**Tor Thogersen**

**DA 460 – Fall 2017**

**Lab 7 - Handout 7 R and Handout 7 SAS**

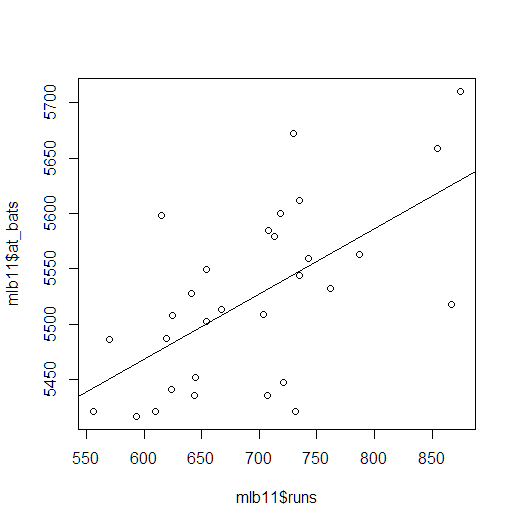
**Part 7 – R Handout - Introduction to linear regression**

**Exercise: 1**

1. What type of plot would you use to display the relationship between runs and one of the other numerical variables?
   1. I would use a scatter plot with linear regression line to show the relationship
2. Plot this relationship using the variable at\_bats as the predictor.

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

> abline(lm(mlb11$at\_bats~mlb11$runs))



1. 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?
   1. Yes, the scatterplot with linear regression could give you that information due to spread of then plots shows a linear progression but the spread of the plots are also very scattered apart which may cause issue with reliability when you use cor( ) function in tandem.

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

[1] 0.610627

**Exercise: 2**

1. 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.
   1. Based on the plot and linear regression line between at\_bats and runs, the trend for at\_bats has an upward progression showing linear relationship between at\_bats and Runs, based on the correlation fuction the relationship does not seem to very strong when you look at that data and all the outliers.

**Exercise: 3**

1. 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?
   1. Sum of Squares: 132793.5

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

Call:

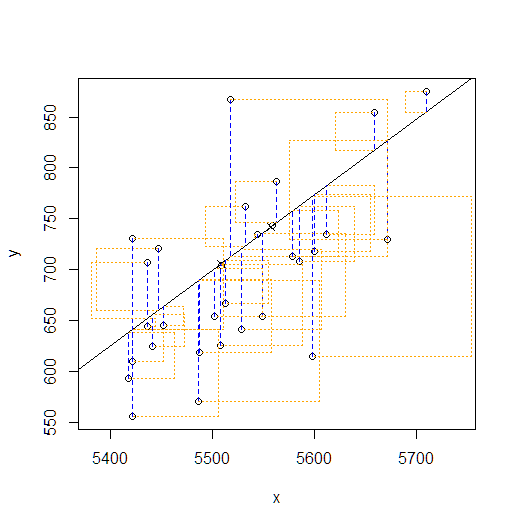
lm(formula = y ~ x, data = pts)

Coefficients:

(Intercept) x

-4643.5137 0.9664

Sum of Squares: 144627.7



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

Call:

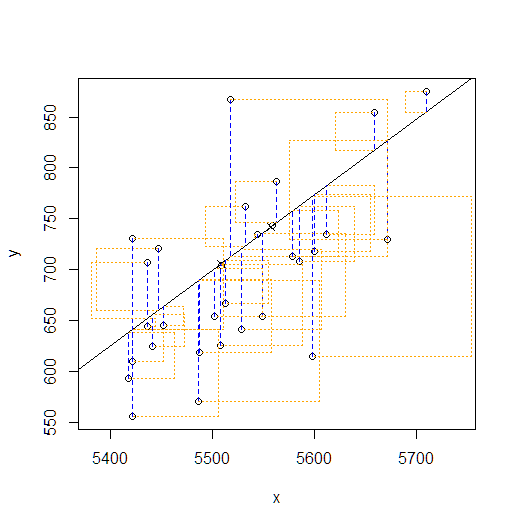
lm(formula = y ~ x, data = pts)

Coefficients:

(Intercept) x

-3370.53 0.74

Sum of Squares: 142173.4>



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

Call:

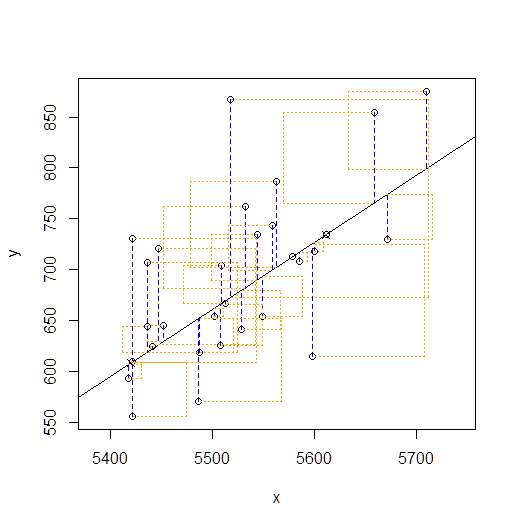
lm(formula = y ~ x, data = pts)

Coefficients:

(Intercept) x

-2949.3524 0.6564

Sum of Squares: 132793.5>



**Exercise: 4**

1. Fit a new model that uses homeruns to predict runs. Using the estimates from the R output, write the equation of the regression line.
   1. y = 415.2389+1.8345\*homerun

1. What does the slope tell us in the context of the relationship between success of a team and its home runs?
   1. Tells you the increase in runs based on each additional homerun5

> 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

**Exercise: 5**

1. 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?
   1. Based on the lsrl they would predict 727 runs for the team

> plot(mlb11$runs ~ mlb11$at\_bats)

> abline(m1)

> m1 <- lm(runs~at\_bats, data = mlb11)

> summary(m1)

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

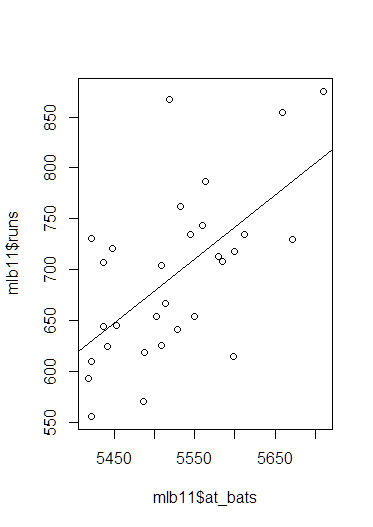
Residual standard error: 66.47 on 28 degrees of freedom

Multiple R-squared: 0.3729, Adjusted R-squared: 0.3505

F-statistic: 16.65 on 1 and 28 DF, p-value: 0.0003388

> -2789.2429+0.6305\*5578

[1] 727.6861



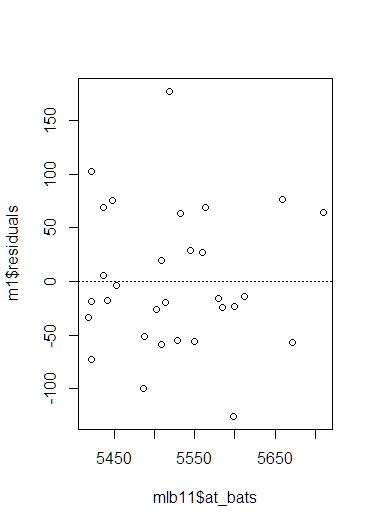
1. Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?
   1. Unable to give that answer based on the data because we do not have 5578 at bats in the chart, we are able to give an accurate estimation based on the data.

**Exercise: 6**

1. Is there any apparent pattern in the residuals plot?
   1. The spread patter seems equal on both sides of line and has no distinct pattern so the data should be linear

> plot(m1$residuals ~ mlb11$at\_bats)

> abline(h = 0, lty = 3)



1. What does this indicate about the linearity of the relationship between runs and at-bats?
   1. Shows the linearity between at\_bats and runs, that if you increase or decrease your at\_bats the run production will also increase or decrease.

