# STATS 413 Hw3

Shu Zhou

2020/10/22

This is the Assignment 3 of STATS 413 Author: Shu Zhou UMID: 19342932

## Exercise 12

**(a.)** Assume $y = \beta_a x + \epsilon_i$ and $x = \beta_b y + \epsilon_j$, Hence, the OLS estimator

$$\hat{\beta_a} = \frac{\sum_i^n x_i y_i}{\sum x_i^2} \tag{1}$$

$$\hat{\beta_b} = \frac{\sum_i^n x_i y_i}{\sum y_i^2} \tag{2}$$

Hence, when the beta denominators are equal $\sum x_i^2 = \sum y_i^2$, the coefficent of estimate are equal

**(b.)**

```
set.seed(100)
x<-rnorm(100)
y<-5*x + rnorm(100)
lmX<-lm(y~x)
lmY<-lm(x~y)
summary(lmX)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05195 -0.43265 -0.07854  0.48583  1.93858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01145    0.07929   0.144    0.885
## x            4.89463    0.07807  62.693   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7929 on 98 degrees of freedom
## Multiple R-squared:  0.9757, Adjusted R-squared:  0.9754
## F-statistic:  3930 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
summary(lmY)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43696 -0.08583  0.01513  0.08913  0.43275
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002211   0.016001  -0.138     0.89
## y            0.199335   0.003180  62.693   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.16 on 98 degrees of freedom
## Multiple R-squared:  0.9757, Adjusted R-squared:  0.9754
## F-statistic:  3930 on 1 and 98 DF,  p-value: < 2.2e-16
```

It is obvious that $\hat{\beta}_a \neq \hat{\beta}_b$

**(c.)**_

```r
set.seed(1)
x = rnorm(100, mean=1000, sd=0.1)
y = x
lmY <- lm(y ~ x)
lmX <- lm(x ~ y)
summary(lmY)
```

```
## Warning in summary.lm(lmY): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.116e-16  5.070e-18  7.940e-18  1.102e-17  9.054e-17
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 9.095e-13  1.042e-13 8.725e+00 7.03e-14 ***
## x           1.000e+00  1.042e-16 9.594e+15  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.316e-17 on 98 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 9.204e+31 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
summary(lmX)
```

```
## Warning in summary.lm(lmX): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
```

```
## lm(formula = x ~ y)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.116e-16  5.070e-18  7.940e-18  1.102e-17  9.054e-17
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 9.095e-13  1.042e-13 8.725e+00 7.03e-14 ***
## y           1.000e+00  1.042e-16 9.594e+15  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.316e-17 on 98 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 9.204e+31 on 1 and 98 DF,  p-value: < 2.2e-16
```

It is obvious that $\hat{\beta}_a = \hat{\beta}_b = 1$

# Exercise 14

### (a.)

```r
set.seed (1)
x1=runif (100)
x2 =0.5* x1+rnorm (100) /10
y=2+2* x1 +0.3* x2+rnorm (100)
```

The form of the regression model is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{3}$$

