Fall 2022 STAT 706 HW 1

Reina Li

Chapter 2.8 : 1a-d, 2a-b, 3a-c, 4a-b, 5, 6a-c

1.

```
# Import data set-----
df_1a <- read.csv("playbill.csv")

# Fit the model to the data-----
model_1a <- lm(CurrentWeek ~ LastWeek,
               data = df_1a)
summary(model_1a)
```

```
##
## Call:
## lm(formula = CurrentWeek ~ LastWeek, data = df_1a)
##
## 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
```

(a)

```
alpha <- 0.05

# method 1
n <- dim(df_1a)[1]
xbar <- mean(df_1a$LastWeek)
ybar <- mean(df_1a$CurrentWeek)
```

```r
sxy <- 0
for (i in 1:n){
  a <- df_1a$LastWeek[i]-xbar
  b <- df_1a$CurrentWeek[i]-ybar
  sxy <- sxy + (a*b)
}

sxx <- 0
for (i in 1:n) {
  a <- df_1a$LastWeek[i]-xbar
  sxx <- sxx + (a^2)
}

bhat_1 <- sxy/sxx

s_squared <- 0
for (i in 1:n) {
  s_squared <- s_squared + (model_1a$residuals[i])^2
}
s_squared <- (1 / (n - 2)) * s_squared
s <- sqrt(s_squared)

#same as summary(model_1a)$coefficients[2,2]
standarderr <- s/sqrt(sxx)

tval <- qt(p = alpha/2, df = n - 2, lower.tail = FALSE)

lowerconf <- bhat_1 - tval * standarderr
lowerconf
```

```
##         1
## 0.9514971
```

```r
upperconf <- bhat_1 + tval * standarderr
upperconf
```

```
##        1
## 1.012666
```

```r
# method 2
confint(model_1a, level = 0.95)[2,]
```

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

Two methods were used to compute the confidence interval for β_1. Method 1 uses the formulas found in the textbook, while method 2 uses a built-in R function. They return the same values for the confidence interval. The confidence interval is (0.9514971, 1.0126658).

1 is a plausible value for β_1 because 1 is within the 95% confidence interval.

(b)

```
# H0 : beta_0 = 10000 against HA : beta_0 != 10000

h_0 <- 10000
h_obs <- summary(model_1a)$coefficients[1,1]
h_obs_standarderr <- summary(model_1a)$coefficients[1,2]

test_stat <- (h_obs - h_0) / h_obs_standarderr
p_val <- 2 * pt(q = test_stat, df = n - 2, lower.tail = FALSE)
p_val
```

```
## [1] 1.248219
```

```
p_val < alpha
```

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

Since the p-value, 1.2482193, is not less than 0.05, we fail to reject the null hypothesis.

(c)

