STAT 4130: Homework3

Wang Boqian

2023-06-21

```r
# if you are using libraries, it's good practice to load them here
library(ggplot2)
```

Question 1

```r
# please do your coding inside a code chunk
# unless otherwise stated, feel free to do all computations in R
# commented code is always appreciated

#print("Hello")
```

1 We use the matrix notation:

$$\begin{bmatrix} Y_1 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ \dots & \dots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{bmatrix} \tag{1}$$

Using the OLS theory:

$$Q(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)^T (Y - X\beta) \tag{2}$$

After differetiating by β:

$$\frac{\partial Q}{\partial \beta} = -2X^T(Y - X\beta) = 0 \tag{3}$$

So the $\hat{\beta}$ is given by:

$$b = \hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{4}$$

In this case:

$$\hat{Y} = Xb = X\left(X^T X\right)^{-1} X^T Y = HY \tag{5}$$

So hat matrix is defined by:

$$H = X\left(X^T X\right)^{-1} X^T \tag{6}$$

$X^T X$ is a 2 × 2 matrix:

$$X^T X = \begin{bmatrix} 1 & X_1 \\ \dots & \dots \\ 1 & X_n \end{bmatrix}^T \begin{bmatrix} 1 & X_1 \\ \dots & \dots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \tag{7}$$

Let $\sum_{i=1}^n X_i = a$ and $\sum_{i=1}^n X_i^2 = b$ So $\left(X^T X\right)^{-1}$ is given by:

$$\left(X^T X\right)^{-1} = \frac{1}{nb - a^2} \begin{bmatrix} b & -a \\ -a & n \end{bmatrix} \tag{8}$$

We can solve H analytically in SLR model:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{k}(x_k - \bar{x}))} \tag{9}$$

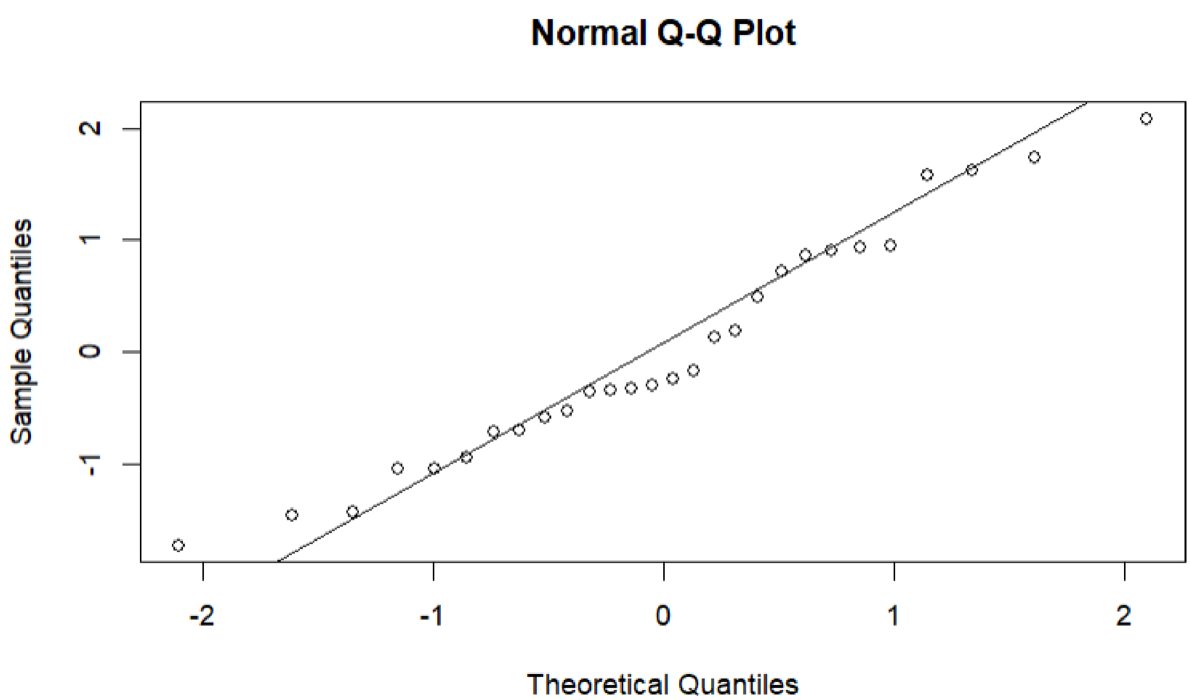
So in SLR model, a leverage point h_{ii} is given by:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2)}{\sum_{i=1}^{k}(x_k - \bar{x}))} \tag{10}$$