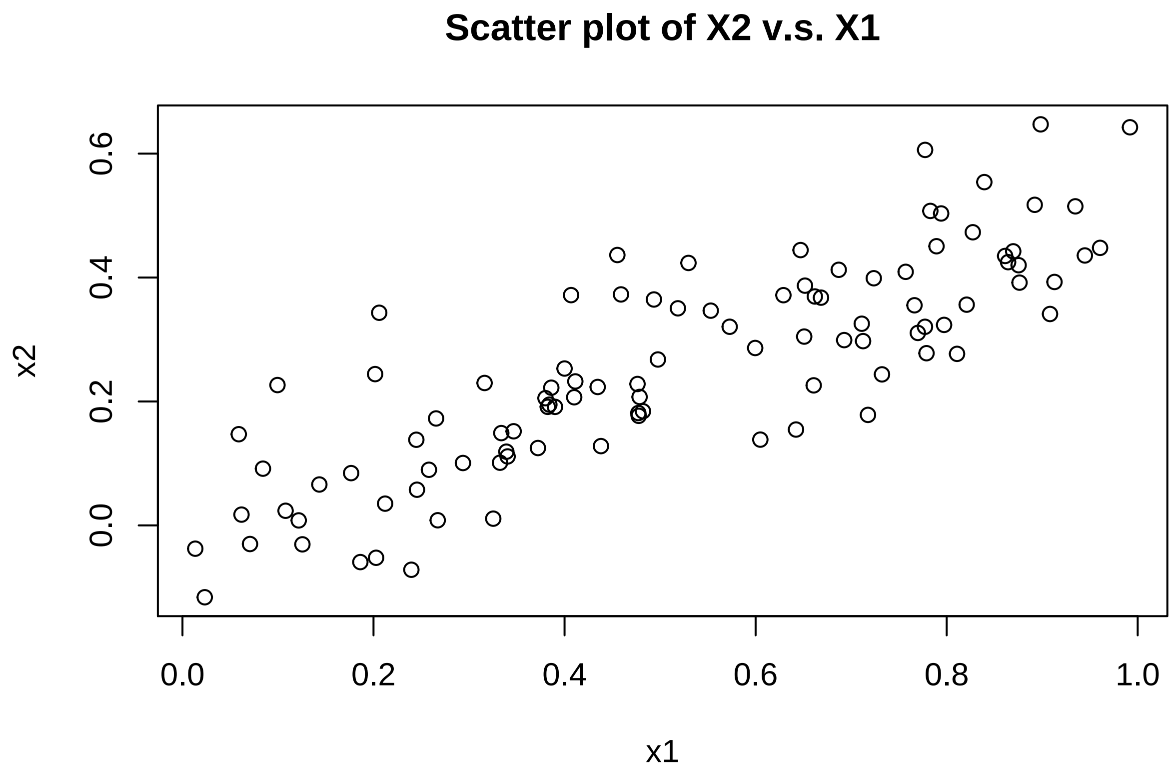
Where $\beta_0 = 2$, $\beta_1 = 2$ and $\beta_2 = 0.3$

### (b.)

```r
cor(x1,x2)
```

```
## [1] 0.8351212
```

```r
plot(x1,x2, main = "Scatter plot of X2 v.s. X1")
```




**(c.)**

```r
model_3 <- lm(y~x1+x2)
summary(model_3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

- $\hat{\beta}_0 = 2.1305$ $(\beta_0 = 2)$
- $\hat{\beta}_1 = 1.4396$ $(\beta_1 = 2)$

- $\hat{\beta}_2 = 1.0097$ ($\beta_2 = 0.3$)

we can reject $H_0 : \beta_1 = 0$; but we cannot reject $H_0 : \beta_2 = 0$

**(d.)**

```r
model_4<- lm(y~x1)
summary(model_4)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