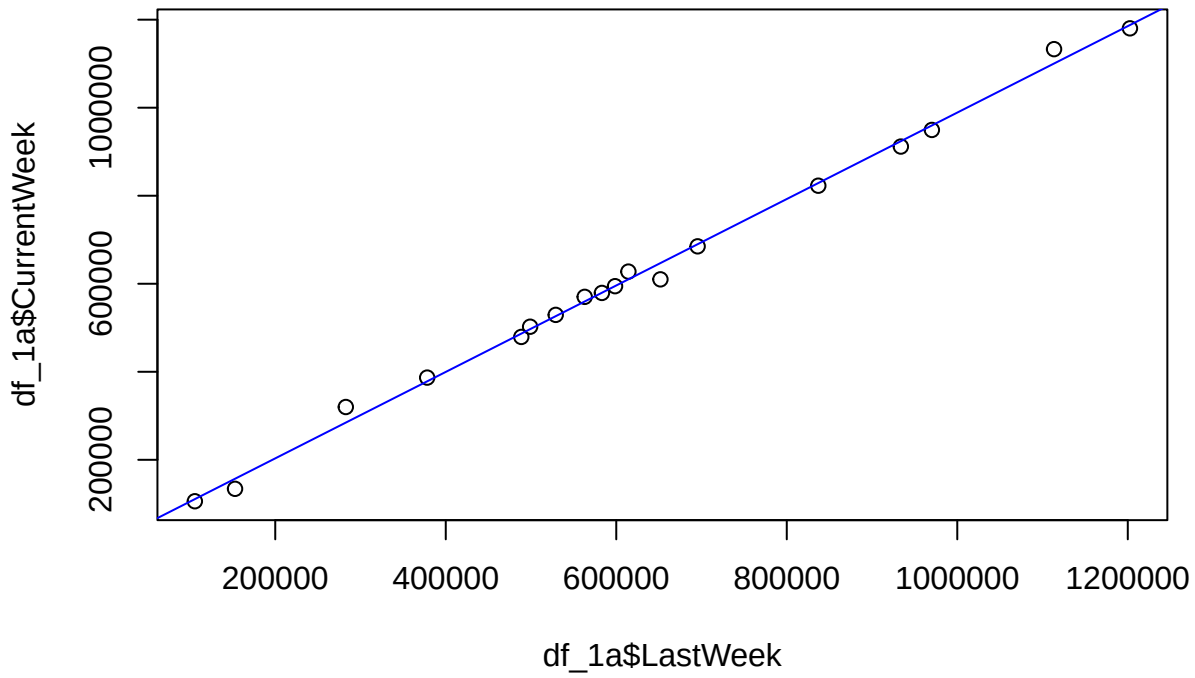
```
predict(model_1a, data.frame(LastWeek = 400000), interval = "prediction", level = 0.95)
```

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

Using the `predict()` function, a 95% prediction interval was calculated for the gross box office results for the current week for a production with \$400,000 in gross box office the previous week. The prediction interval is $(3.5983275 \times 10^5, 4.394422 \times 10^5)$. Since \$450,000 is greater than the upper bound of the prediction interval, \$450,000 is not a feasible value for the gross box office results in the week.

(d)

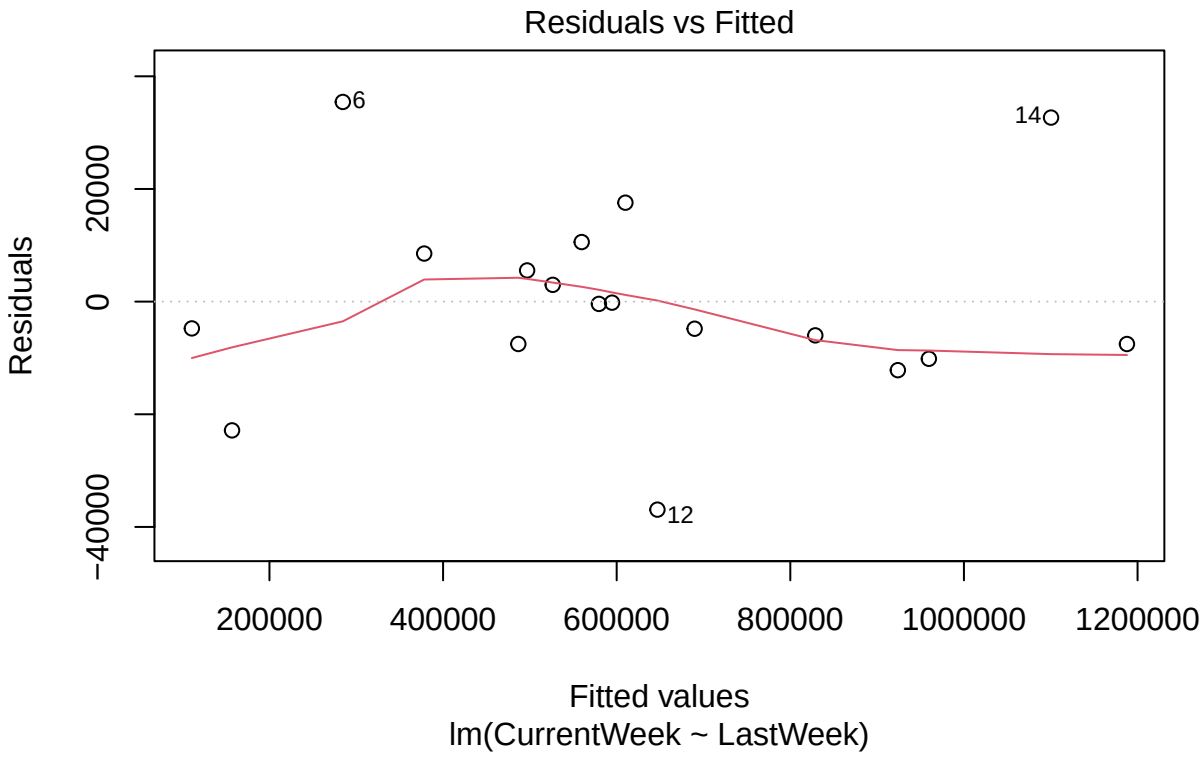
```
plot(df_1a$LastWeek, df_1a$CurrentWeek)
abline(model_1a, col = "blue")
```

