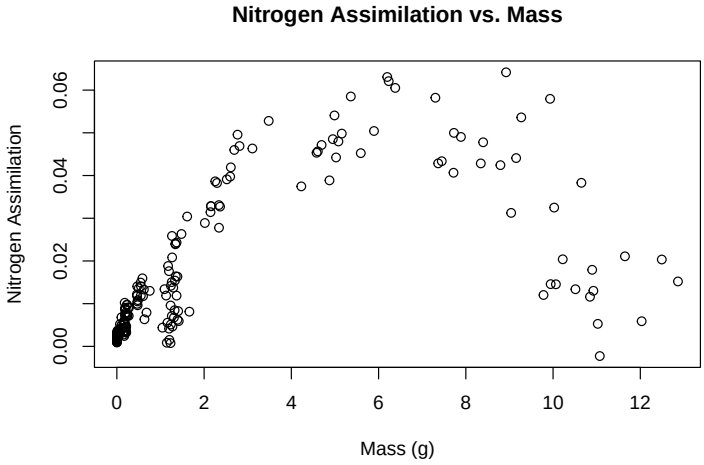
Day4cont

Olivia Wu

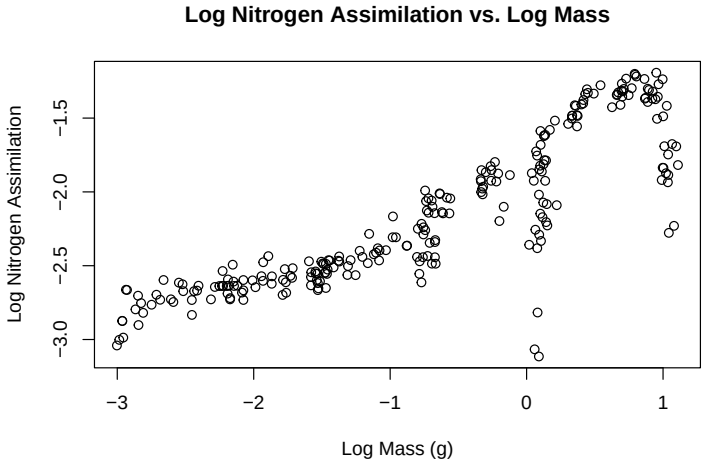
2024-02-20

Problem 1.30

a) The scatterplot shows a strong curve.



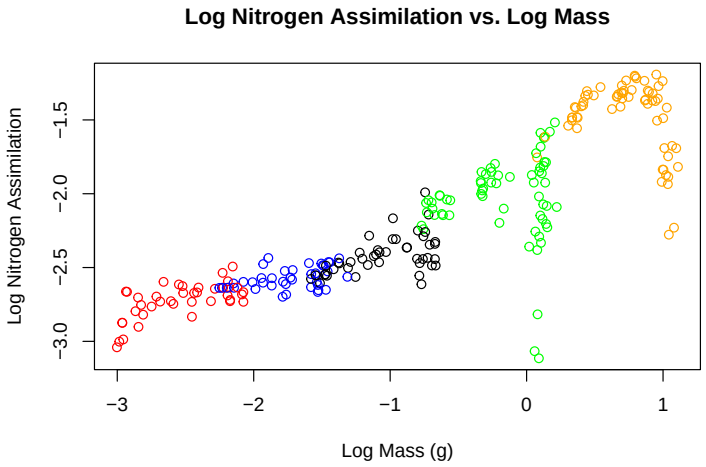
b) The scatterplot is more linear, but there are outliers as the log mass gets closer to 1.



