hw 4

Stone Cai

2024-03-08

R Markdown

Problem 1

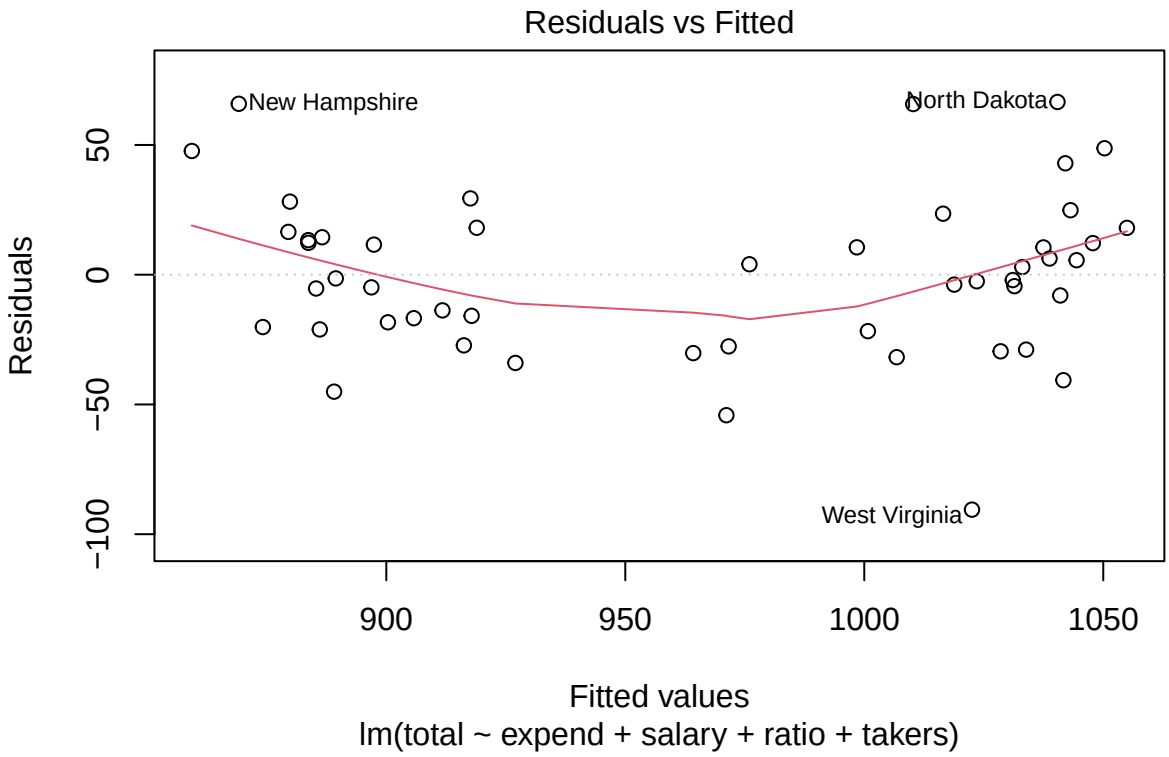
```
library (faraway)
```

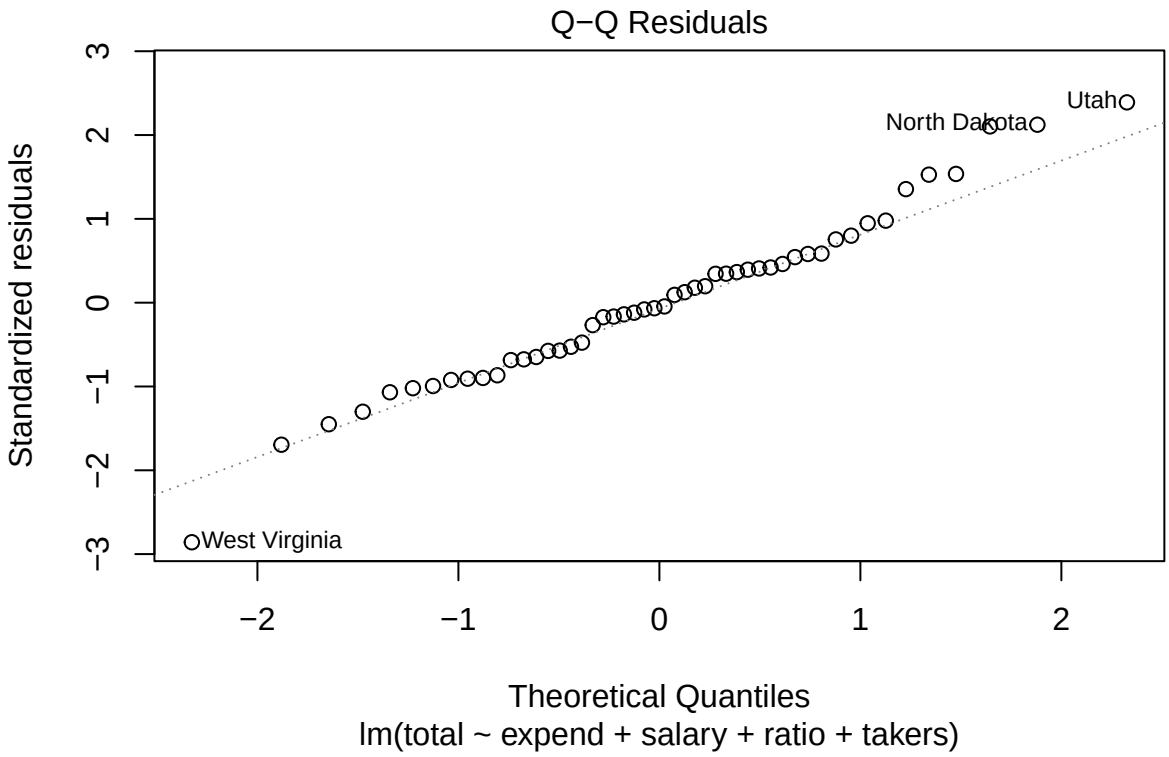
```
## Warning: package 'faraway' was built under R version 4.3.2
```