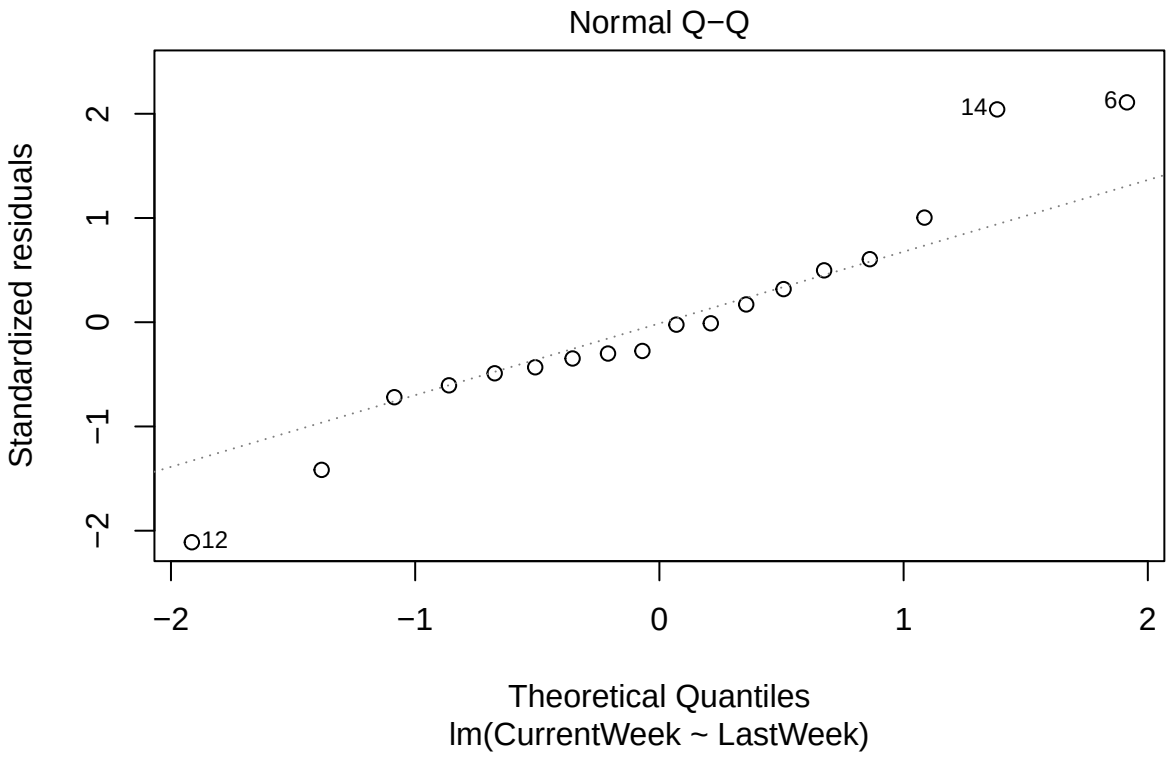


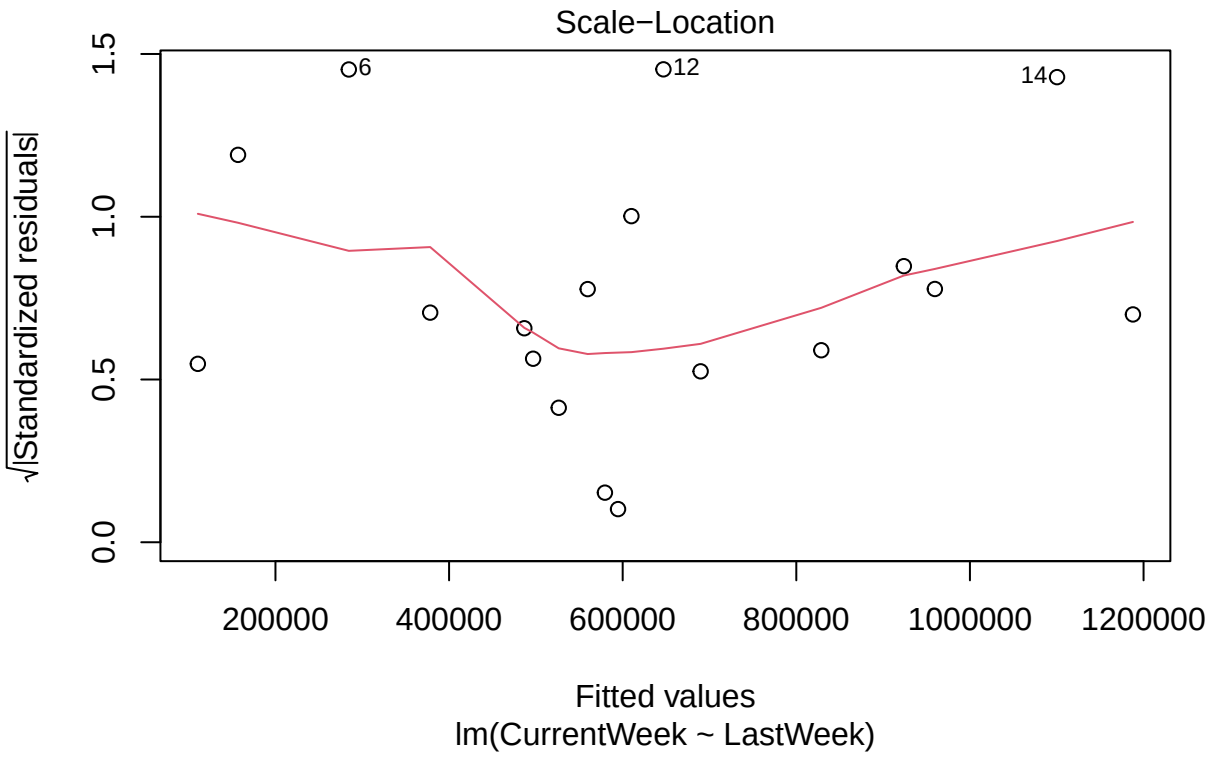

```
summary(model_1a)$r.squared
```

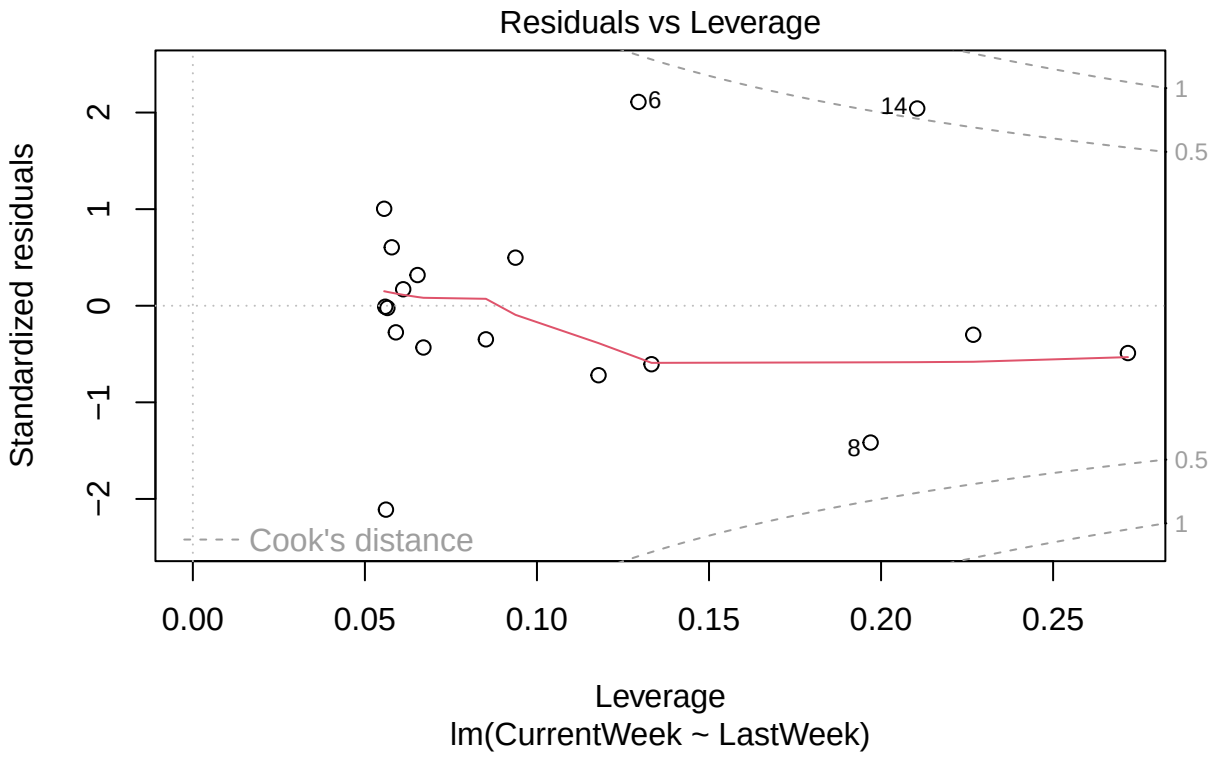
```
## [1] 0.9965589
```

```
plot(model_1a)
```


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


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


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



First, looking at the first plot of the data with a regression line (blue line), there is evidence that we should be doing a simple linear regression on the model. The R^2 value, 0.9965589, is also very close to 1, which means that the model explains almost all the variation in the response variable around its mean. When we look at the Residuals vs