We can reject $H_0 : \beta_1 = 0$

**(e.)**

```r
model_5<- lm(y~x2)
summary(model_4)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

We can reject $H_0 : \beta_2 = 0$

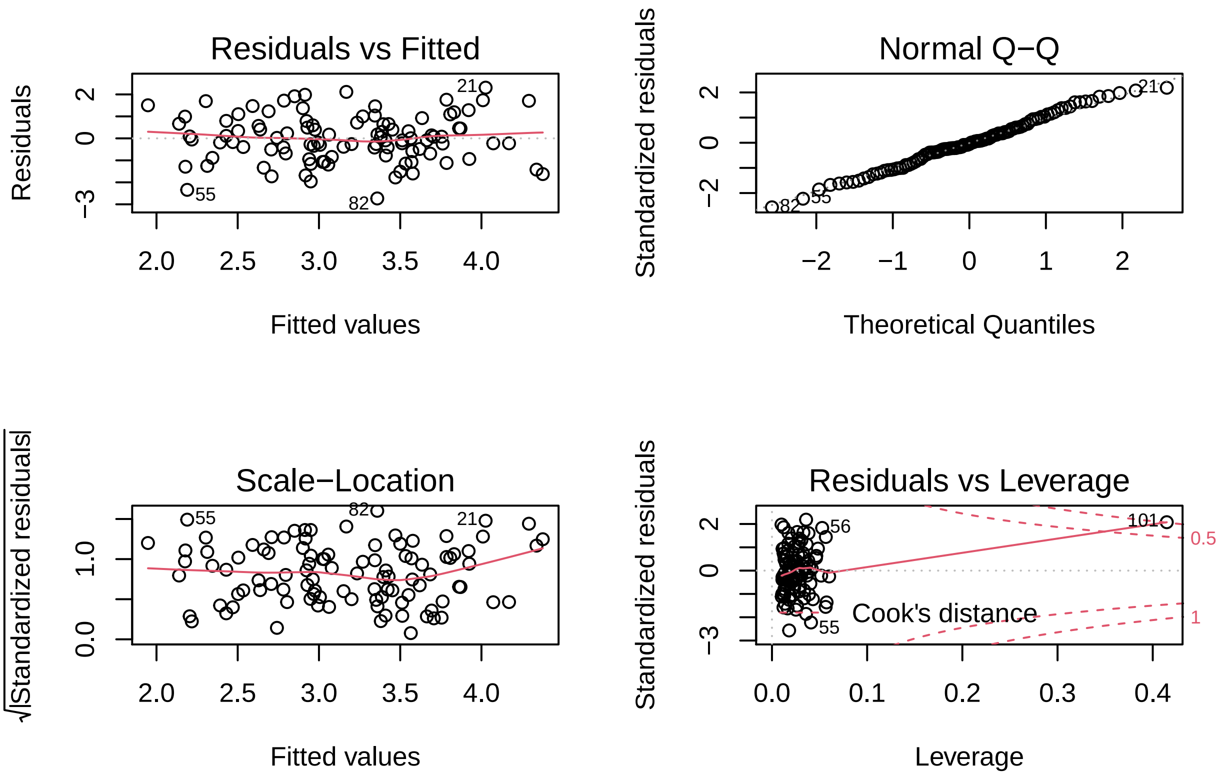
**(f.)** The results from (c.) to (e.) do not contradict each other.

Without the presence of other predictors, both $\beta_1$ and $\beta_2$ are statistically significant. However, x2 does not provide sufficiently new information when fitting a model that already contains x1. Hence,in the presence of other predictors, $\beta_2$ is no longer statistically significant. **(g.)**