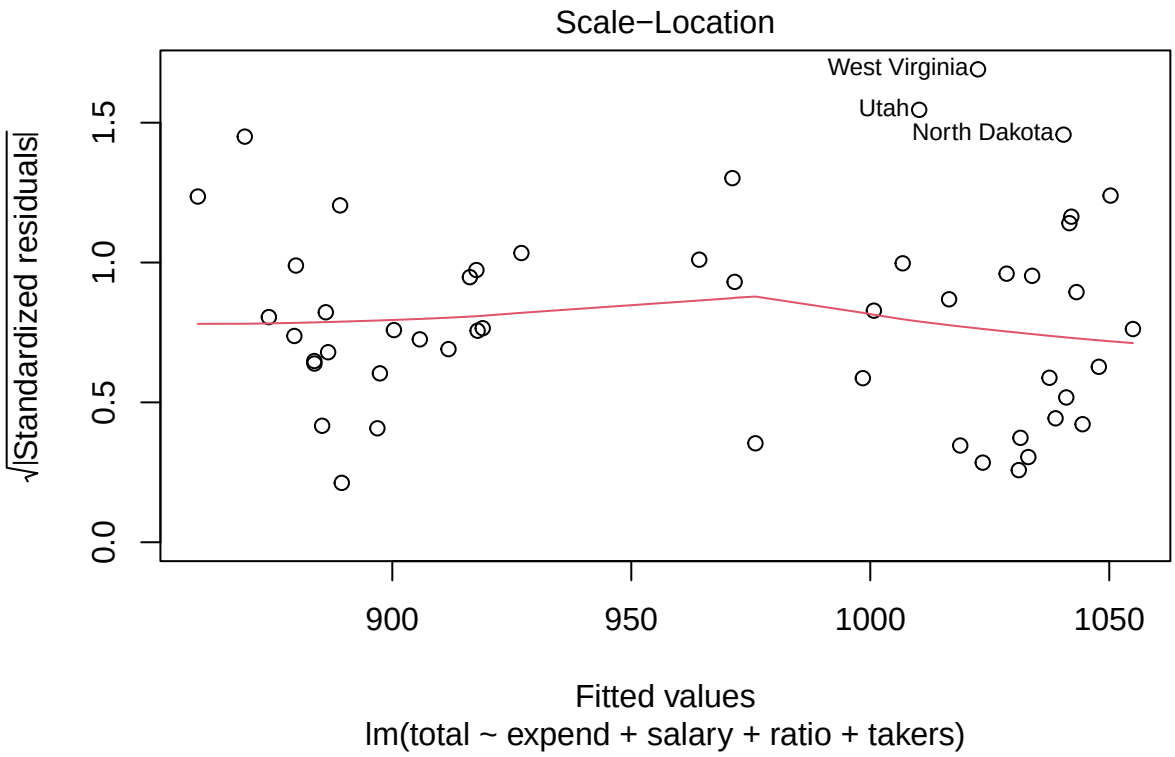
```
 data(sat)

 library(car)
```

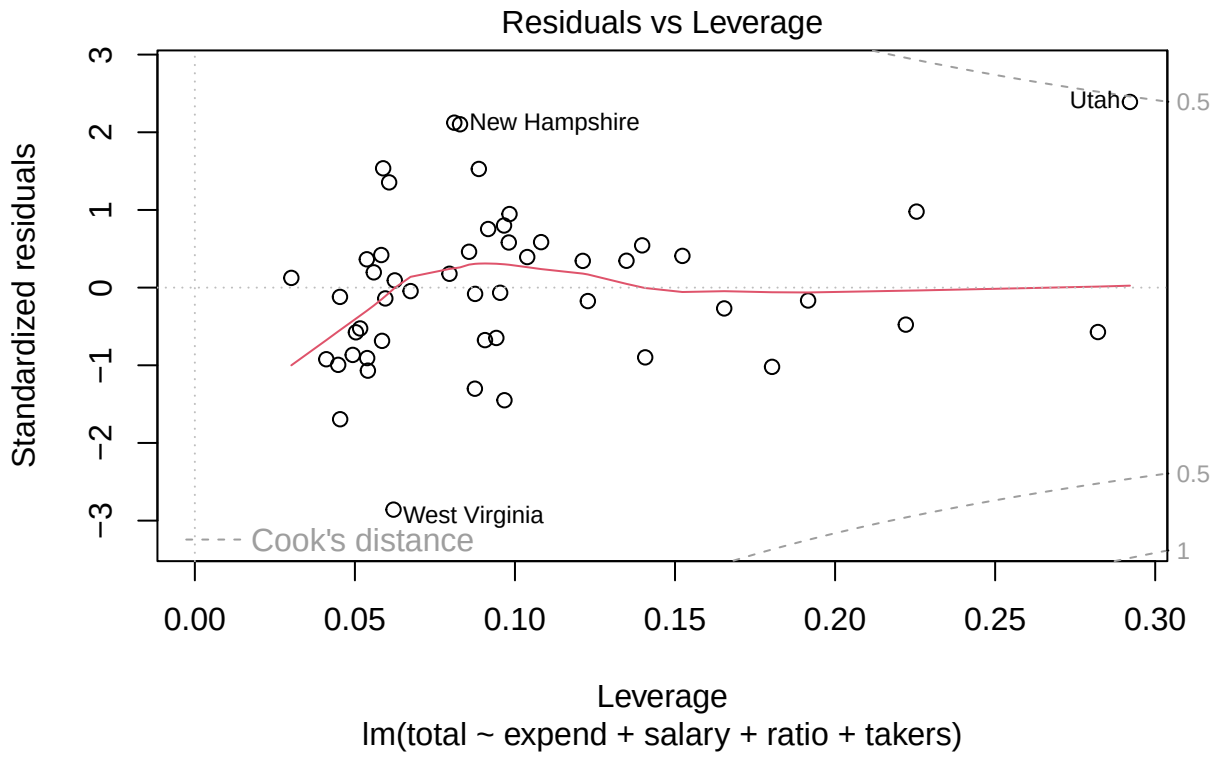
```
## Warning: package 'car' was built under R version 4.3.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.2

##
## Attaching package: 'car'

## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```
sat_mod <- lm( total~ expend+salary+ ratio+ takers, data= sat)

plot (sat_mod)
```


Residuals vs Fitted
New Hampshire
North Dakota
West Virginia
Residuals
50
0
−50
−100
900
950
1000
1050
Fitted values
lm(total ~ expend + salary + ratio + takers)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(total ~ expend + salary + ratio + takers)
Utah
North Dakota
West Virginia
−3
−2
−1
0
1
2
3

