HW1

Getong Zhong

2022-09-22

Problem 1

(a)

```
fit1 <- lm(CurrentWeek ~ LastWeek, data = playbill)
summary(fit1)
```

```
##
## Call:
## lm(formula = CurrentWeek ~ LastWeek, data = playbill)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36926  -7525  -2581   7782  35443
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.805e+03  9.929e+03   0.685    0.503
## LastWeek    9.821e-01  1.443e-02  68.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18010 on 16 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9963
## F-statistic:  4634 on 1 and 16 DF,  p-value: < 2.2e-16
```

```
confint(fit1)[2, ]
```

```
##     2.5 %    97.5 %
## 0.9514971 1.0126658
```

According to the result, we can be 95% confident that with one unit increase of the gross box office results for the current week (in US dollar), the increase of is the gross box office results for the previous week (in US dollar) will be between 0.95 to 1.01. Therefore 1 is a plausible value for $\beta1$.

(b)

```
t.test(playbill$LastWeek,mu=10000)
```

```
##
##  One Sample t-test
##
## data:  playbill$LastWeek
## t = 8.5797, df = 17, p-value = 1.39e-07
## alternative hypothesis: true mean is not equal to 10000
## 95 percent confidence interval:
##  471645.3 772727.9
## sample estimates:
## mean of x
##  622186.6
```

```
summary(fit1)$coef[1, 2]
```

```
## [1] 9929.318
```

```
b <- coef(fit1)[[1]]
b1 <- b + coef(fit1)[[2]] * 10000
se <- summary(fit1)$coef[1, 2]
t <- (b - b1) / se
t>qt(p=0.05, nrow(playbill) - 2, lower.tail = FALSE)
```

```
## [1] FALSE
```

According to the result, t doesn't fall into the rejection region therefore, we don't have enough evidence to reject the null hypothesis. Therefore we cannot conclude that the true mean gross box office results is significantly different from 10000.

(c)

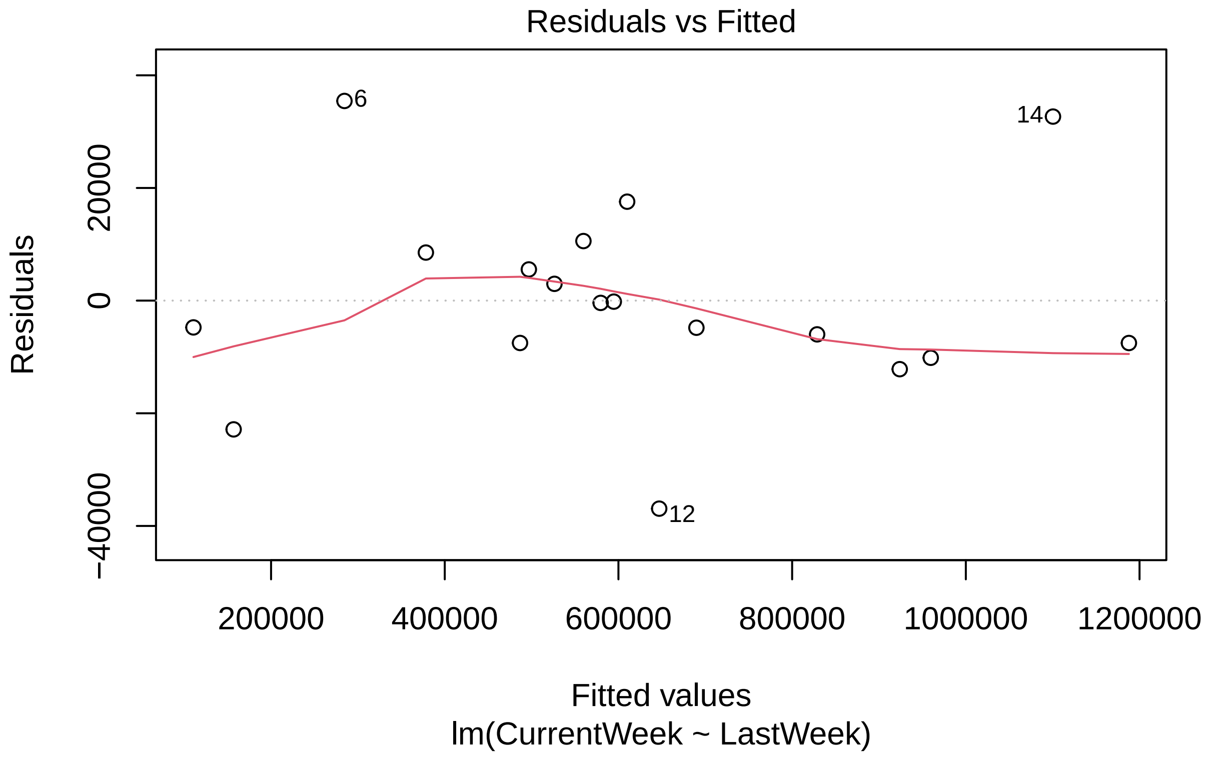
```
predict(fit1, data.frame(LastWeek = 400000), interval = "prediction")
```

