lab7

jason grahn

11/19/2018

# The R stuff

## 1

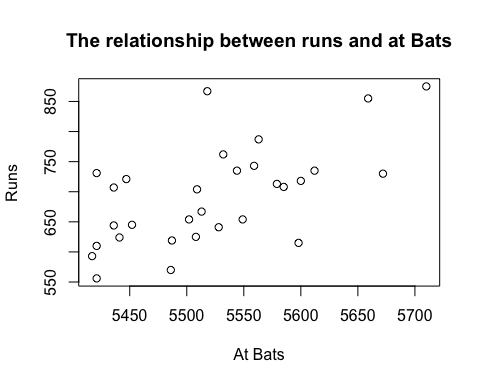
### What type of plot would you use to display the relationship between runs and one of the other numerical variables?

I’d use a scatterplot.

### Plot this relationship using the variable at\_bats as the predictor. Does the relationship look linear? If you knew a team’s at\_bats, would you be comfortable using a linear model to predict the number of runs?

The relationship does *not* appear linear. When we plot a linear line against the scatter, we can see over half the points exist outside the confidence interval of the linear plot. Using ggpubr we also see the Pearson correlation factor is approximatly 0.61.

plot(mlb11$runs ~ mlb11$at\_bats,   
 main = "The relationship between runs and at Bats",   
 xlab = "At Bats", ylab = "Runs")



#or using ggpubr from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r  
library("ggpubr")

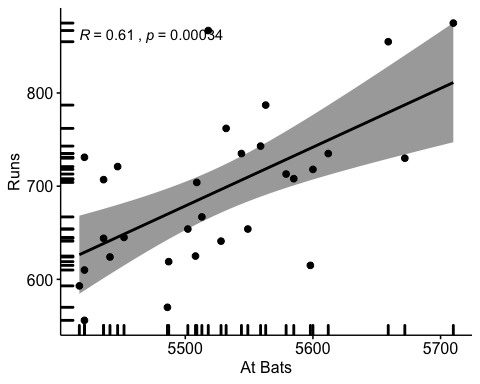
## Loading required package: magrittr

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

ggscatter(mlb11,   
 x = "at\_bats", y = "runs",   
 rug = TRUE,  
 conf.int = TRUE, add = "reg.line",   
 cor.coef = TRUE, cor.method = "pearson",  
 xlab = "At Bats", ylab = "Runs")



# 2

### Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

cor(mlb11$runs, mlb11$at\_bats)

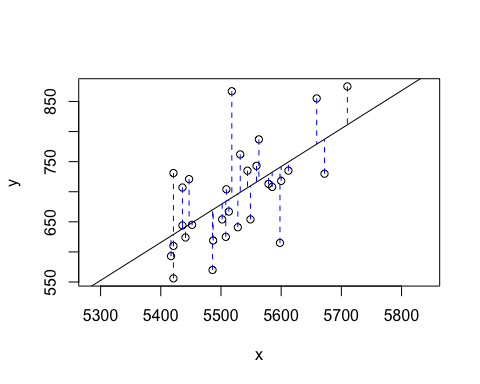
## [1] 0.610627

Our correlation as stated in the scatterplot is 0.61. This is positive but certainly not strong - noting that 0.7 is generally the accepted threshold for a “strong” relationship. There are outliers above and below that are easily identified outside of the confidence band on the plot.

# 3

### Using plot\_ss, choose a line that does a good job of minimizing the sum of squares. Run the function several times. What was the smallest sum of squares that you got? How does it compare to your neighbors?

plot\_ss(x = mlb11$at\_bats, y = mlb11$runs)



## Click two points to make a line.  
   
## Call:  
## lm(formula = y ~ x, data = pts)  
##   
## Coefficients:  
## (Intercept) x   
## -2789.2429 0.6305   
##   
## Sum of Squares: 123721.9