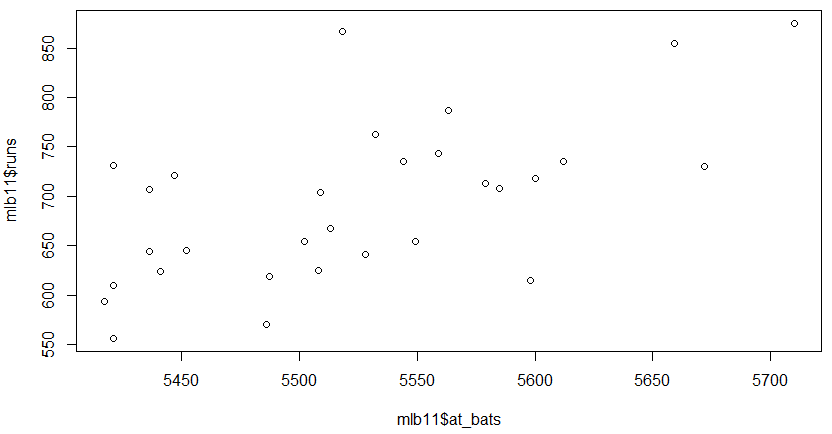
**Lab 8**

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

I would use a scatterplot to display these relationships.

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

The relationship does not have much curvature, and the graph looks moderately positive and linear. There appears to be at least one outlier. I would be hesitant to use a linear model until I check for the three criteria needed for linear regression.

**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.**

The two variables appear positive in direction with a seemingly linear form. When at bats increase, the number of runs increase overall. The strength is moderate, as there is at least one outlier and some variance.

**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, showSquares = TRUE)

Call:

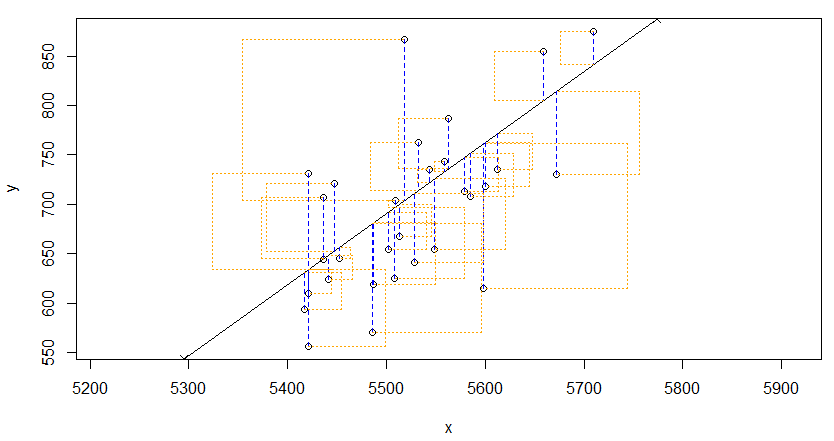
lm(formula = y ~ x, data = pts)

Coefficients:

(Intercept) x

-3262.8522 0.7188

Sum of Squares: 130983.2



**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?**

m2 <- lm(runs ~ homeruns, data = mlb11)

> summary(m2)

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

Y^ = 415.24 + (1.8345 \* homeruns)

Slope predicts that if homeruns increase by 1, then runs will increase by 1.8345.

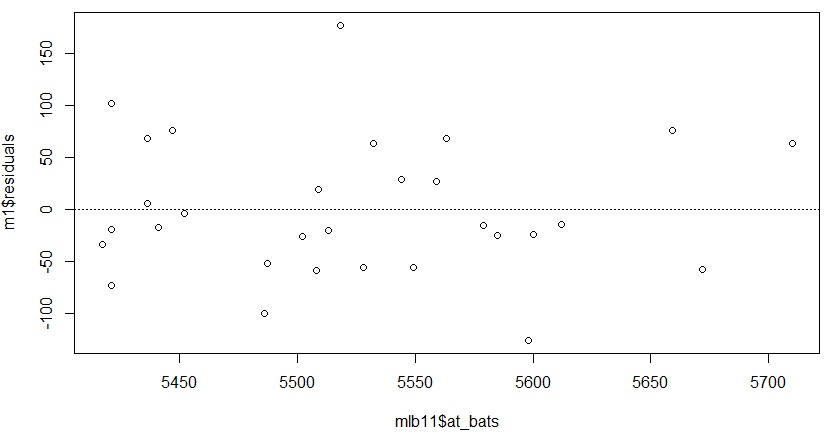
**5) If a team manager saw the least squares regression line and not the actual data, how many runs would he or she predict for a team with 5,578 at-bats? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?**