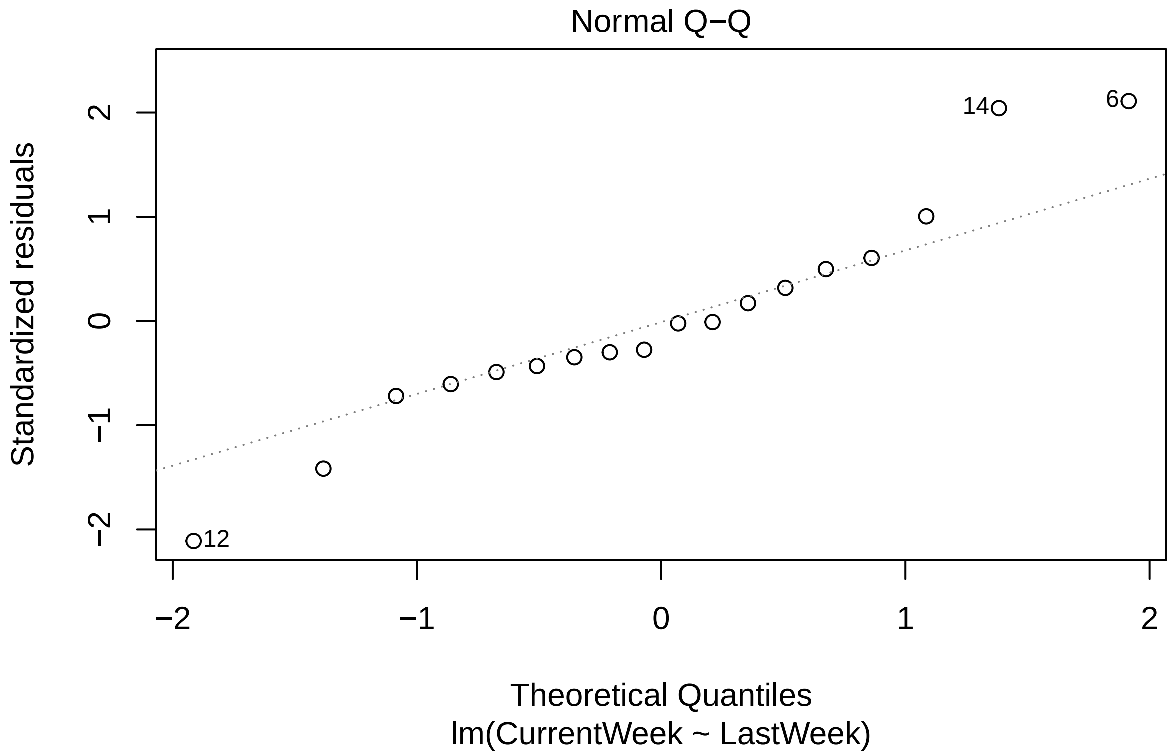
```
##        fit      lwr      upr
## 1 399637.5 359832.8 439442.2
```