Fitted plot, we can see that only 3 values were not predicted well, because they are far away from the line at $x = 0$. Given all this information from the plots, this prediction rule is appropriate.

2.

```
# Import data set-----
df_2a <- read.table("indicators.txt", header = TRUE, sep = "", dec = ".")

# Fit the model to the data-----
model_2a <- lm(PriceChange ~ LoanPaymentsOverdue,
               data = df_2a)
summary(model_2a)
```

```
##
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = df_2a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.6541 -3.3419 -0.6944  2.5288  6.9163
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.5145     3.3240   1.358   0.1933
## LoanPaymentsOverdue  -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

(a)

```
confint(model_2a, level = 0.95)[2,]
```

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

The confidence interval for β_1 is (-4.1634543, -0.3335853).

On the basis of this confidence interval, there is evidence of a significant negative linear association.

(b)

```
predict(model_2a, data.frame(LoanPaymentsOverdue = 4), interval = "confidence", level = 0.95)
```

```
##         fit       lwr       upr
## 1 -4.479585 -6.648849 -2.310322
```

The confidence interval for $\mathrm{E}(Y|X = 4)$ is (-6.6488492, -2.3103215).

0 is not a feasible value for $\mathrm{E}(Y|X = 4)$ because 0 is greater than the upper bound of the confidence interval, so it is outside of the interval.

3.