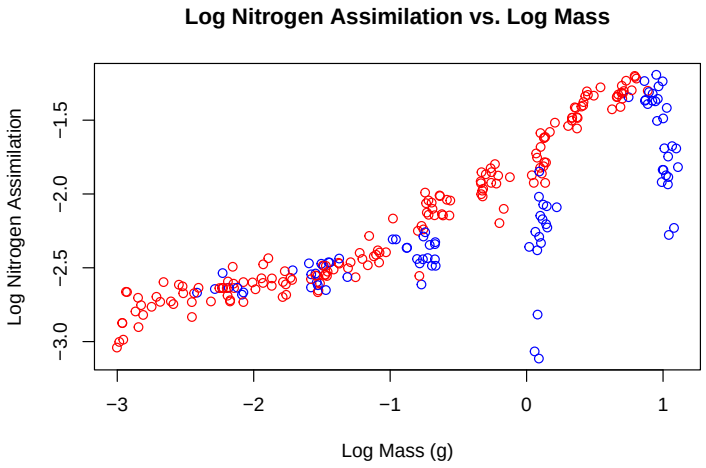
c) I would rather use the plot in part (b) because it is more linear. The equation for this is $\widehat{logNassim} = -1.887 + 0.371 \cdot logMass$

```
##
## Call:
## lm(formula = logNassim ~ logMass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26089 -0.11558  0.02162  0.16725  0.41862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.88738    0.01841 -102.53   <2e-16 ***
## logMass      0.37096    0.01332   27.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2501 on 251 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.7555, Adjusted R-squared:  0.7545
## F-statistic: 775.6 on 1 and 251 DF,  p-value: < 2.2e-16
```

d) Map: 1 (red), 2 (blue), 3 (black), 4 (green), 5 (orange). Each group was mostly linear, but the stage 4 and 5 groups showed more curve.

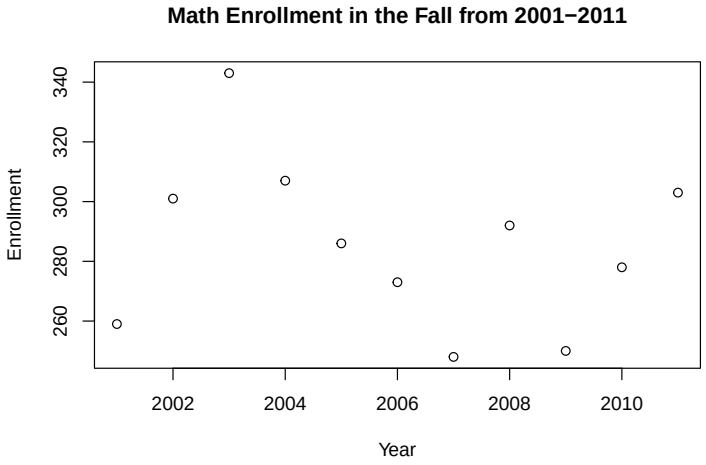


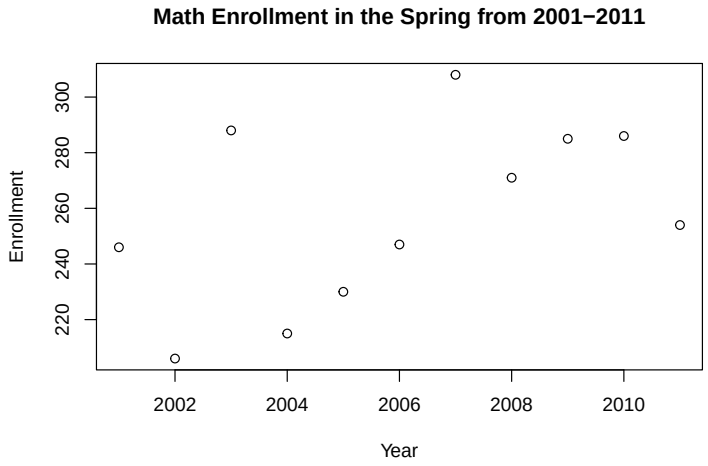
e) Map: Fgp (red), not Fgp (blue). The free-growth period plants had a more linear trend.



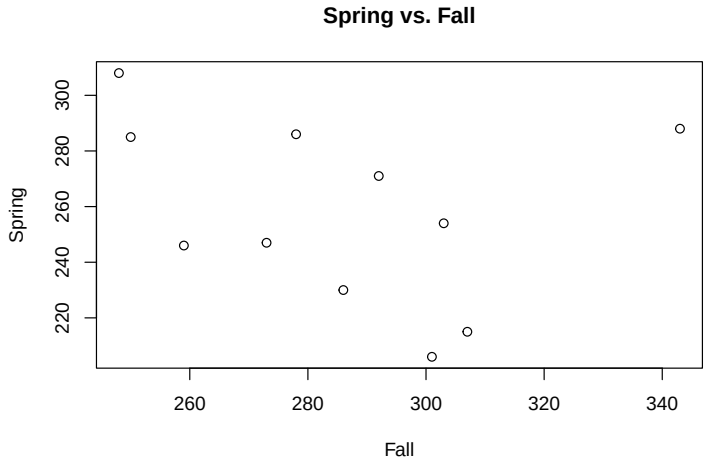
Problem 1.34

a) The trend over time is different for each semester. In the fall, it is a weak negative association. In the spring, it is a weak positive association.