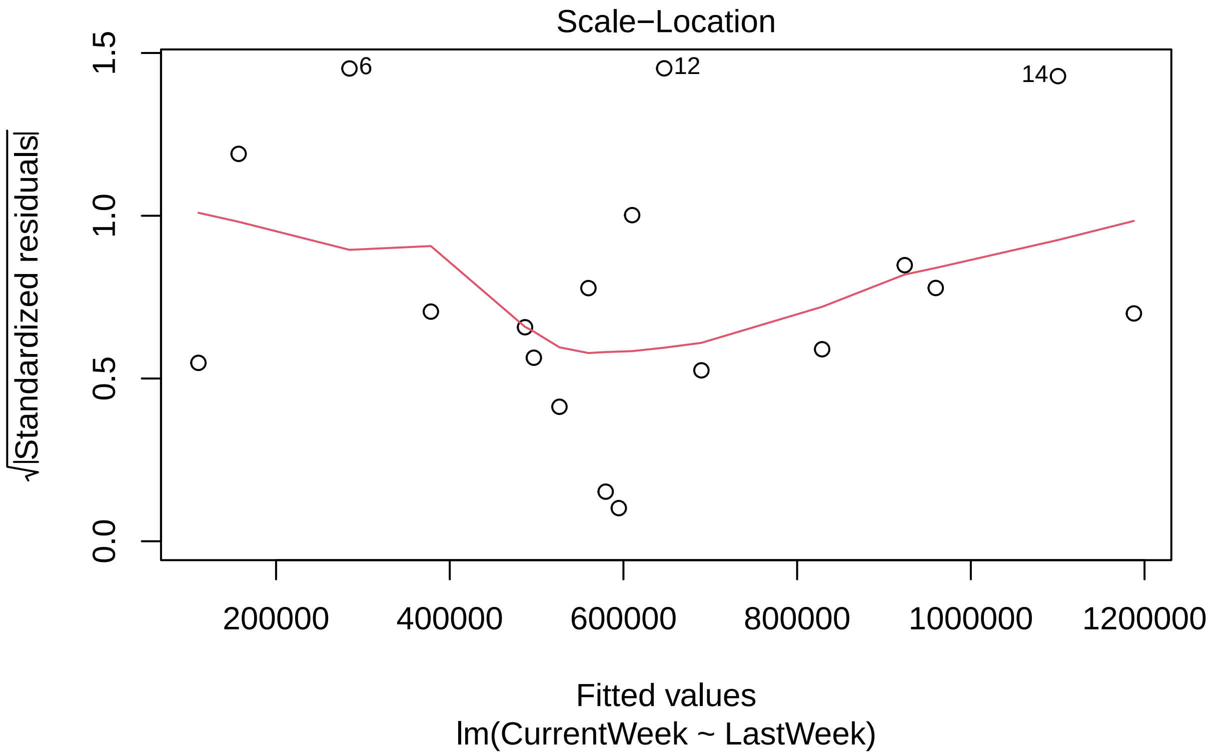
From the result we can conclude that the prediction of 450000 dollar is not a feasible prediction, for a production of 400000 dollar last week, since it fall way beyond the 95% interval of prediction from 359832.8 to 439442.2.