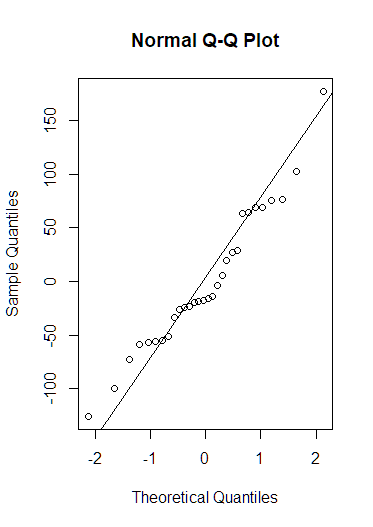
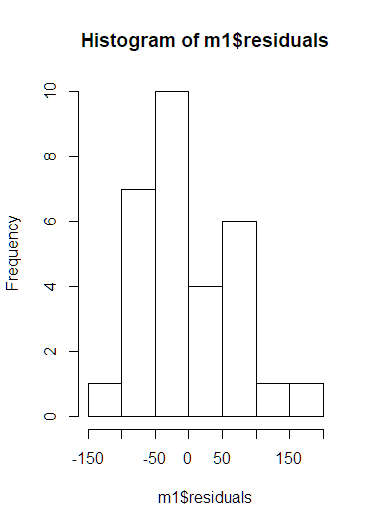
**Exercise: 7**

1. Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?
   1. Yes, histogram shows a approximate normal distribution and data follows the linear line fairly well.

> hist(m1$residuals)

> qqnorm(m1$residuals)

> qqline(m1$residuals)



**Exercise: 8**

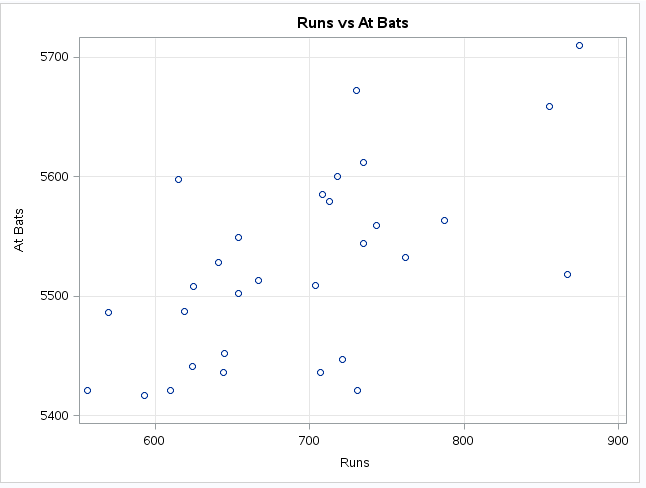
1. Based on the plot in (1), does the constant variability condition appear to be met?
   1. Yes, but we do have a few outliers

**Part 7 – SAS Handout - Introduction to linear regression**

**Exercise 1:**

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.
   1. I would use scatter plot to view the relationship between runs and at\_bats
2. 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?
   1. Yes based on the scatter plot there is an upward trend to suggestion linear regression, we have many outlier plots that could cause issue with accuracy.

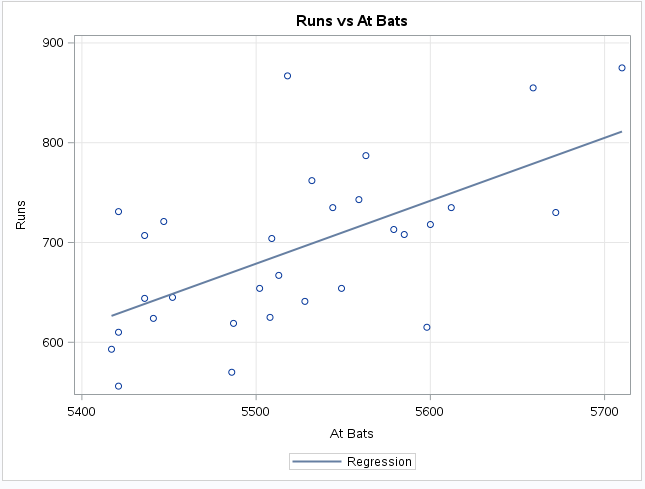
ods graphics / reset imagemap;  
  
proc sgplot data=WORK.MLB11;  
 title "Runs vs At Bats";  
 scatter x=runs y=at\_bats / transparency=0.0 name='Scatter';  
 xaxis grid label="Runs";  
 yaxis grid label="At Bats";  
run;  
  
ods graphics / reset;  
title;



**Exercise 2:**

1. Looking at your plot from the previous exercise, describe the relationship between these two variables. Be sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.
   1. Based on the plot and linear regression line between at\_bats and runs, the trend for at\_bats has an upward progression showing linear relationship between at\_bats and Runs, based on the correlation fuction the relationship does not seem to very strong when you look at that data and all the outliers.