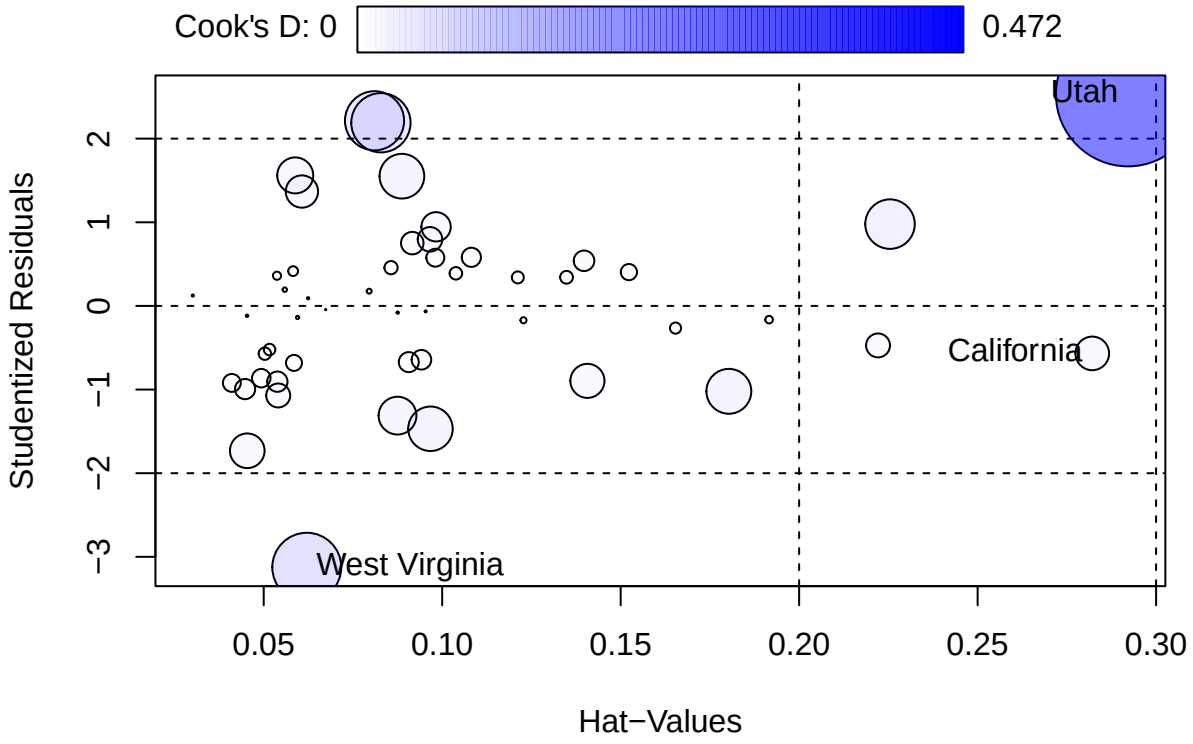

Scale−Location
West Virginia
Utah
North Dakota
√|Standardized residuals|
0.0
0.5
1.0
1.5
900
950
1000
1050
Fitted values
lm(total ~ expend + salary + ratio + takers)




```
outlierTest(sat_mod)
```

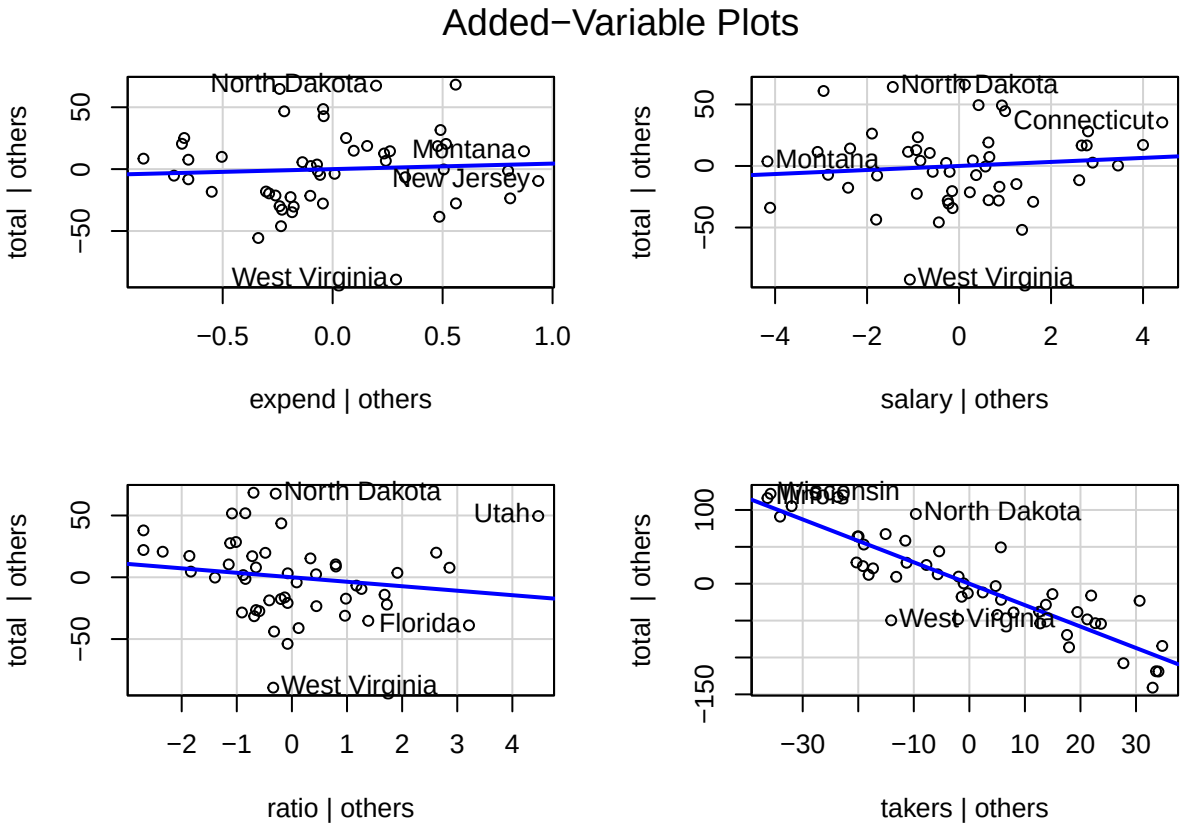
```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##                rstudent unadjusted p-value Bonferroni p
## West Virginia -3.124428          0.0031496      0.15748
```

```
influencePlot(sat_mod)
```




```
##                   StudRes        Hat      CookD
## California     -0.5676458 0.28211791 0.02571304
## Utah            2.5295873 0.29211280 0.47152866
## West Virginia -3.1244283 0.06206536 0.10813954
```

```
avPlots(sat_mod)
```



a.The plot of residuals vs fitted values has non-constant variance across all values of X. There is no violation of homoskedacity.

b) The QQ plot from above has the standardized residuals generally following the normal QQ line with a few small deviations. The assumption of normality is met here.

c). The residuals vs leverage plot shows there are two points with the highest leverage with almost 0.3 leverage. In this case, only Utah has a large Cook's distance of almost 0.5, while the other point has low influence despite being a high leverage point.