-2789.2429 + (0.6305 \* 5579) = 728.3166

This is an overestimate because the Phillies scored 713 runs, which is lower than the prediction.

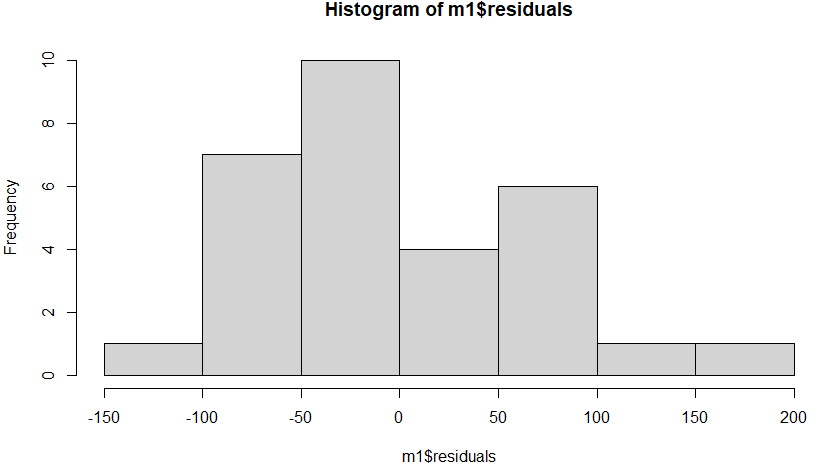
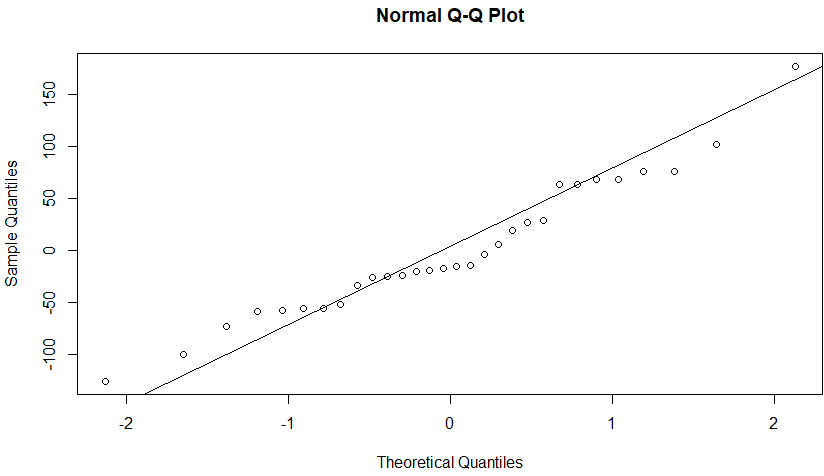
Residual = data – fit = 713 – 728.3166 = -15.3166 runs

**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?**



There does not appear to be much of a pattern, so we can say that there is linearity between runs and at-bats.

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

The nearly normal residuals condition appears to be met, because the sample size is small, and the data appear nearly normally distributed. The qq plot doesn’t have tails that curve to show a skew, and the qqnormsim function shows that simulated normal plots have a very similar shape to the qq plot for the real data.

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

The amount of overestimate and underestimate needs to be nearly-evenly spread out above and under the best fit line. The average amount of error for the residuals appears the same for both sides, so constant variability appears to be met.