```
x1=c(x1 , 0.1)
x2=c(x2 , 0.8)
y=c(y,6)
par(mfrow=c(2,2))
# regression with both x1 and x2
model_6 <- lm(y~x1+x2)
summary(model_6)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

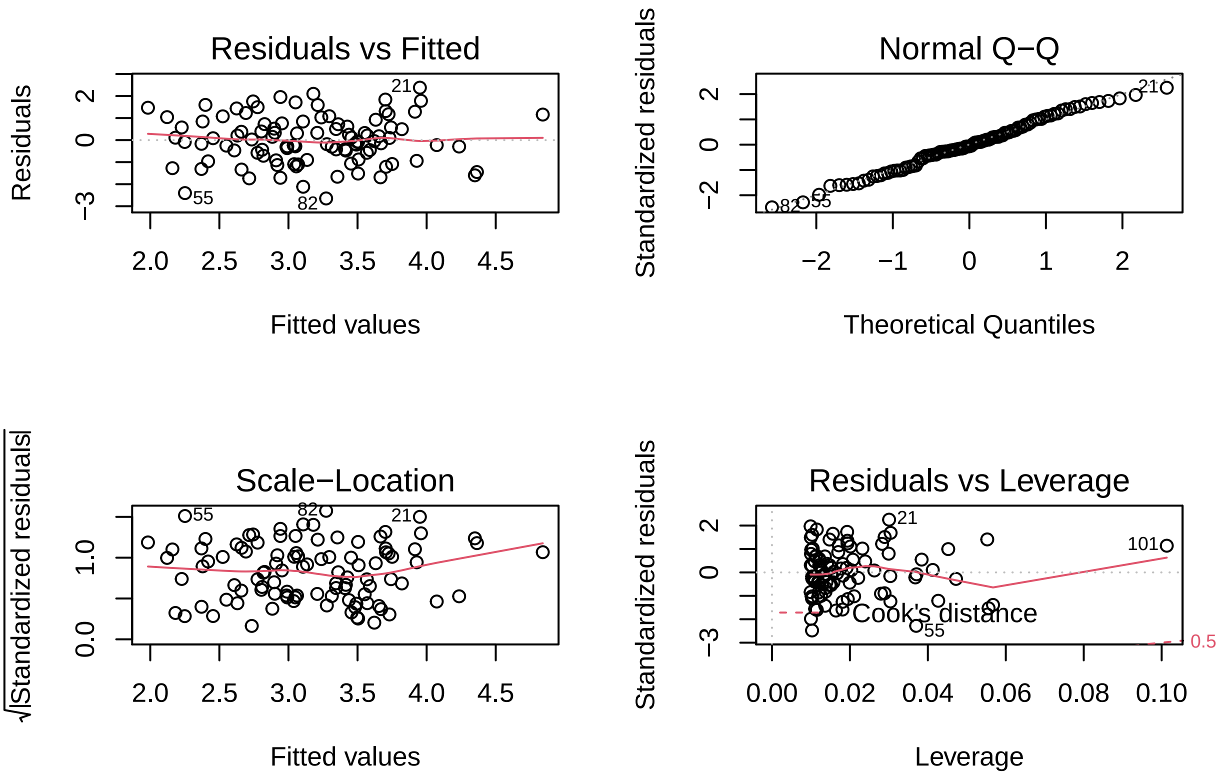
```
plot(model_6)
```




```r
# regression with x1 only
model_7 <- lm(y~x2)
summary(model_7)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

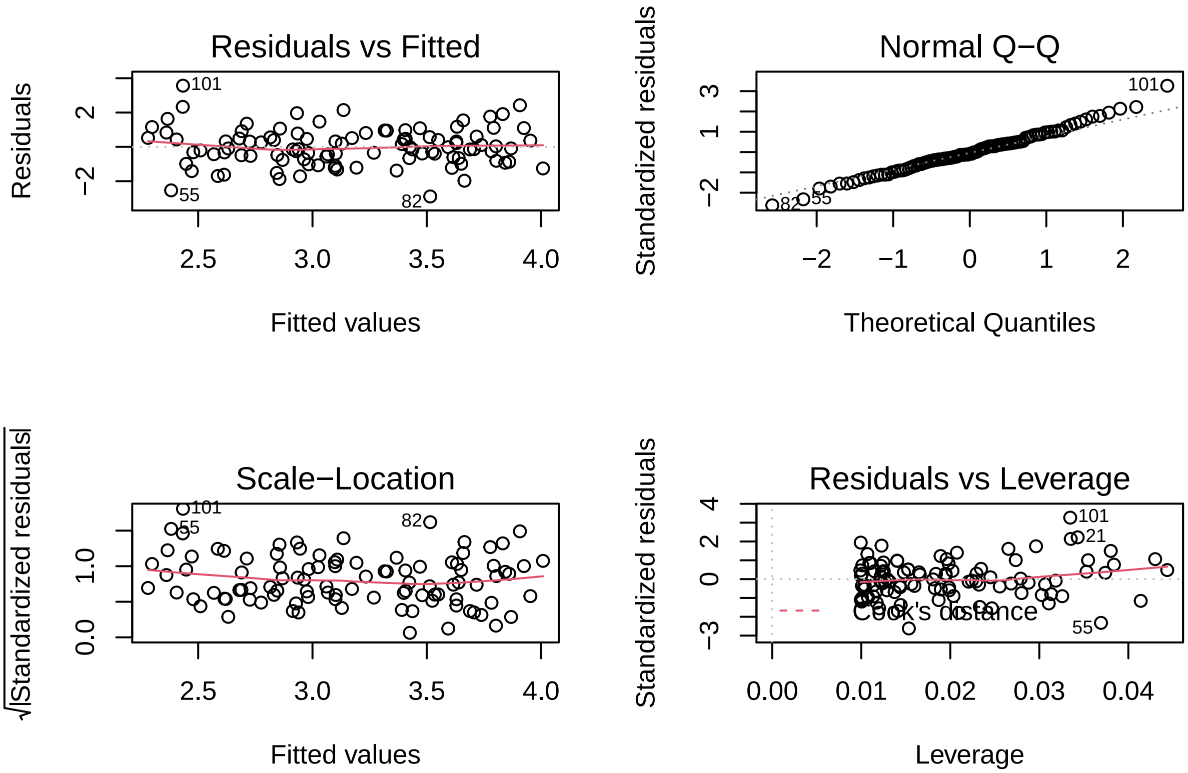
```r
plot(model_7)
```