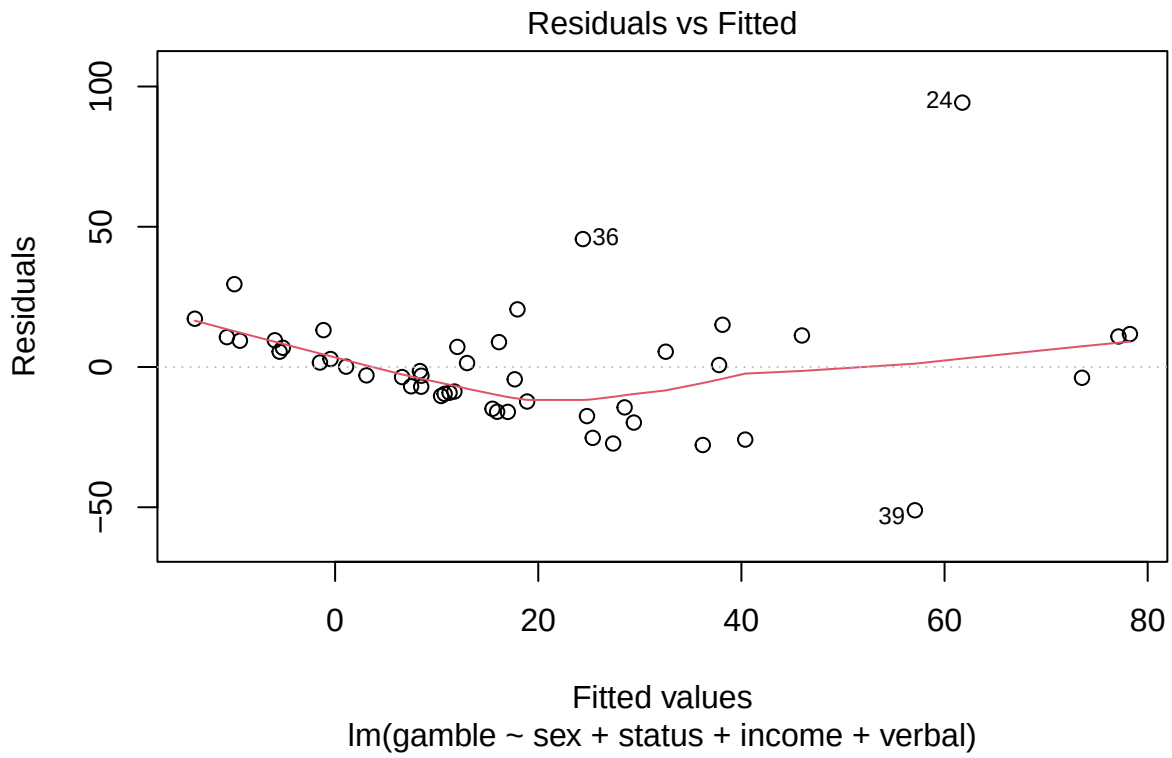
d) There is no true outlier in this model when we apply the outlier test function. The Bonferroni p-value is larger than 0.05 and the largest studentized residual is West Virginia with -3.12.

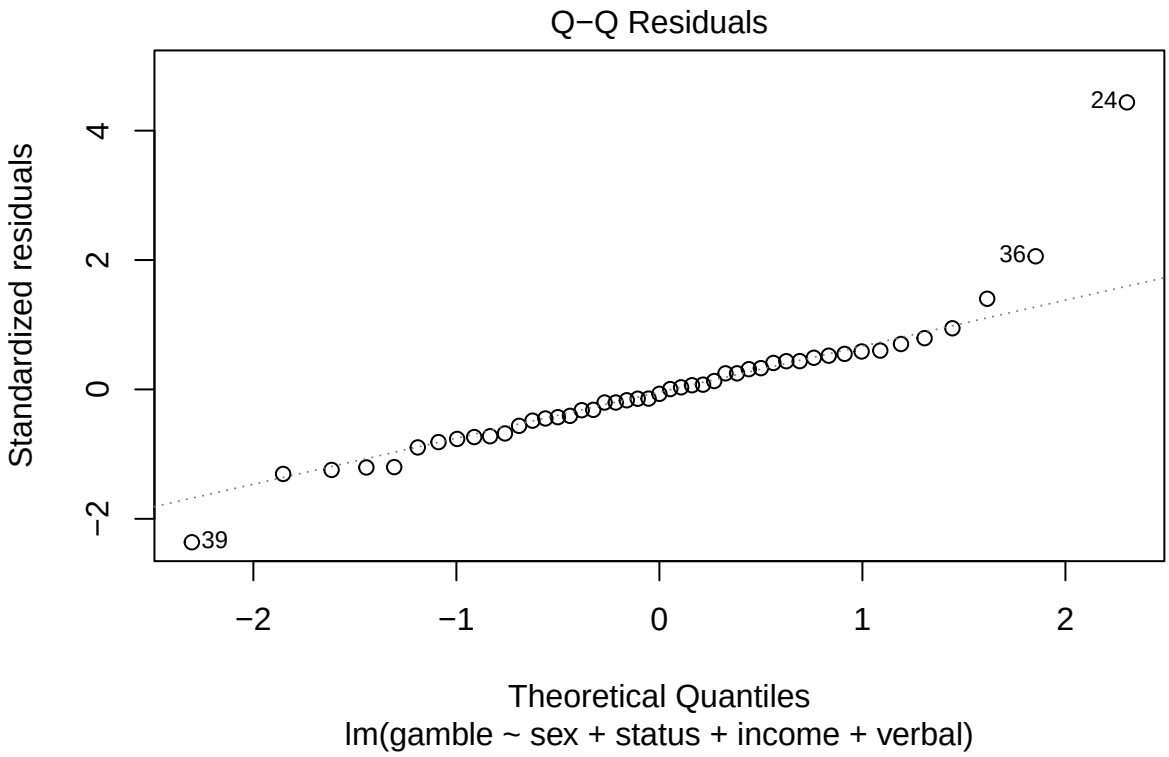
e). The influence plot shows that California and Utah have the highest leverage points, but Utah's Cook's distance of 0.472 is the largest among the points. Since we already determined that no point with the data sets are true outliers, this makes Utah's Cook's distance a little problematic, but still warrants considering exclusion from model.

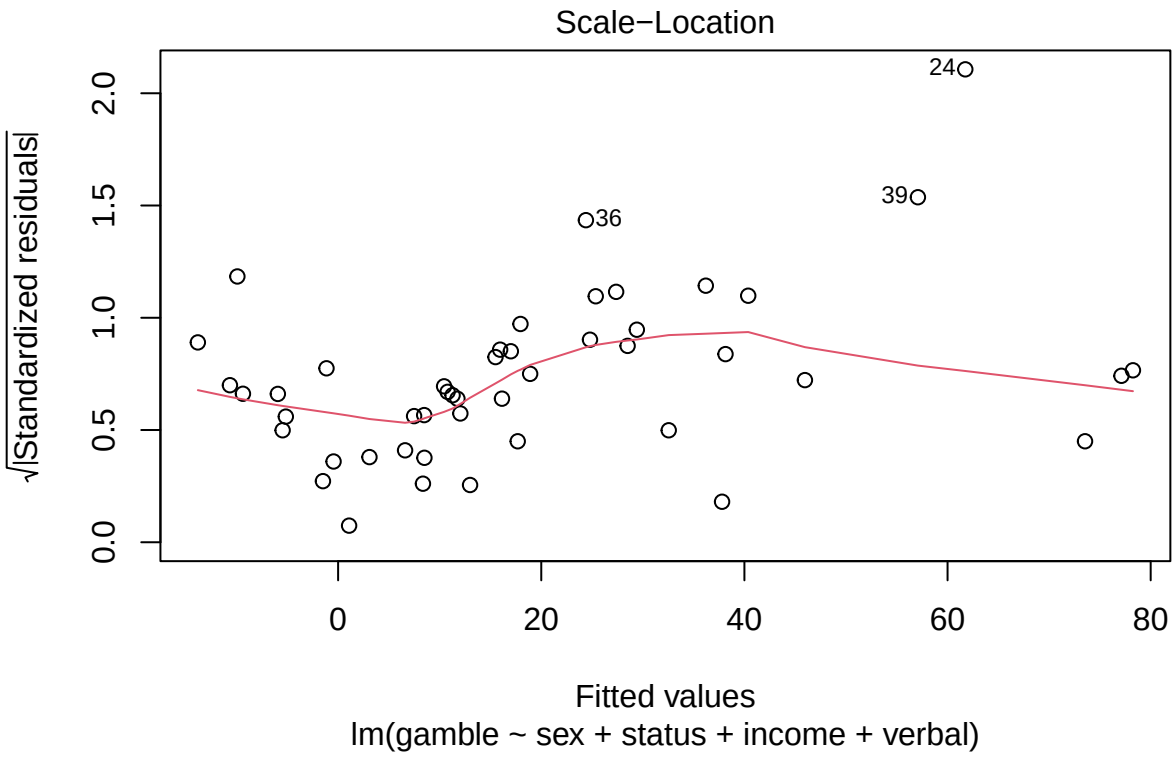
f) From Avplots above, there doesn't necessarily seem to be any structual problems with relationships between predictors and reponse variable total. They all seem to have linear relationships, but to varying degrees of strength. Expend and salary have very weak linear relationships with total. There is a weak negative relationship between ratio and total. Finally, there seems to be a strong negative relationship between total and takers.