**ON YOUR OWN**

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?**

>plot(mlb11$hits, mlb11$runs)

>m3 <- lm(runs ~ hits, data = mlb11)

> summary(m3)

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 \*\*\*

---

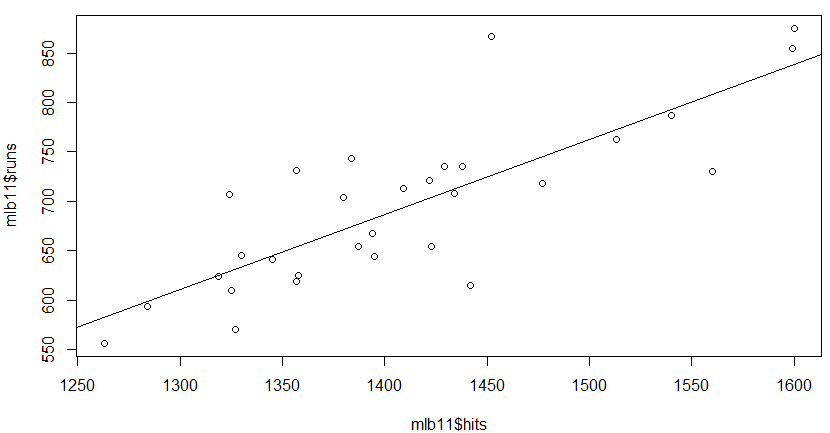
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

> plot(mlb11$hits, mlb11$runs)

> abline(m3)

Yes, there appears to be a stronger linear relationship between hits and runs, and the line of best fit shows this.

1. **How does this relationship compare to the relationship between runs and at\_bats? Use the R22 values from the two model summaries to compare. Does your variable seem to predict runs better than at\_bats? How can you tell?**

This relationship appears stronger than the relationship between runs and at bats. The R-squared value for hits to runs is 64.2%, while for at bats to runs is 37.3%. This means that the proportion of variability for runs is better explained by hits than by at bats.

In addition, the correlation coefficient for hits is stronger than for at bats, when comparing to runs.

cor(mlb11$runs, mlb11$hits) = 0.8012108

cor(mlb11$runs, mlb11$at\_bats) = 0.610627

1. **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).**

cor(mlb11$runs, mlb11$hits) = 0.8012108

cor(mlb11$runs, mlb11$at\_bats) = 0.610627

cor(mlb11$runs, mlb11$homeruns) = 0.7915577

cor(mlb11$runs, mlb11$bat\_avg) = 0.8099859

cor(mlb11$runs, mlb11$strikeouts) = -0.4115312

cor(mlb11$runs, mlb11$stolen\_bases) = 0.05398141

cor(mlb11$runs, mlb11$wins) = 0.6008088

Looking at the correlation coefficients, the two strongest variables appear to be hits and batting average to predict runs, with batting average appearing stronger. I checked it out graphically and numerically.

m4 <- lm(runs ~ bat\_avg, data = mlb11)

summary(m4)

Call:

lm(formula = runs ~ bat\_avg, data = mlb11)

Residuals:

Min 1Q Median 3Q Max

-94.676 -26.303 -5.496 28.482 131.113

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -642.8 183.1 -3.511 0.00153 \*\*

bat\_avg 5242.2 717.3 7.308 5.88e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

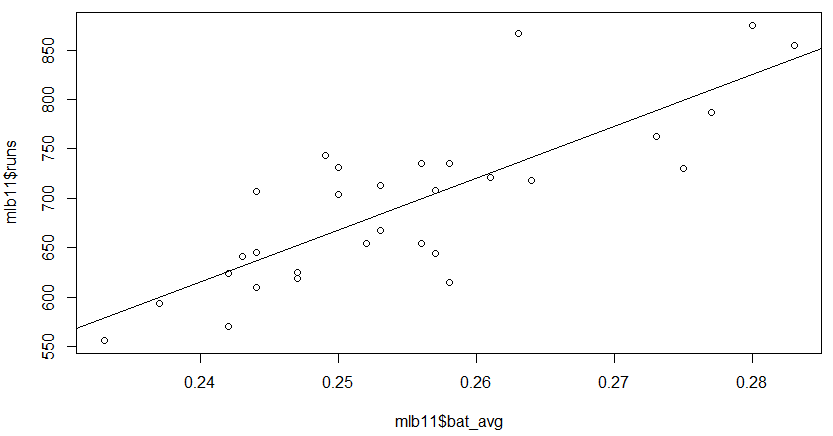
Residual standard error: 49.23 on 28 degrees of freedom