Question 2

```
# Your code
# 2.a
data = read.table('hw3.txt')
lm1 = lm(y ~ x8, data = data)

# 2.b
qqnorm(rstudent(lm1))
qqline(rstudent(lm1))
```




```
# 2.c
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x8         1 178.09 178.092  31.103 7.381e-06 ***
## Residuals 26 148.87   5.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 2.d
confint(lm1,"x8")
```

```
##           2.5 %       97.5 %
## x8 -0.009614347 -0.004435854
```

```r
# 2.e
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x8, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.804 -1.591 -0.647  2.032  4.580
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
## x8          -0.007025   0.001260  -5.577 7.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 26 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5272
## F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

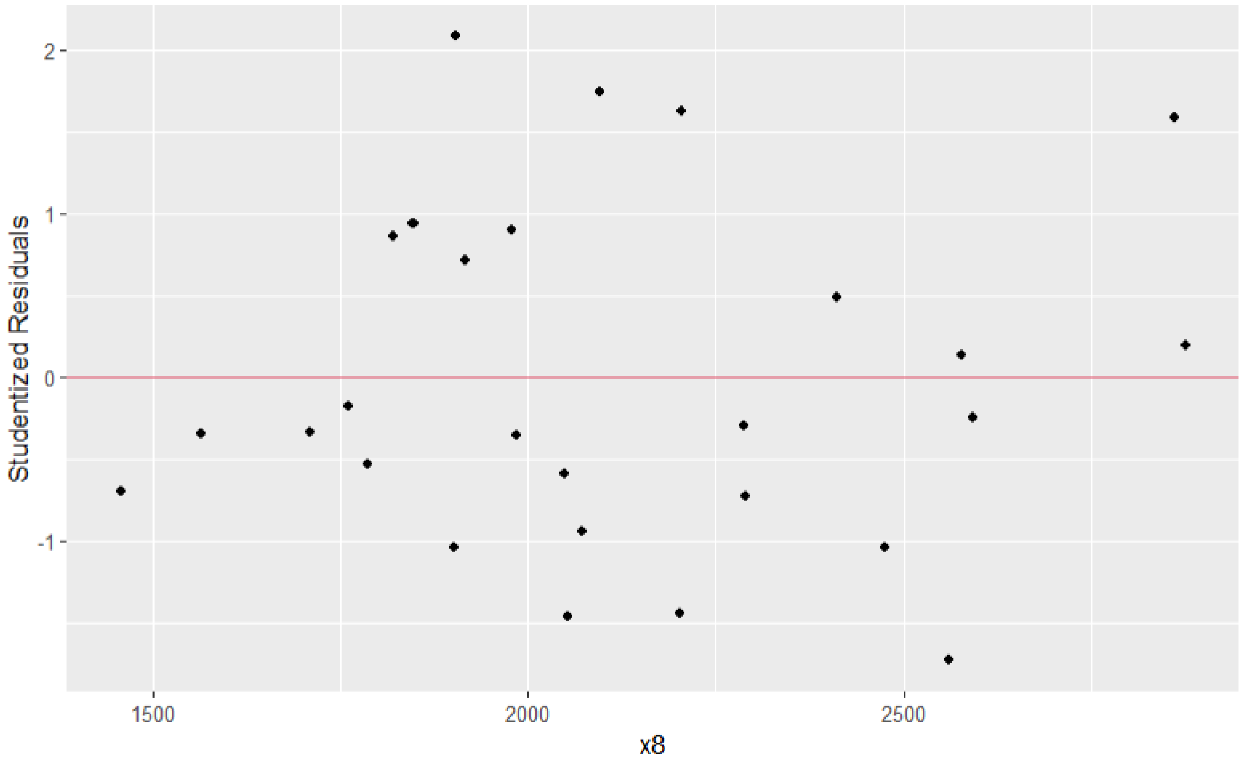
```r
# 2.f
# 95% CI:
predict(lm1, data.frame(x8=2000), interval="confidence")
```