```r
# regression with x2 only
model_8 <- lm(y~x1)
summary(model_8)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```r
plot(model_8)
```



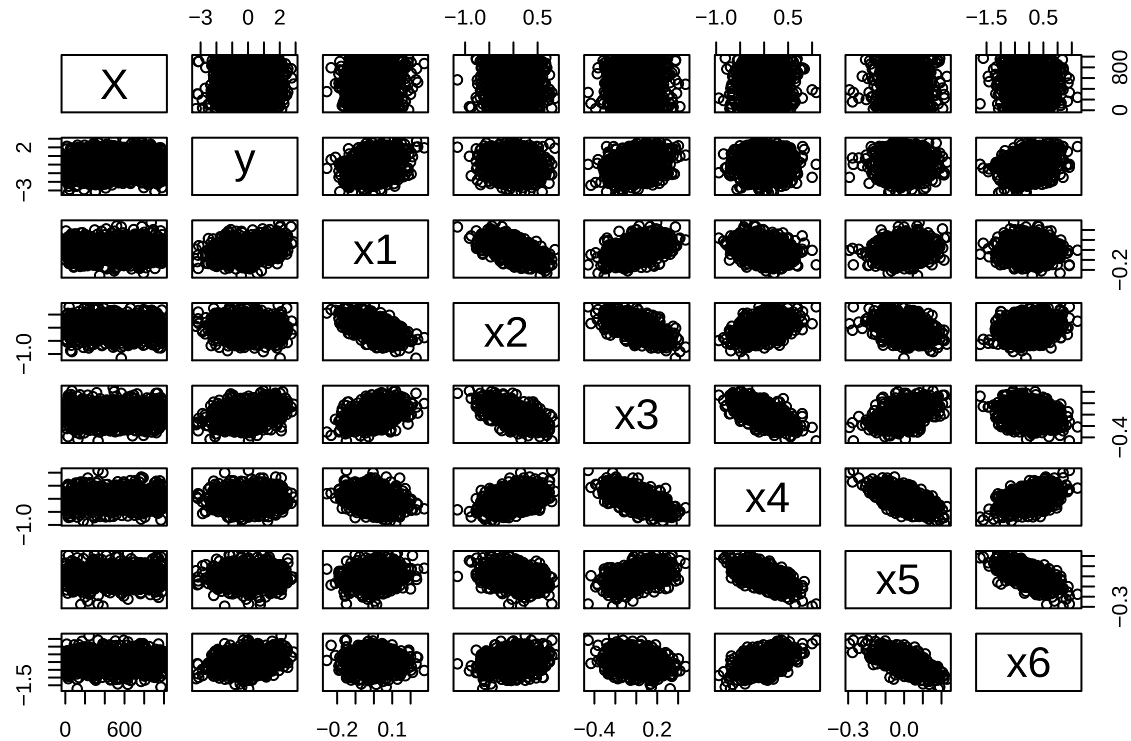

- In the regression with both x1 and x2, we can see that the new observation has the highest leverage and residual, which can be considered as an outlier.
- In the regression with x1, the new observation is still fairly high-leverage and have a large residual, so it can also be considered and an outlier.
- In the regression with x1, the new observation is still fairly high-leverage and have a large residual, so it can also be considered and an outlier.

Hence, for this model, the new observation might not be considered influential, since in all cases it can be regarded as an outlier.

## Exercise 9.1 in ALR

**(9.1.1)**