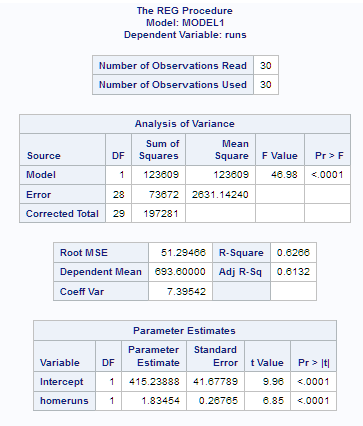
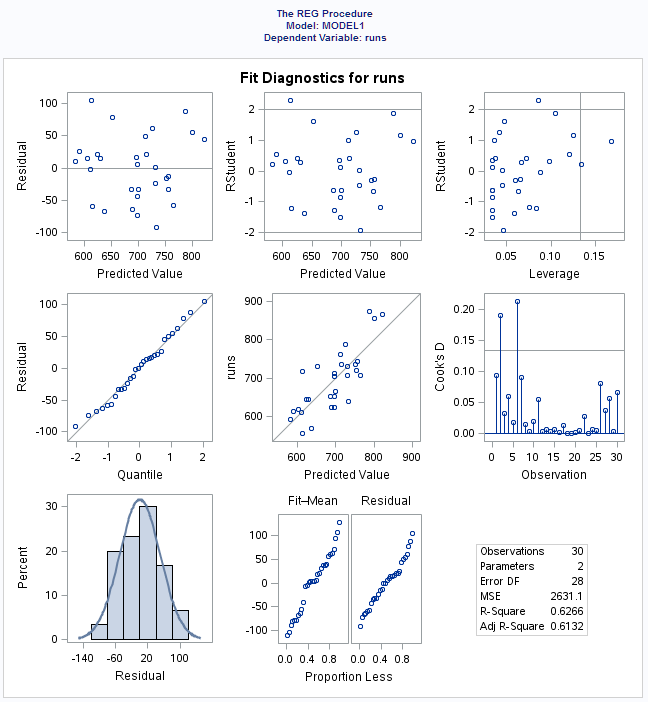
proc sgplot data=WORK.MLB11;  
 title "Runs vs At Bats";  
 reg x=at\_bats y=runs / transparency=0.0 name='regression';  
 xaxis grid label="At Bats";  
 yaxis grid label="Runs";  
run;

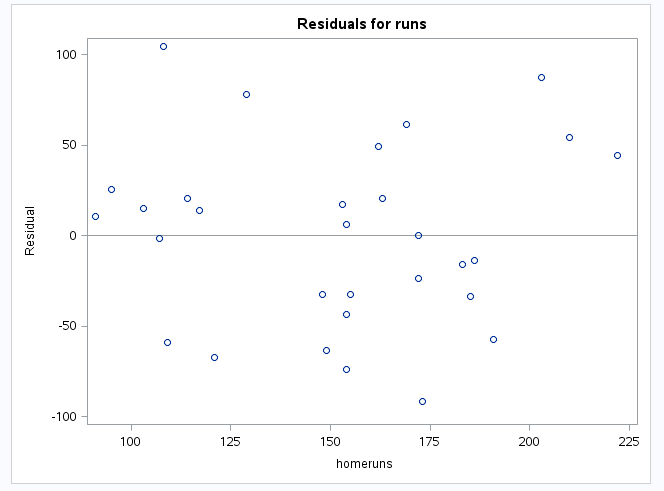
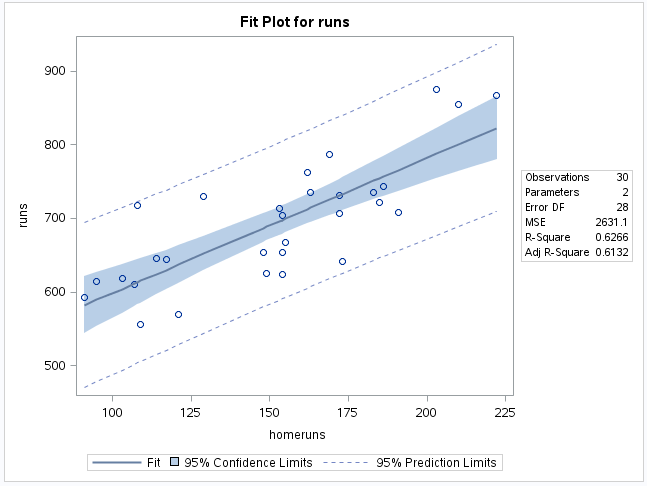