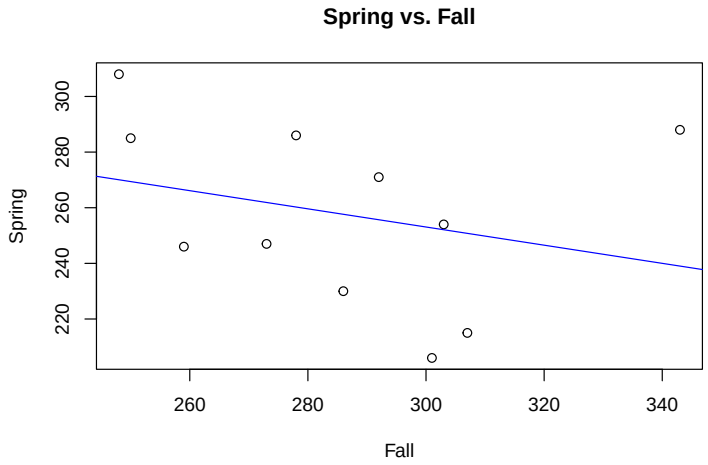
b) I disagree with the statement because the scatterplot of fall vs. spring has a very weak negative linear association.



c) They could be talking about the 2003 year where the Fall enrollment was 343 and the Spring enrollment was 288.

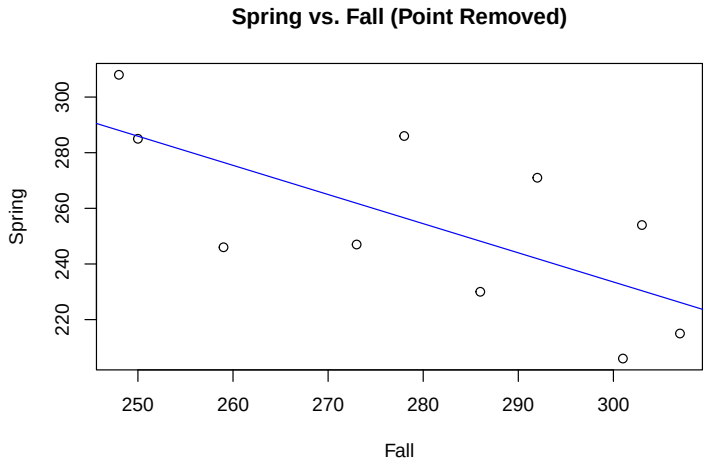
```
##
## Call:
## lm(formula = Spring ~ Fall, data = MthEnr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.740 -24.050   1.913  20.674  48.978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 351.0585   106.4710   3.297  0.00927 **
## Fall         -0.3266     0.3713  -0.880  0.40195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 33.09 on 9 degrees of freedom
## Multiple R-squared:  0.07916,    Adjusted R-squared:  -0.02315
## F-statistic: 0.7737 on 1 and 9 DF,  p-value: 0.4019
```



d) The slope of the fitted line without the point (-1.048) is less than the slope of the fitted line with the point (-0.327). Since the slope changed a lot, this point is influential.

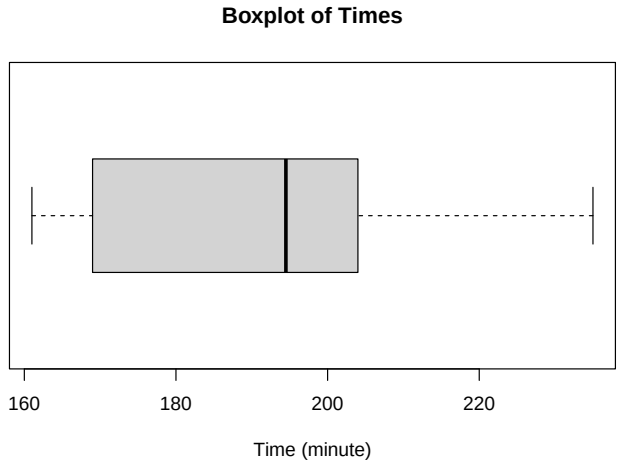
```
##
## Call:
## lm(formula = Spring[Fall != 343] ~ Fall[Fall != 343], data = MthEnr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.500 -17.353  -6.058  22.711  29.418
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       548.0094   106.7236   5.135 0.000891 ***
## Fall[Fall != 343]  -1.0483     0.3805  -2.755 0.024870 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.94 on 8 degrees of freedom
## Multiple R-squared:  0.4868, Adjusted R-squared:  0.4227
## F-statistic: 7.589 on 1 and 8 DF,  p-value: 0.02487
```



