 10 variables, new\_obs shows the most promise with a Multiple R-Squared so close to *1* and a correlation of approximately *0.97*

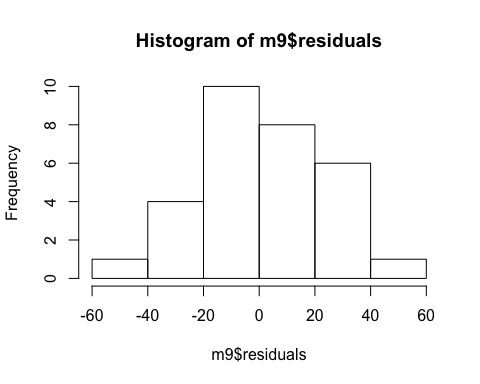
## 5.

### Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.

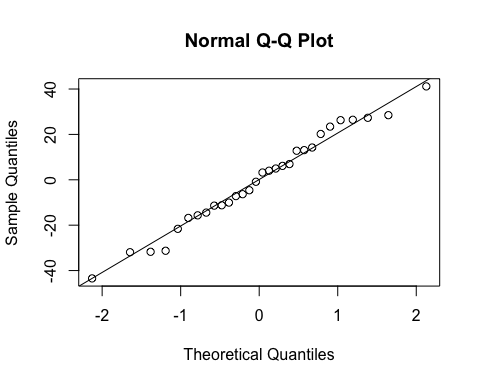
#residual distribution  
plot(m9$residuals ~ mlb11$new\_obs)  
abline(h = 0, lty = 3)

 Residual plot shows linear behaviour with no obvious patterns. The points appear to be randomly dispersed about the zero-axis. A linear regression model seems appropriate modeling.

#histogram of residuals  
hist(m9$residuals)

 The histogram shows nearly-normal distribution of the data, with maybe a slightly-right skew. This provides indication of normal residual conditions.

#normal probability plot of residuals  
qqnorm(m9$residuals)  
qqline(m9$residuals)



The quantile-quantile plot shows linearity of the points, without much distance from the linear line. I would agree that a general linear model would be sufficient for predicting runs based on new\_obs.