```
##       fit      lwr      upr
## 1 7.73805 6.765753 8.710348
```

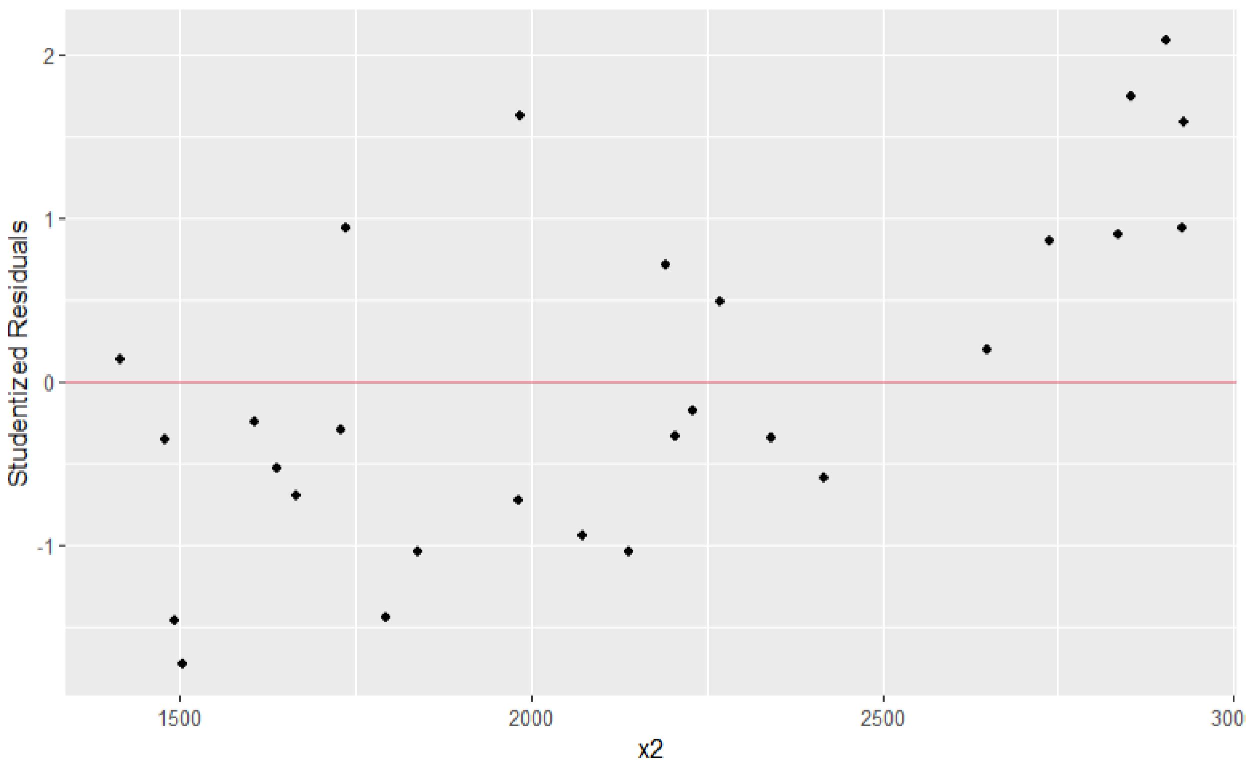
```r
# 95% PI:
predict(lm1, data.frame(x8=2000), interval="prediction")
```