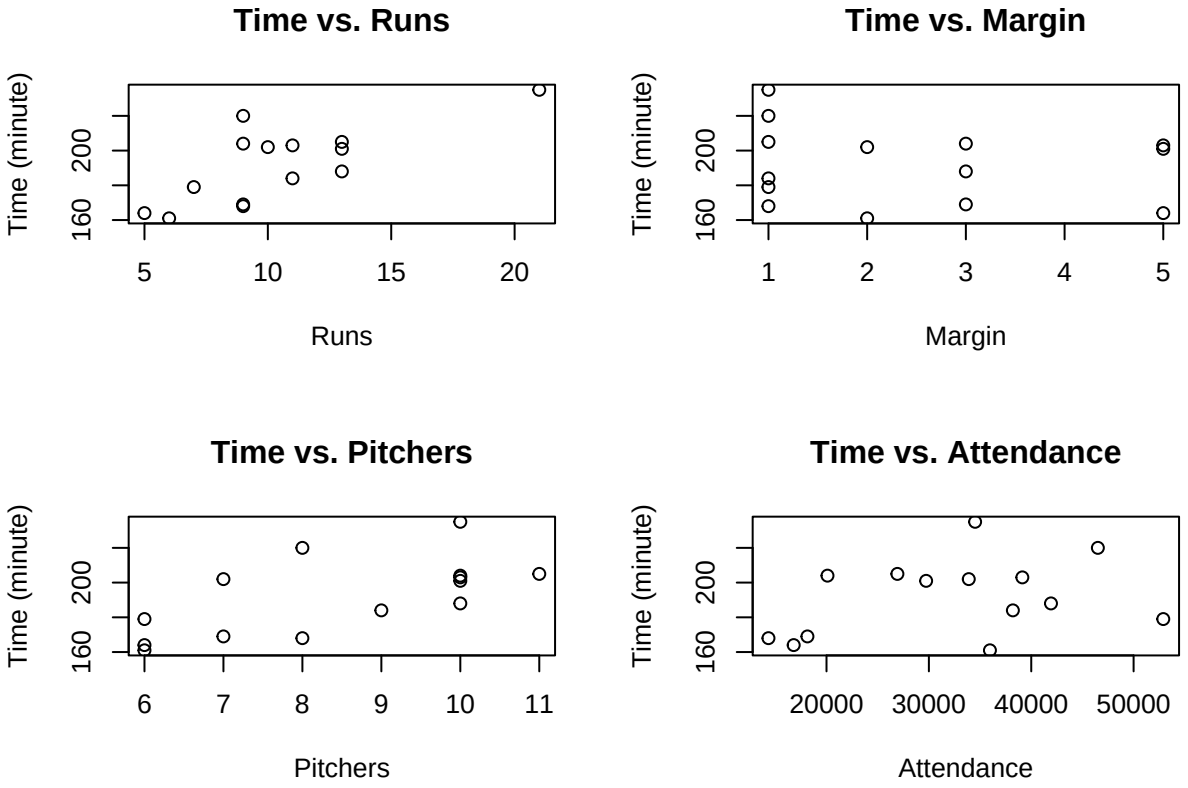
Problem 1.45

a) The times do not have a symmetric distribution. They are centered around 194.5 minutes, and there is a moderate spread.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   161.0   171.5   194.5   191.6   203.8   235.0
```

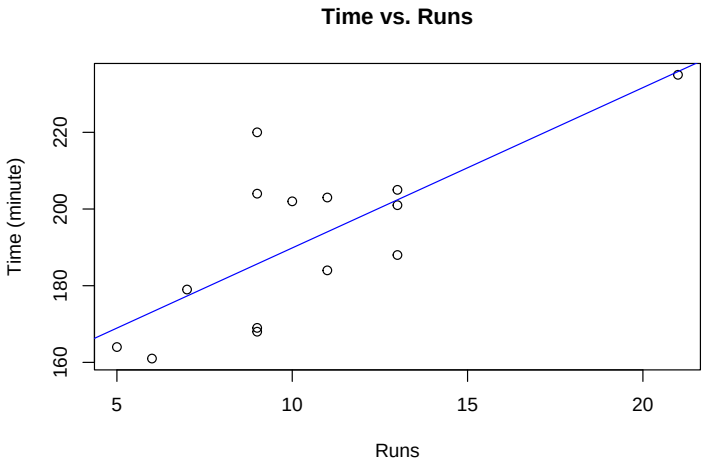