**Exercise 3:**

1. Fit a new model that uses homeruns to predict runs. Using the estimates from the SAS 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?
   1. Y-hat = 415.2389+1.8345\*homerun
   2. That runs will increase or decrease by 1.8356 based on how many at\_bats

proc reg data=mlb11;  
 model runs=homeruns;  
run;   
quit;

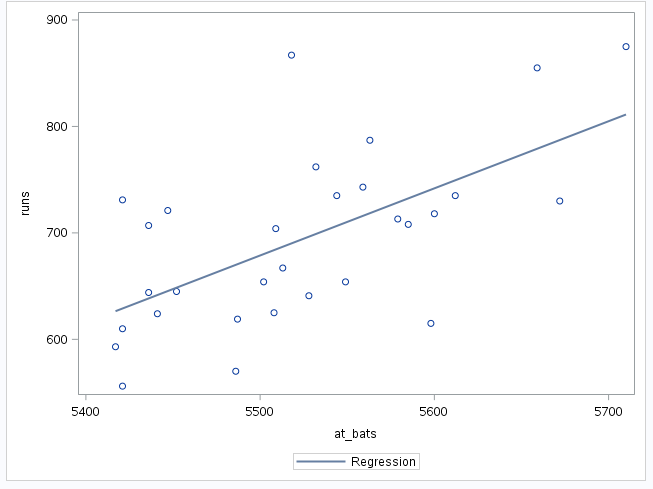
 

**Exercise 4:**

1. 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?
   1. they would predict 727 runs for the team
2. Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?
   1. Unable to give that answer based on the data because we do not have 5578 at bats in the chart, we are able to give an accurate estimation based on the data

proc sgplot data=mlb11;  
 reg x=at\_bats y=runs;  
run;



**Exercise 5:**

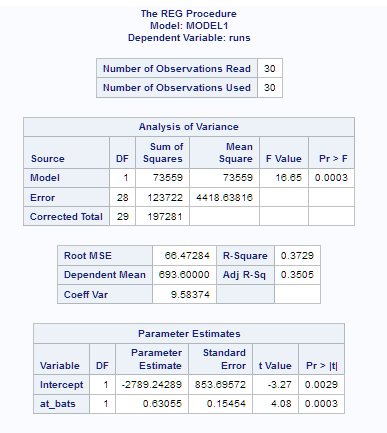
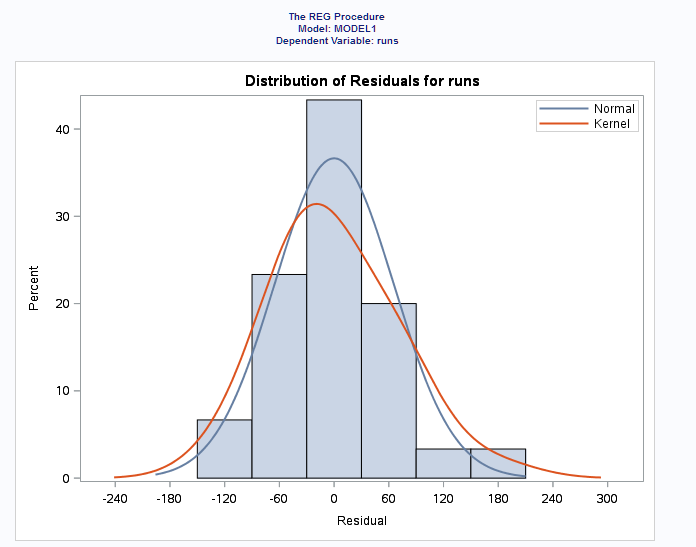
1. Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at\_bats?
   1. The spread patter seems equal on both sides of line and has no distinct pattern so the data should be linear