```
##       fit      lwr      upr
## 1 7.73805 2.724248 12.75185
```

```r
# 2.g
ggplot(data, aes(x=x8, y=rstudent(lm1)))+geom_point()+labs(x = "x8", y= "Studentized Residuals")+geom_hl
```

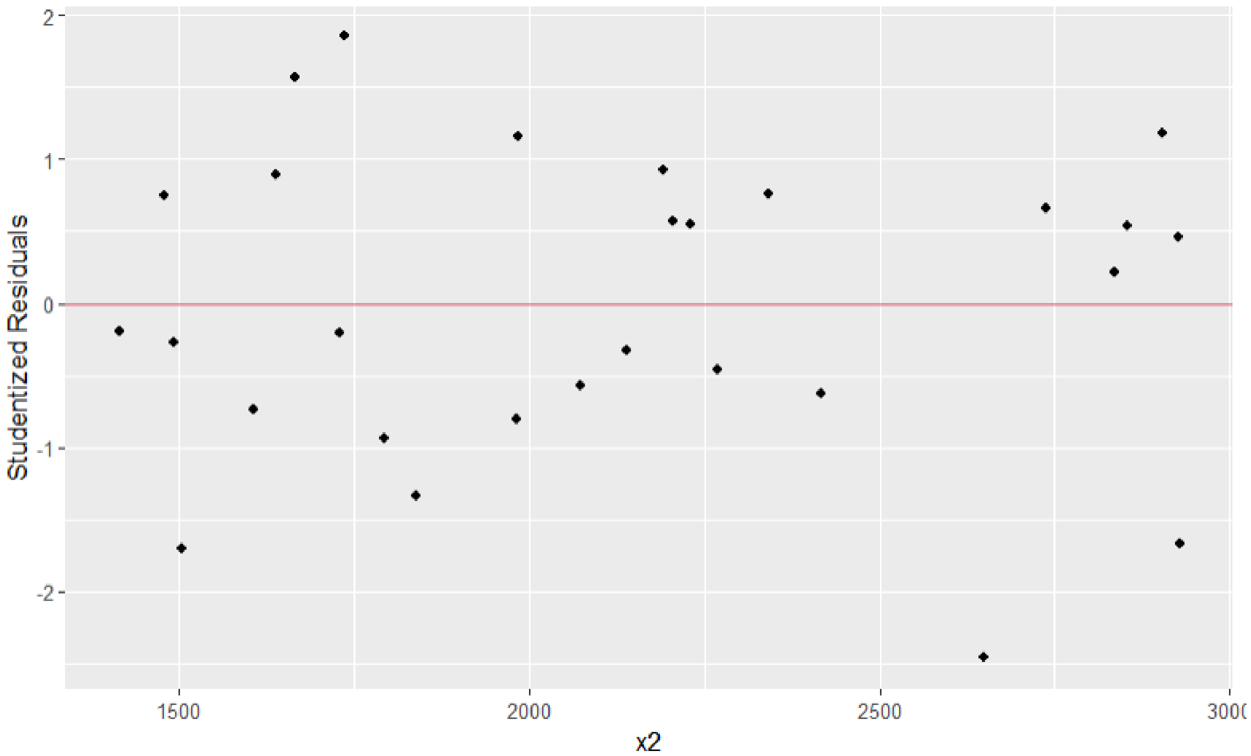