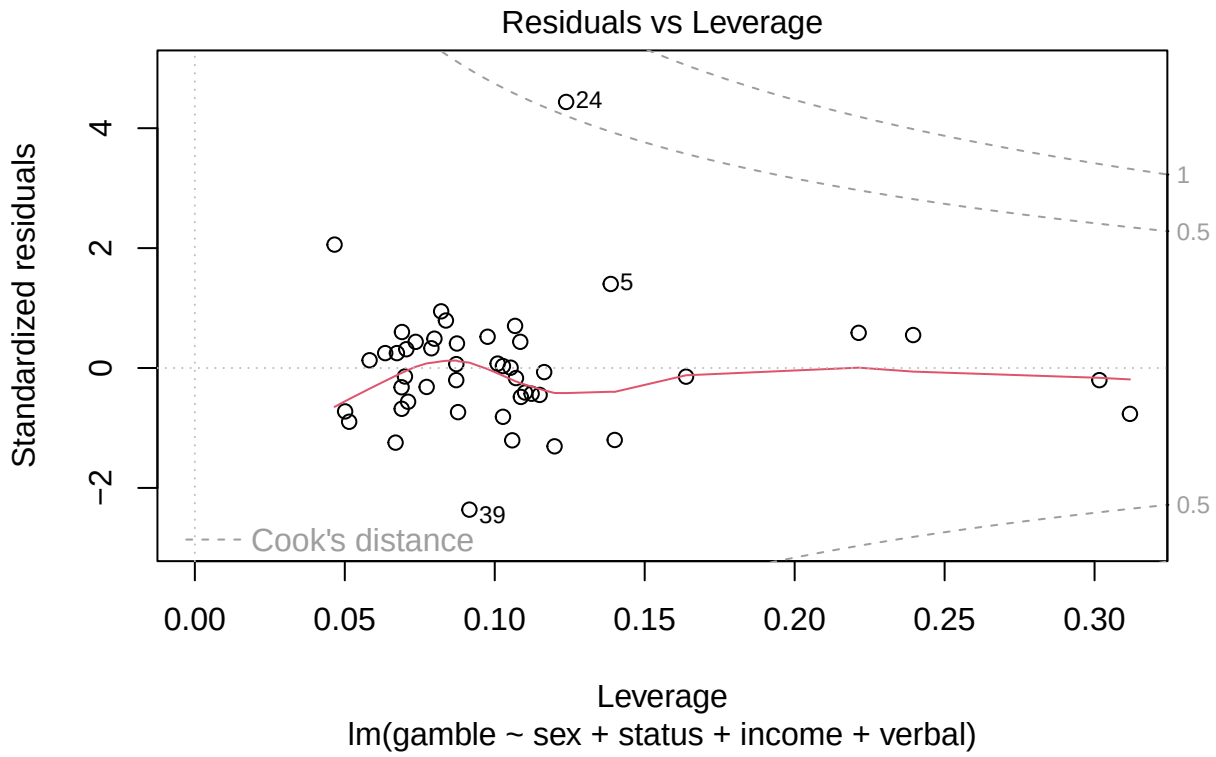
Problem 2

```r
library (faraway)
 data(teengamb)

 library(car)

teen_model<- lm( gamble ~ sex+status+ income+ verbal, data= teengamb)


plot (teen_model)
```




Q−Q Residuals
Standardized residuals
4
2
0
−2
24
36
39
−2
−1
0
1
2
Theoretical Quantiles
lm(gamble ~ sex + status + income + verbal)

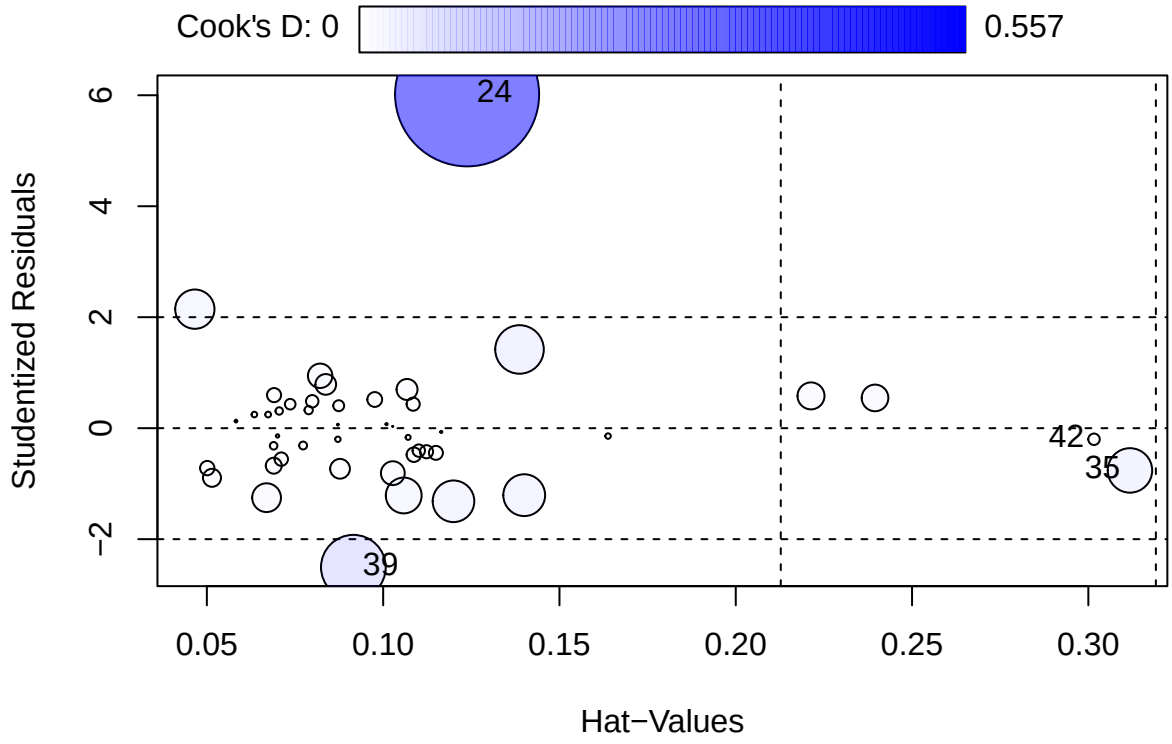

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
24
39
36
0
20
40
60
80
Fitted values
lm(gamble ~ sex + status + income + verbal)




```
outlierTest(teen_model)
```