Multiple R-squared: 0.6561, Adjusted R-squared: 0.6438

F-statistic: 53.41 on 1 and 28 DF, p-value: 5.877e-08

plot(mlb11$bat\_avg, mlb11$runs)

abline(m4)

The scatterplots for hits and batting average appear very similar, which can be the case because their relationships to runs are very similar, but batting average slightly outclasses hits to predict runs. Batting average has a higher correlation coefficient for runs than hits (0.810 vs. 0.801) and has a higher R-squared value (65.6% vs. 64.2%). From this data, we can conclude that batting average predicts runs the best from the seven traditional variables.

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

cor(mlb11$runs, mlb11$new\_onbase) = 0.9214691

cor(mlb11$runs, mlb11$new\_slug) = 0.9470324

cor(mlb11$runs, mlb11$new\_obs) = 0.9669163

Looking at correlation coefficient alone, so far the three new variables have a much stronger linear correlation to runs than the seven traditional variables. I will next look at on-base plus slugging graphically and numerically to see if this holds up.

m5 <- lm(runs ~ new\_obs, data = mlb11)

summary(m5)

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

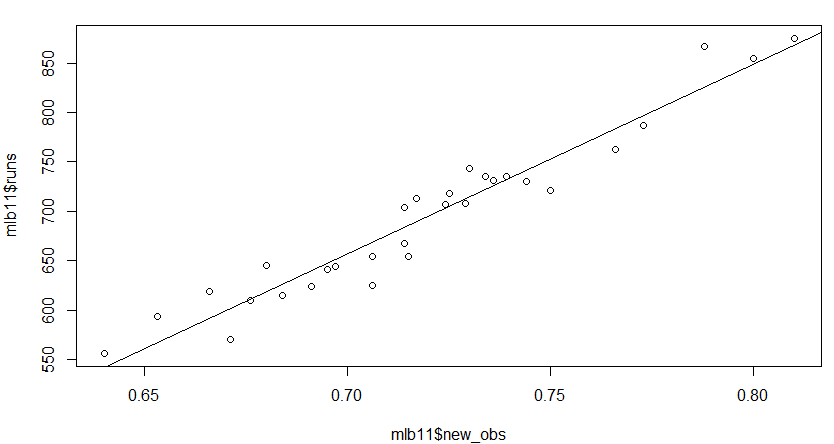
Residual standard error: 21.41 on 28 degrees of freedom

Multiple R-squared: 0.9349, Adjusted R-squared: 0.9326

F-statistic: 402.3 on 1 and 28 DF, p-value: < 2.2e-16

plot(mlb11$new\_obs, mlb11$runs)

abline(m5)