```
# 2.h
ggplot(data, aes(x=x2, y=rstudent(lm1)))+geom_point()+labs(x = "x2", y= "Studentized Residuals")+geom_hl
```

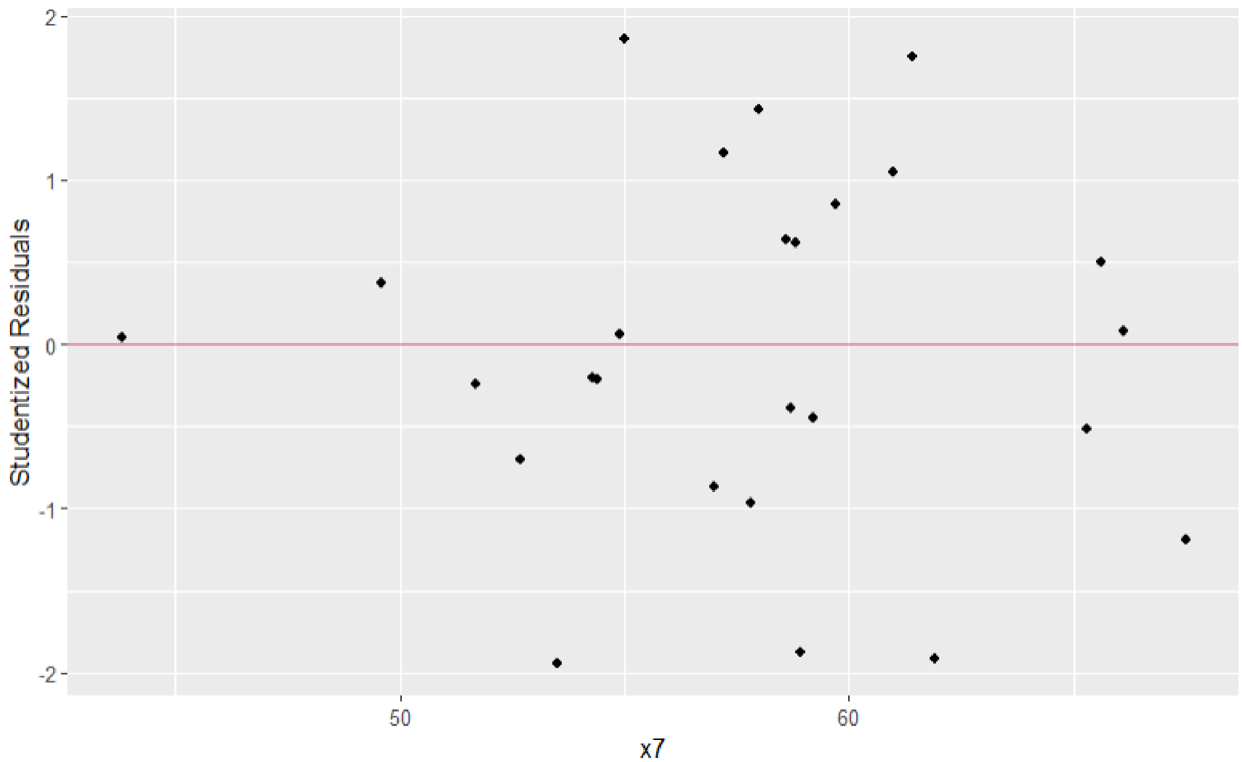