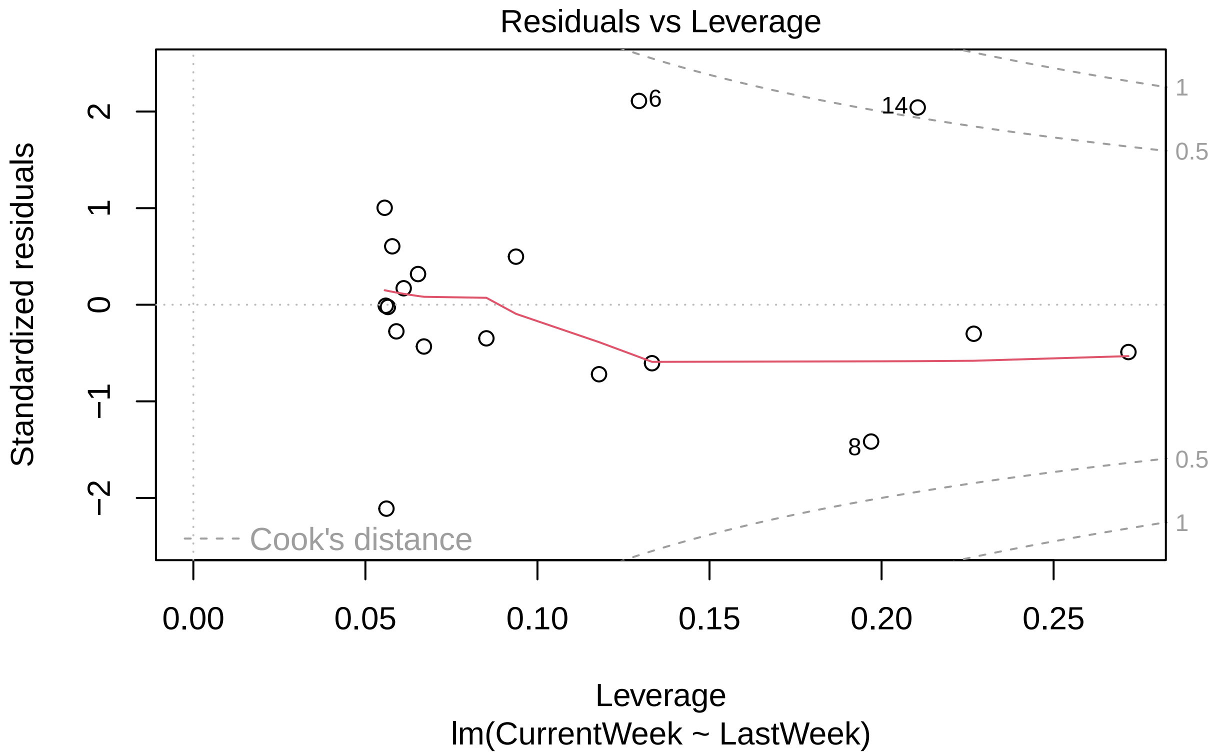
(d)

```
plot(fit1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(CurrentWeek ~ LastWeek)
20000
0
−40000
200000
400000
600000
800000
1000000
1200000
6
14
12


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(CurrentWeek ~ LastWeek)
−2
−1
0
1
2
−2
−1
0
1
2
12
14
6


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
200000
400000
600000
800000
1000000
1200000
Fitted values
lm(CurrentWeek ~ LastWeek)
6
12
14



I think it is appropriate prediction method since it seems like the given information shows there is a almost perfect equalness between this week's gross box office results and previous week's gross box office results. However, worth to mention that there still have three residuals from above plots,it might affect the accuracy of prediction in some sort of degree.

Problem 2

(a)

```
fit2<-lm(PriceChange ~ LoanPaymentsOverdue, data = indicators)
confint(fit2)[2, ]
```

```
##      2.5 %     97.5 %
## -4.1634543 -0.3335853
```

Because the 95% confidence interval for $\beta 1$ is from -4.16 to -0.33, we have enough evidence to believe that there is a significant negative linear association

(b)

```
predict(fit2, data.frame(LoanPaymentsOverdue = 4), interval = "prediction")
```

```
##         fit       lwr      upr
## 1 -4.479585 -13.13784 4.178667
```

0% is not a reasonable estimate for x=4, since the 95% confidence limit is far below 0.

Problem 3

(a)

```
fit3<-lm(Time ~ Invoices, data = invoices)
summary(fit3)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6417099  0.1222707   5.248 1.41e-05 ***
## Invoices    0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

```
b0<-coef(fit3)[[1]]
b0_se<-0.1222707
b0_t<-5.248
p3_95 <- c(b0 - 1.96 * b0_se, b0 + 1.96 * b0_se)
```