b) The plot with Runs is most linear.

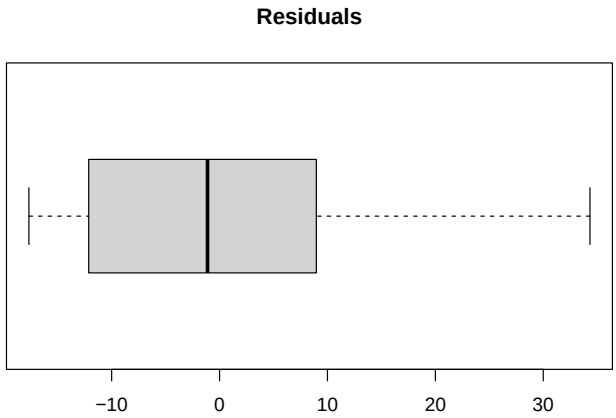


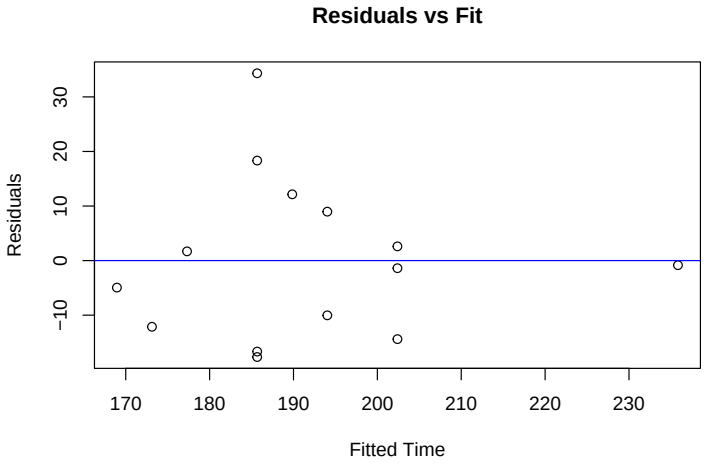
c) The best predictor variable is the number of runs in a game. The regression equation is $\widehat{Time} = 148.043 + 4.181 \cdot Runs$. On average, an increase of 1 run tends to result in an increase of 4.181 minutes of Time of the game.