```
# 2.i
# We try to apply FS method:
```

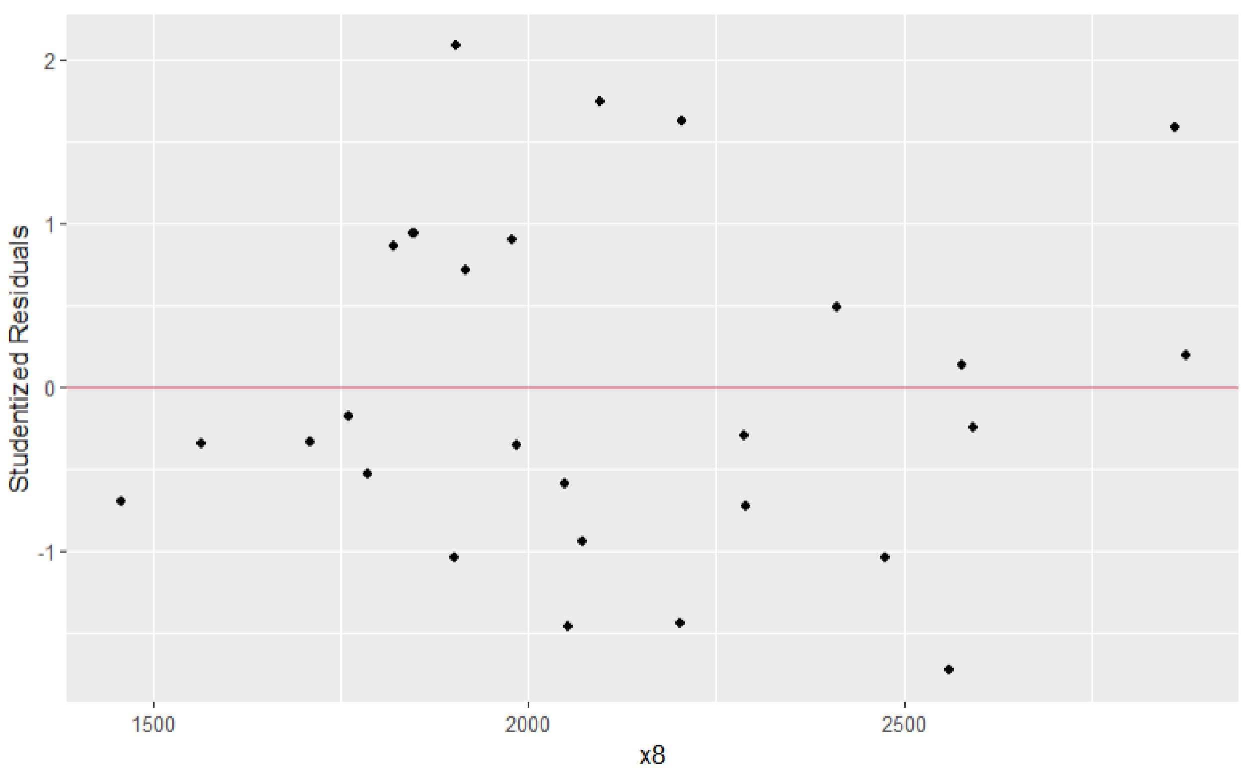
```
lm3 = lm(formula = y ~ 1, data = data)
# First iteration:
lm4 = lm(formula = y ~ x8, data = data)
# Second iteration:
lm5 = lm(formula = y ~ x2 + x8, data = data)
# Third iteration:
lm6 = lm(formula = y ~ x2 + x7 + x8, data = data)
# Verification:
lm7 = lm(formula = y ~ x2, data = data)
lm8 = lm(formula = y ~ x7, data = data)
# Component x2:
ggplot(data, aes(x=x2, y=rstudent(lm7))) + geom_point() +xlab("x2") + ylab("Studentized Residuals") +geo
```




```
# Component x7:
ggplot(data, aes(x=x7, y=rstudent(lm8))) + geom_point() +xlab("x7") + ylab("Studentized Residuals") +geo
```




```r
# Component x8
ggplot(data, aes(x=x8, y=rstudent(lm4))) + geom_point() +xlab("x8") + ylab("Studentized Residuals") +ge
```



2b. We assume that the errors are normally distributed based on the normality assumption, howeverm the Normal Q-Q Plot suggests a light-tailed data.

2c. The p-value of the regression is 7.381e-06, which is far less than 0.05. So we reject H0 and conclude that the regression is significant.

2d. The 95% confident interval of the slope is [-0.009614347,-0.004435854].

2e. As the R-squared is 0.5447, the total variability in y is explained by this model is 54.47%.

2f. The 95% confidence interval is [6.765753,8.710348], while the 95% prediction interval is [2.724248,12.75185].

2g. Interpretation: This is roughly an ideal case as the residuals are evenly distributed around 0.

2h. Yes. As the plot suggests, the linear fit of lm2 is accurate and the variance is evenly distributed. With a constant variance, the model can be improved with x2 added to the model.