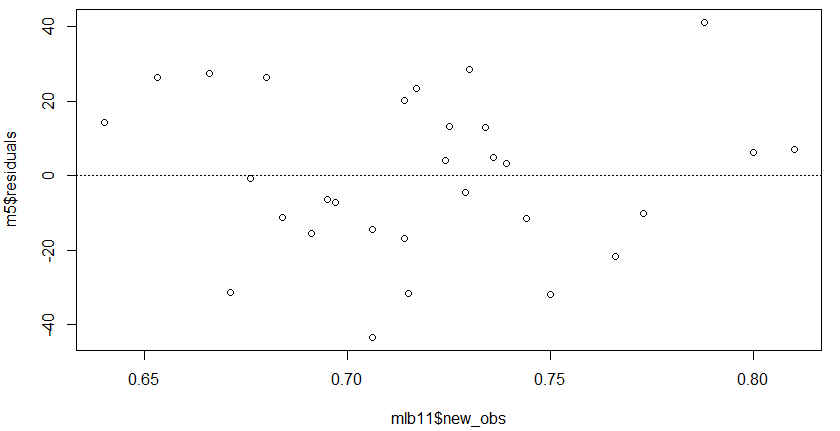


The scatterplot for OPS vs. runs is the tightest spread out of all the variables, as there is very little variance and no outliers. The on-base plus slugging variable appears to be the strongest predictor of runs, since it has the highest correlation coefficient of all variables, and its R-squared value is the highest at 93.5%. This means that 93.5% of the proportion of variability for runs can be attributed to on-base plus slugging. This makes sense because a player’s ability to get on base and slug for power can be very helpful to get runs.

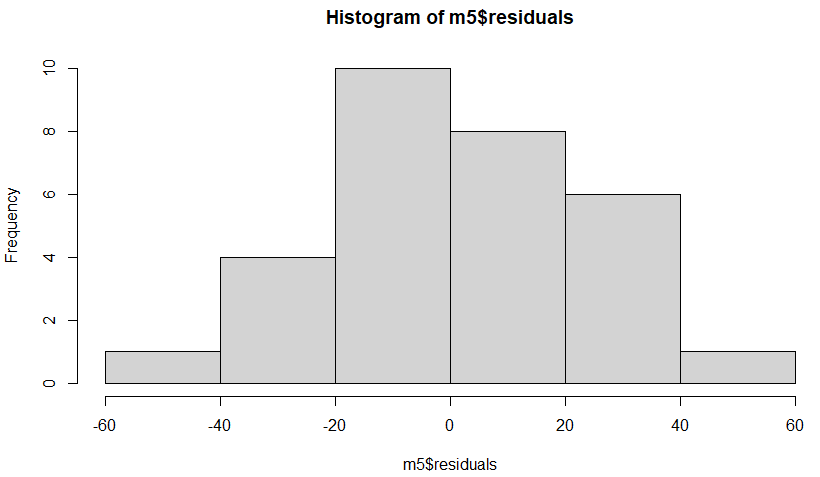
1. **Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.**

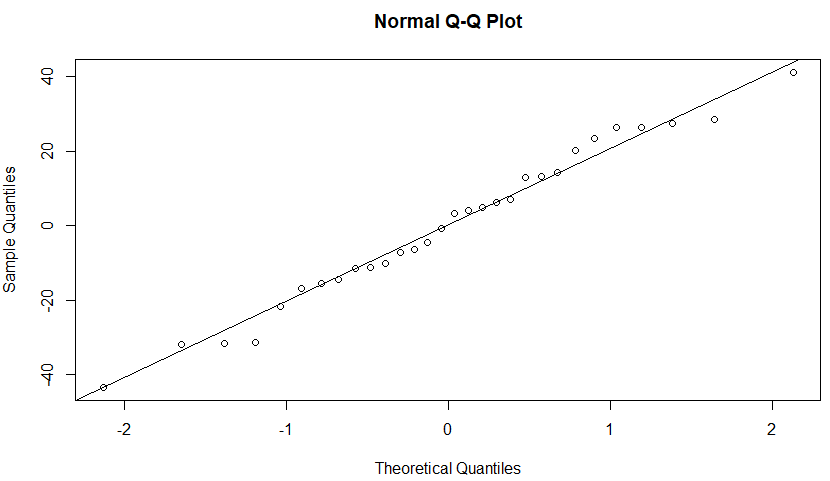
plot(m5$residuals ~ mlb11$new\_obs)

abline(h = 0, lty = 3)



hist(m5$residuals)

qqnorm(m5$residuals)

qqline(m5$residuals)

Looking at the model diagnostics, the variable OPS follows all the criteria for regression. It has linearity because the scatterplot is linear with no outliers and a strong correlation, as well as random patterns for the residual plot which also support linearity. The histogram and qq-plot both show the residuals are very normal, which check off nearly normal residuals. Looking at the residual plot for constant variability, the residuals on both sides appear to have very similar standard errors overall and don’t follow a pattern, so we can say there is constant variability. On both sides, no residuals stray further than 40 units away from the y= 0 line.