lab7

jason grahn

11/19/2018

# The R stuff

## 1

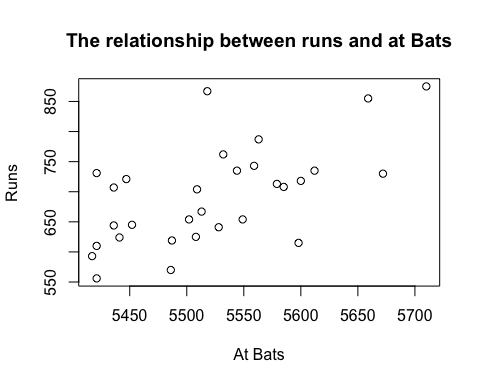
### What type of plot would you use to display the relationship between runs and one of the other numerical variables?

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")

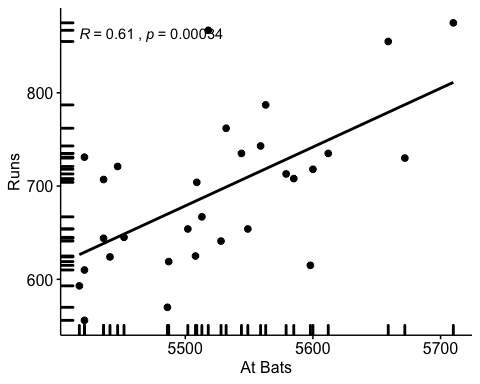
## Loading required package: magrittr

##   
## Attaching package: 'magrittr'

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

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

runs.bats.scatter <- 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")  
  
runs.bats.scatter



## 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

#or tidy style with moderndive::get\_correlation  
mlb11 %>%   
 get\_correlation(runs ~ at\_bats)

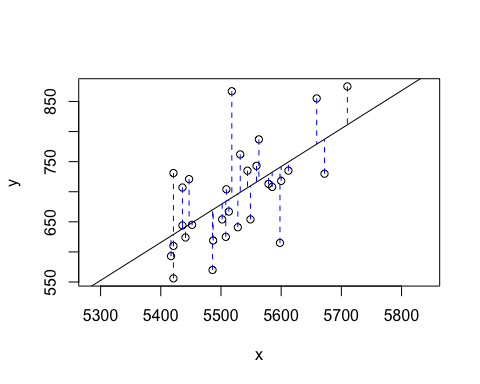
## # A tibble: 1 x 1  
## correlation  
## <dbl>  
## 1 0.611

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)



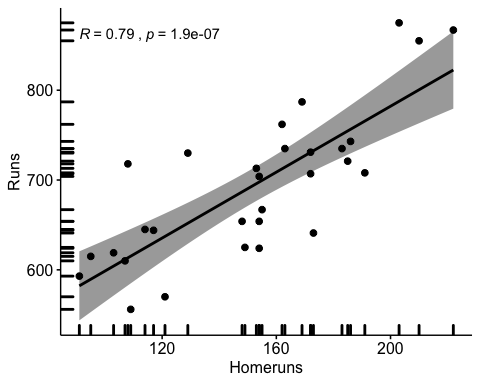
## Click two points to make a line.  
   
## Call:  
## lm(formula = y ~ x, data = pts)  
##   
## Coefficients:  
## (Intercept) x   
## -2789.2429 0.6305   
##   
## Sum of Squares: 123721.9

This seems to be the smallest sum of squares I can find. I’m sure there’s a more automated way to calculate this rather then hunt for this.

## 4

### Fit a new model that uses homeruns to predict runs. Using the estimates from the R output, write the equation of the regression line. What does the slope tell us in the context of the relationship between success of a team and its home runs?

ggscatter(mlb11,   
 x = "homeruns", y = "runs",   
 rug = TRUE,  
 conf.int = TRUE, add = "reg.line",   
 cor.coef = TRUE, cor.method = "pearson",  
 xlab = "Homeruns", ylab = "Runs")



m1 <- lm(runs ~ homeruns, data = mlb11)  
summary(m1)

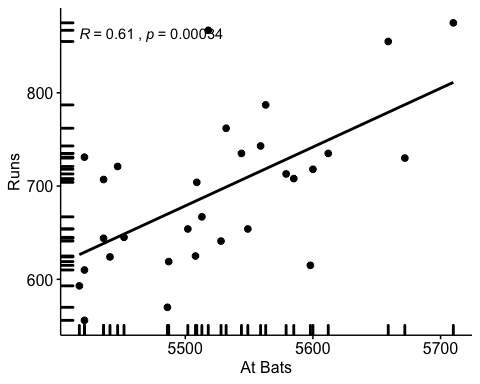
##   
## Call:  
## lm(formula = runs ~ homeruns, data = mlb11)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -91.615 -33.410 3.231 24.292 104.631   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 415.2389 41.6779 9.963 1.04e-10 \*\*\*  
## homeruns 1.8345 0.2677 6.854 1.90e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 51.29 on 28 degrees of freedom  
## Multiple R-squared: 0.6266, Adjusted R-squared: 0.6132   
## F-statistic: 46.98 on 1 and 28 DF, p-value: 1.9e-07

The formula is runs = 415.2389 + 1.8345 \* homeruns. There’s a positive relationship between homeruns and runs and the slope for homerunes of 1.8345 tells us that for every homerun, we can expect approximately 1.8 runs. The question explicitly asks about the “success of a team” - unfortunately this is extrapolation on the model as is. If we wanted to chart “success” we should first correlate *wins* to *runs*; or at least predict *wins* instead of *runs*.