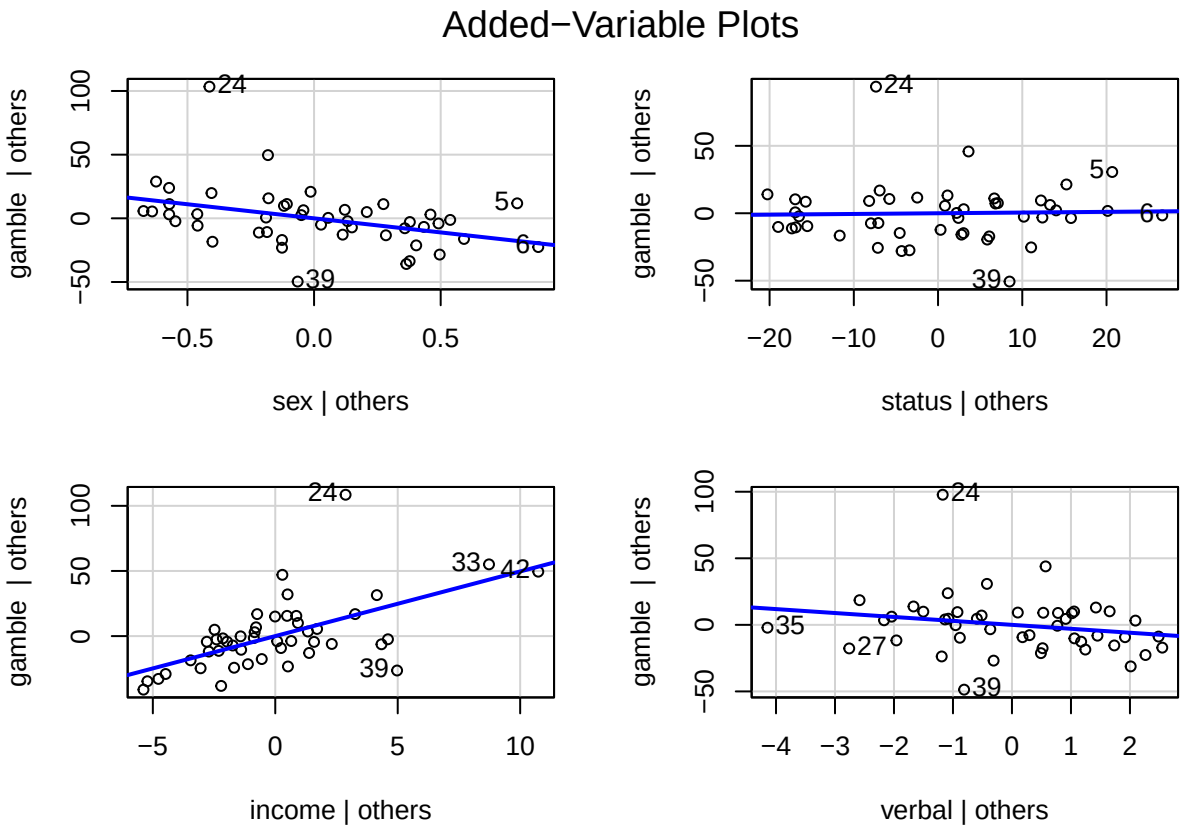
```
##    rstudent unadjusted p-value Bonferroni p
## 24 6.016116         4.1041e-07   1.9289e-05
```

```
influencePlot(teen_model)
```




```
##       StudRes        Hat      CookD
## 24  6.0161163 0.12380463 0.55650113
## 35 -0.7612557 0.31180294 0.05304304
## 39 -2.5060898 0.09155208 0.11244983
## 42 -0.1999795 0.30160877 0.00353499
```

```
avPlots(teen_model)
```



a. From the plot of residuals vs fitted values, I see there is heteroskedacity. The values are clumped for lower x fitted values, but variance grows for larger fitted x values which is a classic example of fanning out. Thus, our assumption of homoskedacity is violated.

b. From the QQ plot, i see that although many points do follow the QQ norm line, the residuals seem to follow a a skewed tail distribution and thus normality assumption is possibly violated if we include all points from the data.

c. To check for large leverage points, I consult the leverage vs residual plot. Here we have several points that have high leverage. There are 2 points with leverage > 0.3, but the plot also shows that their cooks distance (influence) is relatively small and non problematic.Point 24 has lower leverage, but a much large cooks distance which could indicate large influence (to be addressed later).

d. From the outlier test, there is one observation with the largest studentized residual of 6.016 and p-value less than 0.05. Observation 24 is an outlier.

e). As mentioned in part C), by utilizing the influence plot function, I see that observation 24 the outlier we found also indeed has large influence. It's cook distance of 0.557 is clearly the largest amongst the data points and is influential. Two large leverage points we found in Part b) (observations 35 and 42) do not have a large Cook's distance and are not influential.

f) From the avPlots, I can graph response variable against each predictor to analyze the structure of relationship between response and each predictor. Looking at the plot, it seems all relationships are indeed linear and a transformation of predictor or response is not needed. Gamble and sex have weak negative relationship, gamble and status has weak linear positive relationship, there is strong positive relationship between gamble and income, and there is a weak negative relationship between gamble and verbal.

Problem 3

```
df1 <- data.frame (x_values  = c(1.1, 1.4, 2.3, 3.7, 4.4, 2.7,3.3,4.1, 5.0, 3.2),
                   y_values = c(2.3, 2.9, 5.1, 7.6, 8.9, 5.7, 6.2, 8.0,11, 15.7 ) )

print(df1)
```

```
##    x_values y_values
## 1       1.1      2.3
## 2       1.4      2.9
## 3       2.3      5.1
## 4       3.7      7.6
## 5       4.4      8.9
## 6       2.7      5.7
## 7       3.3      6.2
## 8       4.1      8.0
## 9       5.0     11.0
## 10      3.2     15.7
```

```
lm1<- lm(y_values~x_values, data=df1)
summary(lm1)
```