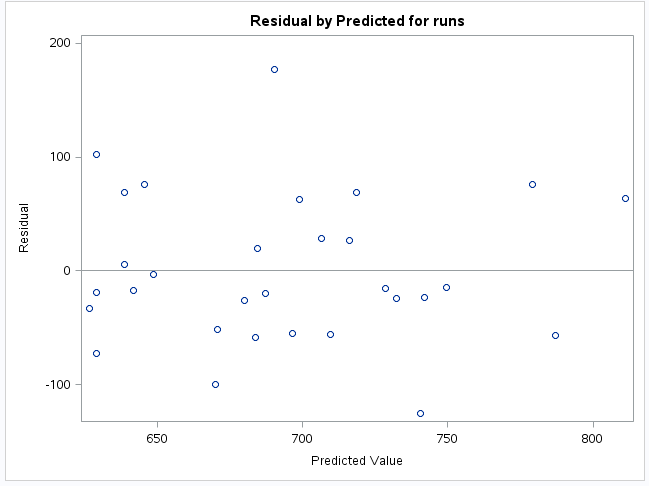
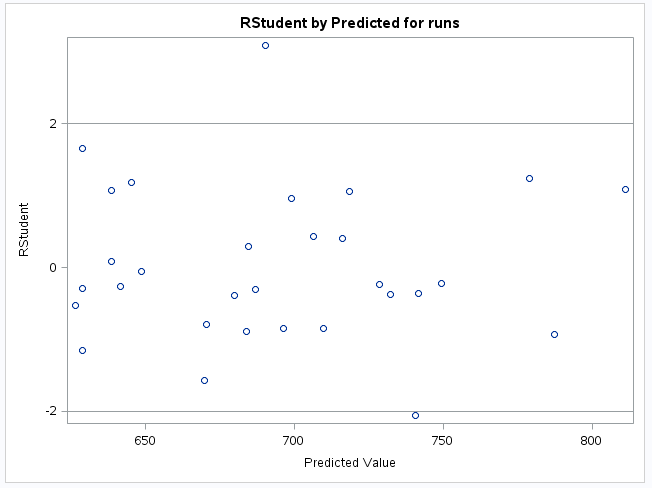
proc reg data=mlb11 plots=diagnostics(unpack);

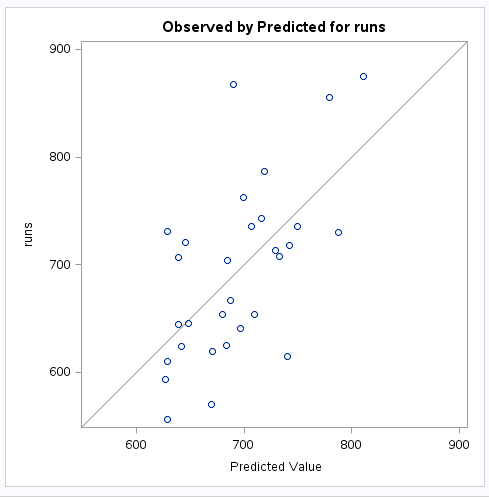
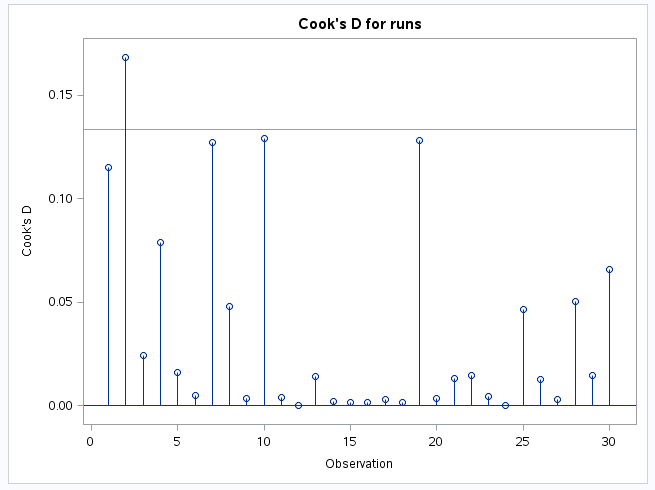
model runs=at\_bats;