The 95% confidence interval for the start-up time is: 0.4020593, 0.8813605

(b)

```
t.test(invoices$Invoices,mu=0.01)
```

```
##
##  One Sample t-test
##
## data:  invoices$Invoices
## t = 9.5177, df = 29, p-value = 2.001e-10
## alternative hypothesis: true mean is not equal to 0.01
## 95 percent confidence interval:
##  102.0930 157.9736
## sample estimates:
## mean of x
##  130.0333
```

```
summary(fit3)$coef[1, 2]
```

```
## [1] 0.1222707
```

```
b <- coef(fit3)[[1]]
b1 <- b + coef(fit3)[[2]] * 10000
se <- summary(fit3)$coef[1, 2]
t <- (b - b1) / se
t>qt(p=0.05, nrow(invoices) - 2, lower.tail = FALSE)
```

```
## [1] FALSE
```

We don't have enough evidence to reject the null hypothesis, therefore we cannot conclude that the true mean process time is significantly different from 0.01.

(c)

```
b0<-coef(fit3)[[1]]
b1<-coef(fit3)[[2]]
rse<-0.3298
fit3
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Coefficients:
## (Intercept)     Invoices
##     0.64171      0.01129
```

```
df<-nrow(invoices) - 2
rss <- rse^2 * df
mse <- rss / nrow(invoices)
est<- b0 + b1 * 130
err <- qt(0.975, 28) * sqrt(mse) * sqrt(1 + 1 / nrow(invoices)) # since x0 = xbar
upr <- est + err
lwr <- est - err
```

The 95% prediction interval for the time taken to process 130 invoices is: $[1.4461777, 2.7730696]$.

Problem 4

(a)

We are trying to prove that

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Since

$$Y_i = \beta x_i + e_i,$$

Then

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2$$

take derivative at the both sides of the equation we get:

$$\frac{\partial}{\partial \beta} \text{RSS} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta} x_i) = 0 \iff \tag{1}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta} \sum_{i=1}^{n} x_i^2 = 0 \iff \tag{2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i}. \qquad \square$$

(b)

1.
$$E(\hat{\beta}|X) = E\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right) = \frac{\sum_{i=1}^{n} x_i E(y_i)}{\sum_{i=1}^{n} x_i^2} = \beta \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta$$

2.
$$\text{Var}(\hat{\beta}|X) = \text{Var}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right) = \frac{\sum_{i=1}^{n} x_i^2 \sigma^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

3. Since X is normal distributed, combine (1) and (2) we can get (3)

Problem 5

(d) is correct. RSS describes the sum of square of the distance between the actual value and predict value, therefore model 2 has a higher RSS since the gap between each point and the regression line is larger than in model 1; SSreg describes how well a regression model represents the modeled data, therefore model 1 has higher SSreg since the points on model 1 more fitted the regression line.

Problem 6

(a)

$$y_i - \hat{y}_i = (y_i - \hat{y}_i) - \bar{y} + \bar{y} = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) = (y_i - \bar{y}) - (\hat{\beta} x_i - \hat{\beta} \bar{x}) = (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})$$

(b)

$$\hat{y}_i - \bar{y} = \hat{\beta} x_i - \hat{\beta} \bar{x} = \hat{\beta}(x_i - \bar{x})$$

(c)

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\hat{\beta}_1(x_i - \bar{x}) =$$

$$\hat{\beta}_1\left(\sum_{i=1}^{n} y_i(x_i\bar{x}) - \hat{\beta}_0\sum_{i=1}^{n} x_i - \bar{x} - \hat{\beta}_1\sum_{i=1}^{n} x_i(x_i - \bar{x})\right) =$$

$$\hat{\beta}_1(\mathrm{SXY} - 0 - \hat{\beta}_1\mathrm{SXX}) =$$

$$\hat{\beta}_1(SXY - SXY) = 0$$