```
##
## Call:
## lm(formula = Time ~ Runs, data = bball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.670 -11.604  -1.117   7.378  34.330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148.043     11.995  12.342 3.53e-08 ***
## Runs           4.181      1.081   3.868  0.00224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.34 on 12 degrees of freedom
## Multiple R-squared:  0.5549, Adjusted R-squared:  0.5178
## F-statistic: 14.96 on 1 and 12 DF,  p-value: 0.002237
```



d) The residual plot shows an even scatter, but the boxplot implies a slight right skew.

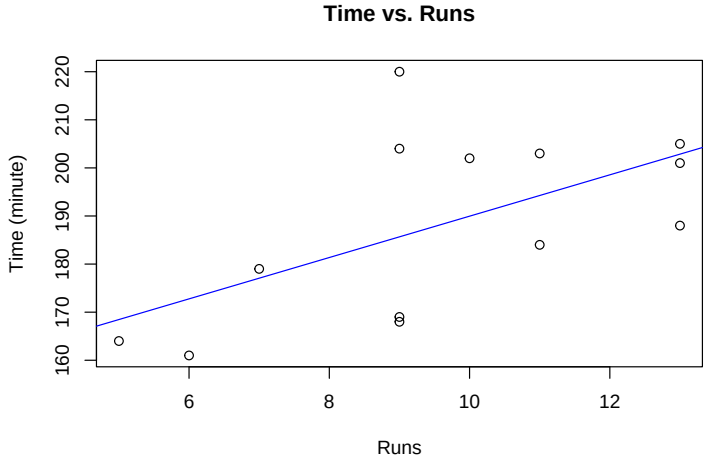




Problem 1.46

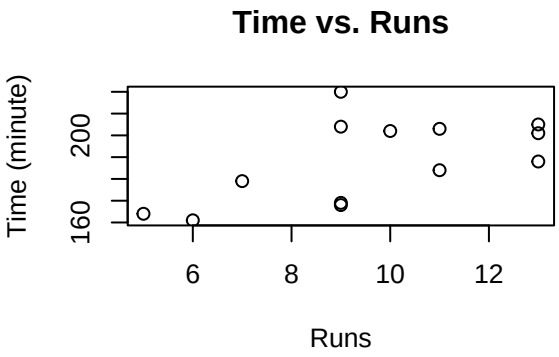
a) The CIN-MIL point has an extreme x-value and it is far from the mean value. It has leverage. However, the slope of the regression line without the CIN-MIL point would be 4.299 compared to 4.181. The slope would not change much, so the CIN-MIL point is not too influential.

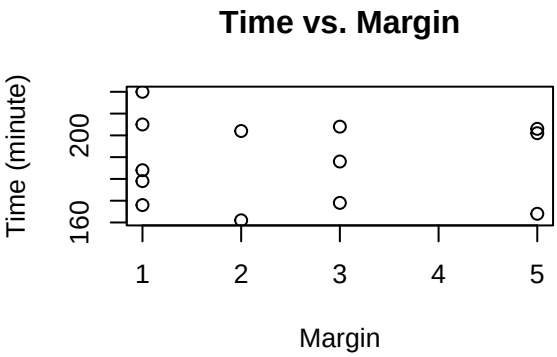
```
##
## Call:
## lm(formula = Time[Time != 235] ~ Runs[Time != 235], data = bball)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.662 -11.766  -1.858   8.740  34.338
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         146.972     17.673   8.316 4.51e-06 ***
## Runs[Time != 235]     4.299      1.779   2.416   0.0342 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.02 on 11 degrees of freedom
## Multiple R-squared:  0.3468, Adjusted R-squared:  0.2874
## F-statistic: 5.839 on 1 and 11 DF,  p-value: 0.03422
```

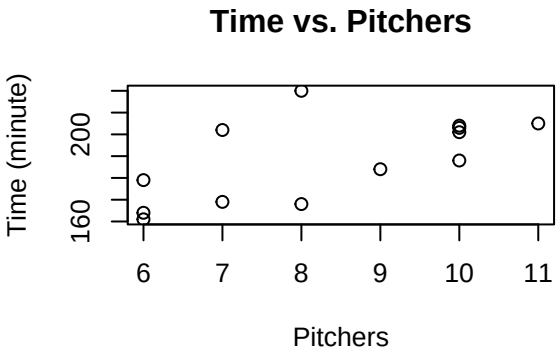


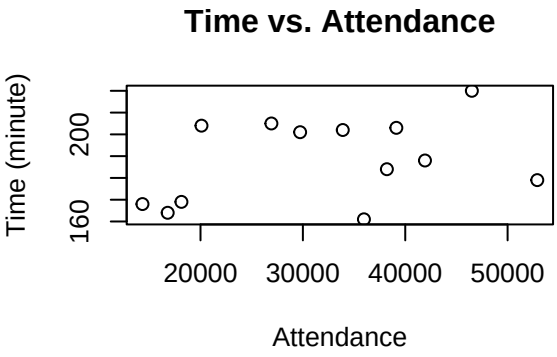
b) look to part a)

c) Removing the point had little change on the linearity of the other graphs. The CIN-MIL point has an influence on distinguishing the more linear pattern between Time vs. Runs and Time vs. Pitchers


Time vs. Runs
Time (minute)
160
200
6
8
10
12
Runs


Time vs. Margin
Time (minute)
160
200
1
2
3
4
5
Margin


Time vs. Pitchers
Time (minute)
160
200
6
7
8
9
10
11
Pitchers


Time vs. Attendance
Time (minute)
160
200
20000
30000
40000
50000
Attendance