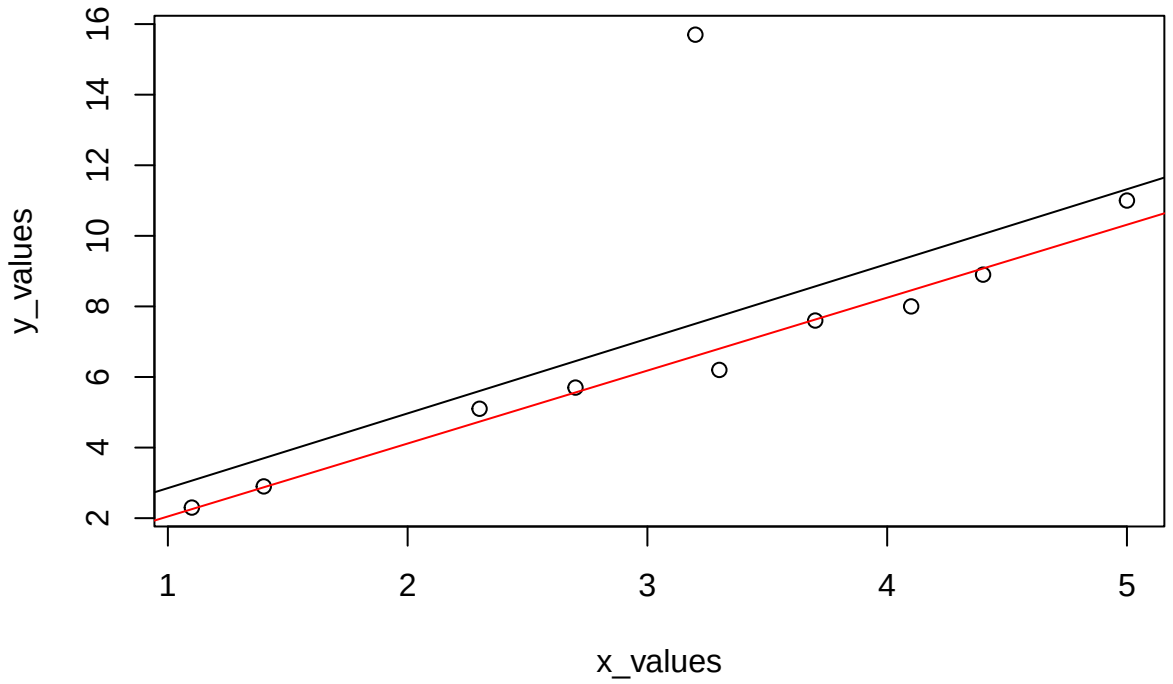
```
##
## Call:
## lm(formula = y_values ~ x_values, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5211 -1.1047 -0.7805 -0.5655  8.1906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7337     2.7119   0.271   0.7936
## x_values      2.1174     0.8113   2.610   0.0311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.078 on 8 degrees of freedom
## Multiple R-squared:  0.4599, Adjusted R-squared:  0.3924
## F-statistic: 6.812 on 1 and 8 DF,  p-value: 0.03113
```

```
df2 <- data.frame (x_values  = c(1.1, 1.4, 2.3, 3.7, 4.4, 2.7,3.3,4.1, 5.0),
                   y_values = c(2.3, 2.9, 5.1, 7.6, 8.9, 5.7, 6.2, 8.0,11 ) )

lm2<- lm(y_values~x_values, data=df2)
summary(lm2)
```

```
##
## Call:
## lm(formula = y_values ~ x_values, data = df2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.60151 -0.17498  0.02541  0.13857  0.68494
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01892    0.37031  -0.051    0.961
## x_values     2.06680    0.11027  18.743 3.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4183 on 7 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.9777
## F-statistic: 351.3 on 1 and 7 DF,  p-value: 3.056e-07
```

```r
plot(df1)
abline(lm1)
abline(lm2, col="red")
```



Here we plotted 9 points that have a strong correlation and have a strong adjusted R^2 of 0.977.The 10th point, has a nearly median x-value (low leverage), but is clearly an outlier. When we plot the black regression line with all 10 points, we obtain a value of 2.11 for $\hat{\beta}_1$ and 0.733 for $\hat{\beta}_0$. The p-value for the $\hat{\beta}_1$ coefficient is significant, but overall adjusted R^2 is only 0.3924 for the model. If we remove the outlier point, the regression line essentially shifts down. Here we have value of 2.07 for $\hat{\beta}_1$ and -0.01 for $\hat{\beta}_0$. The slopes of both lines are essentially unchanged, but the value of intercept is larger in original regression line to account for large outlier. Slope doesn't seem to change much given the outlier is a low leverage point. The regression coefficent for the model without the outlier has a much lower p-value.

problem 4

```
df4 <- data.frame (x_values  = c(1.1, 1.4, 2.3, 3.7, 4.4, 2.7,3.3,4.1, 5.0, 30),
                   y_values = c(2.2, 2.7, 4.7, 7.4, 8.7, 5.4, 6.6, 8.2,10.1, 61) )

print(df4)
```

```
##    x_values y_values
## 1       1.1      2.2
## 2       1.4      2.7
## 3       2.3      4.7
## 4       3.7      7.4
## 5       4.4      8.7
## 6       2.7      5.4
## 7       3.3      6.6
## 8       4.1      8.2
## 9       5.0     10.1
## 10     30.0     61.0
```

```
lm4<- lm(y_values~x_values, data=df4)
summary(lm4)
```

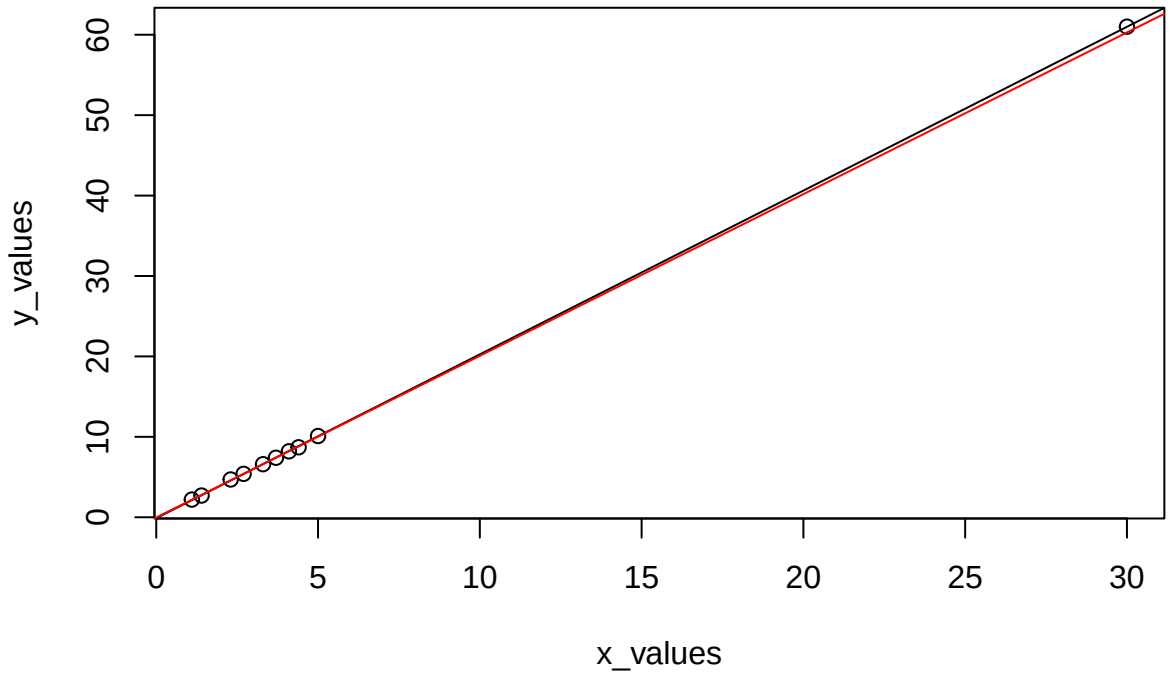
```
##
## Call:
## lm(formula = y_values ~ x_values, data = df4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.148745 -0.034100  0.002511  0.025470  0.128139
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.11234    0.03022  -3.717  0.00589 **
## x_values     2.03661    0.00302 674.420  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07788 on 8 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.548e+05 on 1 and 8 DF,  p-value: < 2.2e-16
```

```
df5<- data.frame (x_values  = c(1.1, 1.4, 2.3, 3.7, 4.4, 2.7,3.3,4.1, 5.0 ),
                   y_values = c(2.2, 2.7, 4.7, 7.4, 8.7, 5.4, 6.6, 8.2,10.1) )

lm5<- lm(y_values~x_values, data=df5)
summary(lm5)
```

```
##
## Call:
## lm(formula = y_values ~ x_values, data = df5)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.113436 -0.010309 -0.001969  0.020965  0.108456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03243    0.06560  -0.494    0.636
## x_values     2.01042    0.01953 102.915 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0741 on 7 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9992
## F-statistic: 1.059e+04 on 1 and 7 DF,  p-value: 2.155e-12
```

```
plot(df4)
abline(lm4)
abline(lm5, col="red")
```