```
Rpdata<-read.csv("Rpdata.csv")
pairs(Rpdata)
```




```r
cor(Rpdata)
```

```
##                X           y          x1           x2          x3          x4
## X   1.000000000  0.05447269  0.03286048 -0.009940518  0.01728578 -0.01476303
## y   0.054472693  1.00000000  0.27576161 -0.081404813  0.28757845  0.02805570
## x1  0.032860477  0.27576161  1.00000000 -0.614978331  0.42582264 -0.25458872
## x2 -0.009940518 -0.08140481 -0.61497833  1.000000000 -0.63525309  0.46282824
## x3  0.017285783  0.28757845  0.42582264 -0.635253087  1.00000000 -0.61864216
## x4 -0.014763030  0.02805570 -0.25458872  0.462828237 -0.61864216  1.00000000
## x5  0.005223843 -0.00872376  0.14966775 -0.320733362  0.43166796 -0.64672789
## x6 -0.019799741  0.28570393 -0.05134618  0.252121520 -0.25838336  0.48735067
##               x5          x6
## X   0.005223843 -0.01979974
## y  -0.008723760  0.28570393
## x1  0.149667747 -0.05134618
## x2 -0.320733362  0.25212152
## x3  0.431667964 -0.25838336
## x4 -0.646727889  0.48735067
## x5  1.000000000 -0.64830627
## x6 -0.648306268  1.00000000
```

From the scatter plot and correlation, we can see tht the correlations between x1 and x2, x2 and x3, x3 and x4, x4 and x5, x5 and x6 are all relatively high, which might cause colinearity. **(9.1.2)**

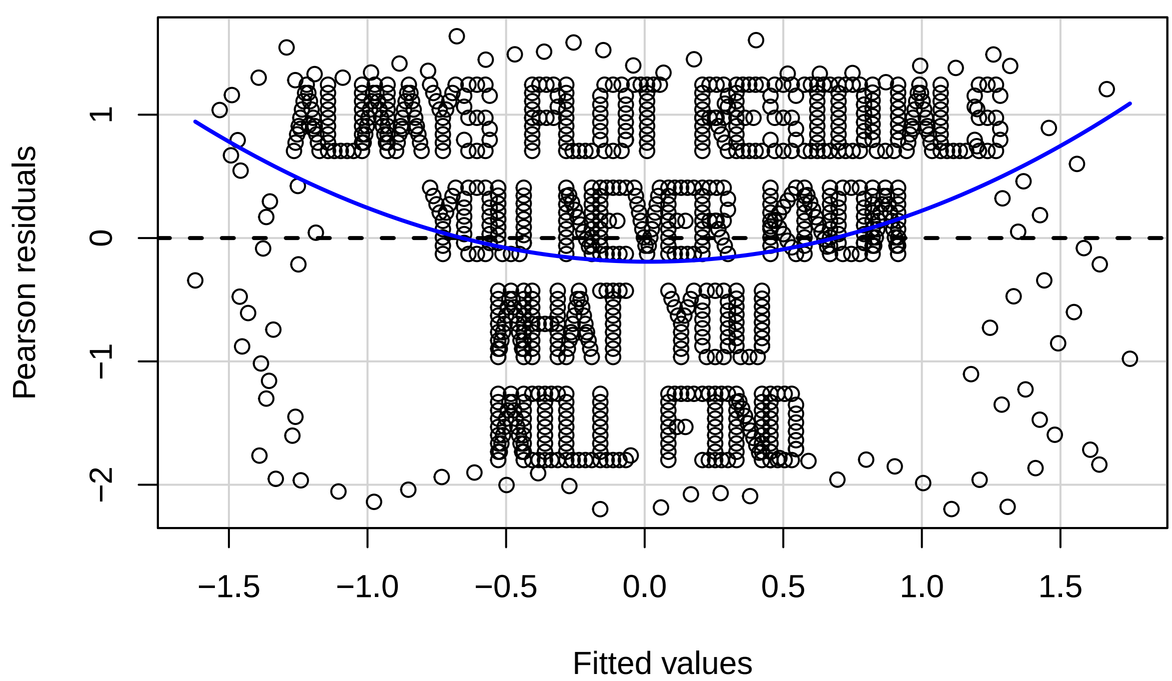
```r
lm.model<-lm(y~x1+x2+x3+x4+x5+x6,data=Rpdata)
summary(lm.model)
```

```
##
```

```
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1977 -0.7631  0.1729  0.8851  1.6359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02481    0.03188   0.778    0.437
## x1           4.14061    0.50954   8.126 1.32e-15 ***
## x2           1.01233    0.15522   6.522 1.11e-10 ***
## x3           3.99614    0.32663  12.234  < 2e-16 ***
## x4           0.96045    0.16657   5.766 1.09e-08 ***
## x5           3.75122    0.64726   5.796 9.17e-09 ***
## x6           0.95390    0.08561  11.142  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.003 on 983 degrees of freedom
## Multiple R-squared:  0.3112, Adjusted R-squared:  0.307
## F-statistic: 74.03 on 6 and 983 DF,  p-value: < 2.2e-16
```

There is nothing strange with the regression coefficients of variables, but the coefficient of intercept is non-significant. **(9.1.3)**

```
residualPlot(lm.model)
```



The plot says "Always plot residuals, you never know what you will find." But the residuals itself is strange since it is has some relationship with the fitted values.

## Exercise 9.4 in ALR

**(9.4.1)**

$$h_{ij} = x_i^T (X^T X)^{-1} x_i = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX} \tag{4}$$

Hence for the leverages $h_{ii}$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} \tag{5}$$

**(9.4.2)** The cases with large leverage will have extreme $(x_i - \bar{x})^2$ values, hence the values on the extremely left or right side of the scatter plot will have high leverage values. **(9.4.3)** We simply let n = 1, hence in this case $x_i = \bar{x}$ Hence

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} = 1 + 0 = 1 \tag{6}$$