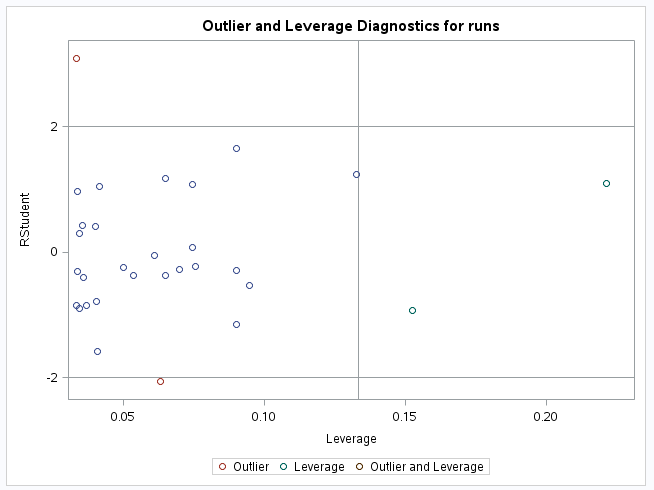
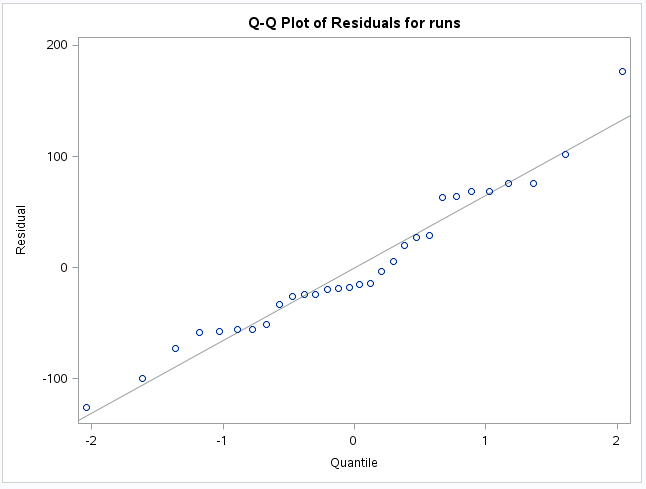
run;

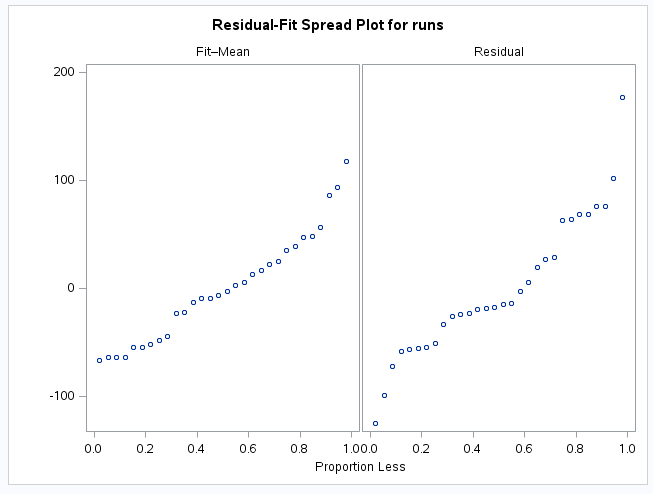
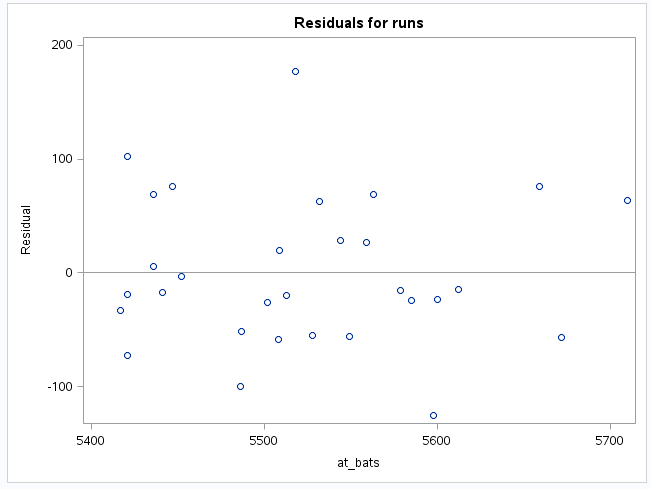
quit;