## 5

### If a team manager saw the least squares regression line and not the actual data, how many runs would he or she predict for a team with 5,578 at-bats? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

#we already built this graph, so let's just call it again.  
runs.bats.scatter



They’d have approximately 700 runs if they were JUST looking at the chart. This is an underestimate. Using lm() to generate a model of 0.6305 \* at bats - 2789.2429, we’d calculate approximately 728 runs.

There is no 5578 at bats available for inference, but the Philadelphia Phillies had 713 runs with 5579. This gives us grounds for comparision to conclude the model overestimated the runs by approximately 15 runs (728 - 713 = 15).

runs.bats.model <- lm(runs ~ at\_bats, data = mlb11)   
summary(runs.bats.model)

##   
## Call:  
## lm(formula = runs ~ at\_bats, data = mlb11)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -125.58 -47.05 -16.59 54.40 176.87   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2789.2429 853.6957 -3.267 0.002871 \*\*   
## at\_bats 0.6305 0.1545 4.080 0.000339 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 66.47 on 28 degrees of freedom  
## Multiple R-squared: 0.3729, Adjusted R-squared: 0.3505   
## F-statistic: 16.65 on 1 and 28 DF, p-value: 0.0003388

0.6305\*5578 - 2789.2429

## [1] 727.6861

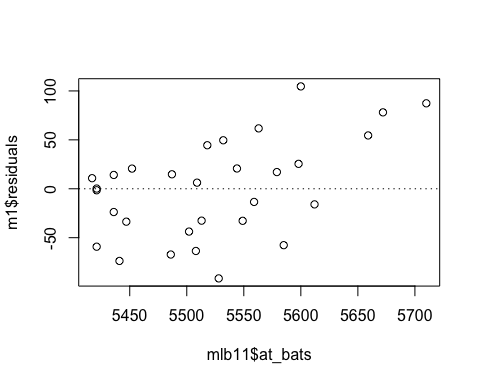
mlb11 %>%   
 filter(at\_bats > 5570 & at\_bats < 5600)

## team runs at\_bats hits homeruns bat\_avg strikeouts  
## 1 Houston Astros 615 5598 1442 95 0.258 1164  
## 2 Baltimore Orioles 708 5585 1434 191 0.257 1120  
## 3 Philadelphia Phillies 713 5579 1409 153 0.253 1024  
## stolen\_bases wins new\_onbase new\_slug new\_obs  
## 1 118 56 0.311 0.374 0.684  
## 2 81 69 0.316 0.413 0.729  
## 3 96 102 0.323 0.395 0.717

## 6

### Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

plot(m1$residuals ~ mlb11$at\_bats)  
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0

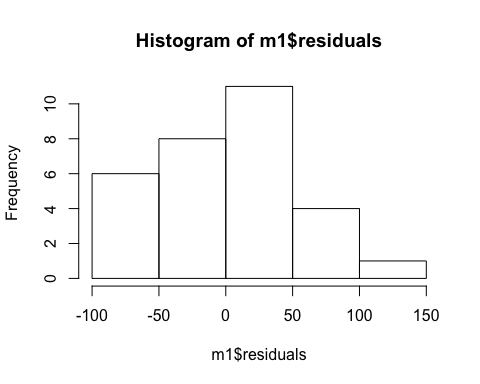


There’s no obvious pattern in the residual plot. They appear randomly dispersed around the zero axis. The relationship appears generally linear, though I’m not a huge fan of the right side and that it only has data points above the axis. A linear regression model would probably be appropriate for the data.

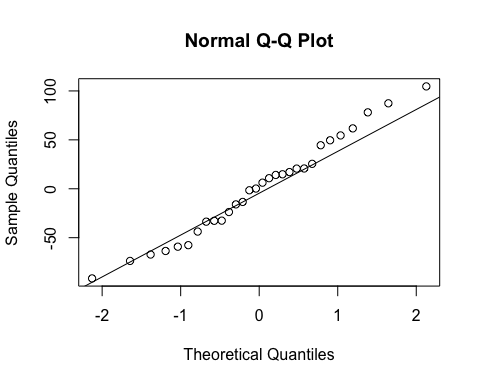
## 7

### Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

hist(m1$residuals)



qqnorm(m1$residuals)  
qqline(m1$residuals) # adds diagonal line to the normal prob plot



The histogram show a nearly-normal distribution, which indicates linearity and a LM model would be acceptable. The normal probability (qq) plot shows some stepwise motion and the right tail is higher than the left in relation to the diagonal, but overall it isn’t too terribly far from the line. I’d say normal residual conditions are met.

## 8

### Based on the plot in (1), does the constant variability condition appear to be met?

Points placed aound the least squares line look acceptably constant. I’d conclude the condition of constant variability has been met, yes.