Here we have 10 points that follow a regression line very closely.Our 10th point is a point that follows the regression pretty closely, but has a extremely large x value of 30, when most of the data ranges from 1.1-5. When we plot both lines, we have very similar lines with $\hat{\beta}_1$ coefficient being roughly 2 for both lines and with similar intercepts. As we can see, a large leverage point is not necessarily problematic if it isn't an outlier. For both lines, the standard error for the $\hat{\beta}_1$ coefficient is very small and both have large t-values.

problem 5

```r
df6 <- data.frame (x_values  = c(1.1, 1.4, 2.3, 3.7, 4.4, 2.7,3.3,4.1, 5.0, 30),
                   y_values = c(3.3, 4, 7, 10, 12.4, 6.3, 9.8, 13,14.8, 3.2) )

print(df6)
```

```
##    x_values y_values
## 1       1.1      3.3
## 2       1.4      4.0
## 3       2.3      7.0
## 4       3.7     10.0
## 5       4.4     12.4
## 6       2.7      6.3
## 7       3.3      9.8
## 8       4.1     13.0
## 9       5.0     14.8
## 10     30.0      3.2
```

```r
lm6<- lm(y_values~x_values, data=df6)
summary(lm6)
```

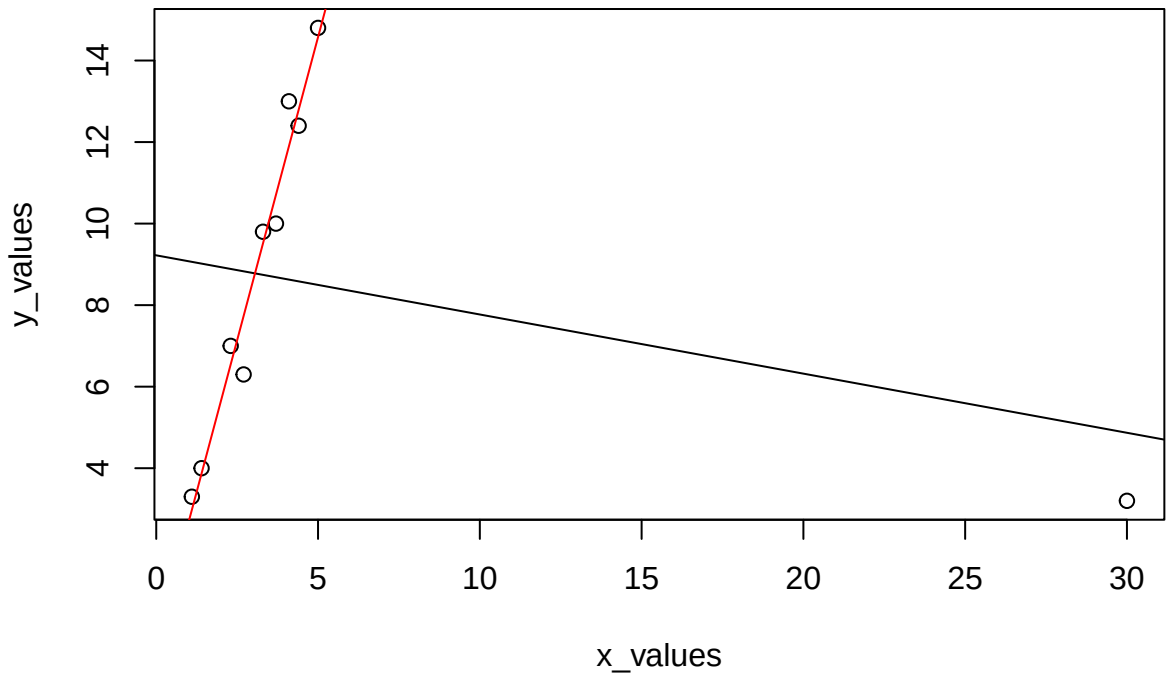
```
##
## Call:
## lm(formula = y_values ~ x_values, data = df6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7619 -2.3693 -0.3058  3.1915  6.3039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.2215     1.6657   5.536  0.00055 ***
## x_values     -0.1451     0.1664  -0.872  0.40878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.293 on 8 degrees of freedom
## Multiple R-squared:  0.08674,    Adjusted R-squared:  -0.02742
## F-statistic: 0.7598 on 1 and 8 DF,  p-value: 0.4088
```

```r
df7 <- data.frame (x_values  = c(1.1, 1.4, 2.3, 3.7, 4.4, 2.7,3.3,4.1, 5.0),
                   y_values = c(3.3, 4, 7, 10, 12.4, 6.3, 9.8, 13,14.8) )

lm7<- lm(y_values~x_values, data=df7)
summary(lm7)
```

```
##
## Call:
## lm(formula = y_values ~ x_values, data = df7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4331 -0.3880  0.2280  0.3244  1.1040
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2951     0.7007  -0.421    0.686
## x_values      2.9734     0.2087  14.250 1.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7915 on 7 degrees of freedom
## Multiple R-squared:  0.9667, Adjusted R-squared:  0.9619
## F-statistic: 203.1 on 1 and 7 DF,  p-value: 1.992e-06
```

```r
plot(df6)
abline(lm6)
abline(lm7, col="red")
```



Here, i chose 9 points with a strong positive correlation with large adjust R^2 of 0.96 and an extremely large outlier and leverage point (30, 3.2). With all 10 points, if we look at the R output summary, the original regression line suggests a negative correlation between y and x values. Additionally, we see the p-value of the $\hat{\beta}_1$ coefficient is not significant.If we takeout the large outlier+leverage point, then we see the remaining 9 points actually have a very strong positive correlation with a significant $\hat{\beta}_1$ with a slope close to 3. Here we can demonstrate that a large outlier and leverage point causes the largest problem as it drastically shifts the regression line of the true model. The first regression line also has a positive intercept value that accounts for large outlier+leverage point, in the second lane this intercept value is nearly zero.