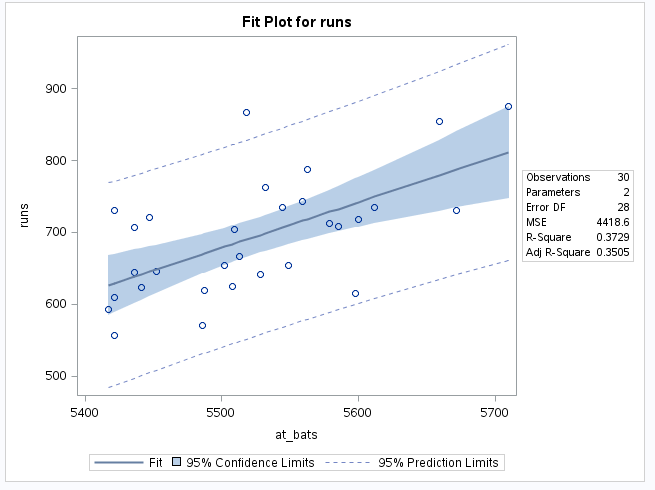
 

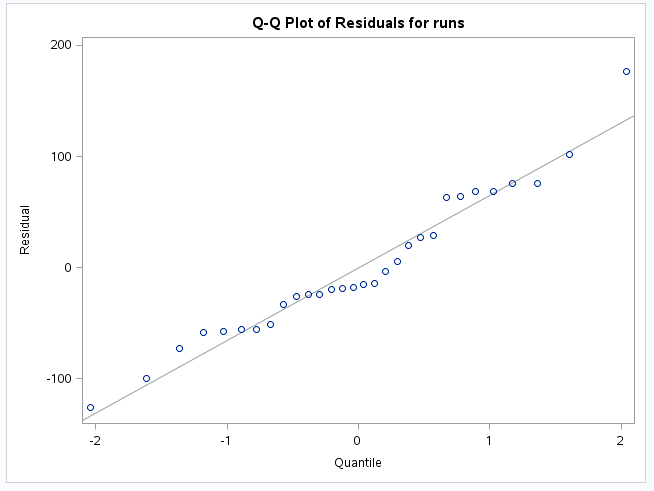
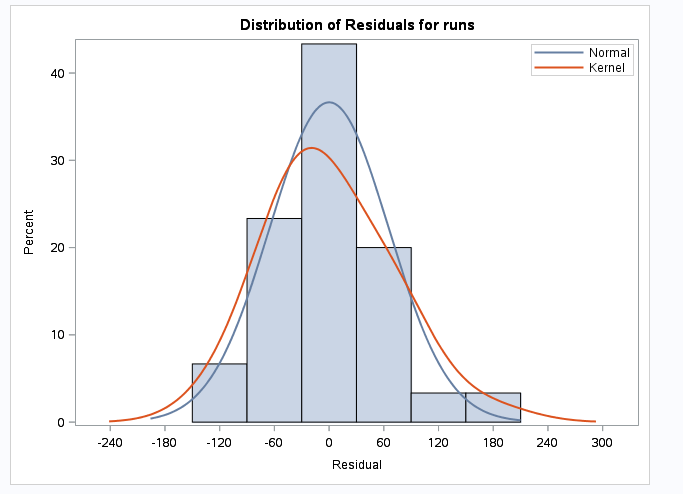
 



**Exercise 6:**

1. Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?
   1. Yes, histogram shows a approximate normal distribution and data follows the linear line fairly well.



**Exercise 7:**

1. Based on the plot in (1), does the constant variability condition appear to be met?
   1. Yes, but we do have a few outliers