```
# Import data set-----
df_3a <- read.table("invoices.txt", header = TRUE, sep = "", dec = ".")

# Fit the model to the data-----
model_3a <- lm(Time ~ Invoices,
               data = df_3a)
summary(model_3a)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = df_3a)
##
## 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
```

(a)

```
bhat_0 <- summary(model_3a)$coefficients[1,1]
standarderr <- summary(model_3a)$coefficients[1,2]
lowerconf <- bhat_0 - 1.96 * standarderr
lowerconf
```

```
## [1] 0.4020592
```

```
upperconf <- bhat_0 + 1.96 * standarderr
upperconf
```

```
## [1] 0.8813605
```

The confidence interval for the start-up time, β_0 is (0.4020592, 0.8813605).

(b)

```
# H0 : beta_1 = 0.01 against HA : beta_1 != 0.01

h_0 <- 0.01
h_obs <- summary(model_3a)$coefficients[2,1]
h_obs_standarderr <- summary(model_3a)$coefficients[2,2]

test_stat <- (h_obs - h_0) / h_obs_standarderr
p_val <- 2 * pt(q = test_stat, df = n - 2, lower.tail = FALSE)
p_val
```

```
## [1] 0.134072
```

```
p_val < alpha
```

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

Since the p-value, 0.134072, is not less than 0.05, we fail to reject the null hypothesis.

(c)

```
n <- dim(df_3a)[1]
bhat_0 <- summary(model_3a)$coefficients[1,1]
bhat_1 <- summary(model_3a)$coefficients[2,1]
rse <- summary(model_3a)$sigma
rss <- rse ^ 2 * (n - 2)
mse <- rss / n

#point estimate
ptestimate <- bhat_0 + bhat_1 * 130
ptestimate
```

```
## [1] 2.109624
```

```
lowerpred <- ptestimate - (qt(1 - (alpha / 2), n - 2) * sqrt(mse) * sqrt(1 + (1 / n)))
lowerpred
```

```
## [1] 1.446231
```

```
upperpred <- ptestimate + (qt(1 - (alpha / 2), n - 2) * sqrt(mse) * sqrt(1 + (1 / n)))
upperpred
```

```
## [1] 2.773016
```

A point estimate for the time taken to process 130 invoices is 2.1096236 hours.

The prediction interval is (1.4462313, 2.7730159).

4.

(a)

Using assumption (1) : Y is related to x by the simple linear regression model $Y_i = \beta x_i + e_i (i = 1, 2, ..., n)$ and using $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

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

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

$$\frac{\partial}{\partial\beta}\left(\sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2\right)$$

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

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

$$\sum_{i=1}^{n}x_i(y_i - \hat{\beta}x_i) = 0$$

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

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

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

$$\hat{\beta}\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_iy_i$$

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

(b)

(i) From the question, we know that $\hat{\beta} = \frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}$

$$E(\hat{\beta}|X)$$

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

$$= E\left(\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}\right)$$

$$= \frac{E(\sum_{i=1}^{n} x_i y_i)}{E(\sum_{i=1}^{n} x_i^2)}$$

$$= \frac{\sum_{i=1}^{n} x_i E(y_i)}{\sum_{i=1}^{n} x_i^2}$$

From assumption (1), $Y_i = \beta x_i + e_i (i = 1, 2, ..., n)$, then $E(y_i) = \beta x_i$

$$= \frac{\sum_{i=1}^{n} x_i * \beta x_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \beta \frac{\sum_{i=1}^{n} x_i * x_i}{\sum_{i=1}^{n} x_i^2}$$

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

$$= \beta$$

(ii) From the question, we know that $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

$$Var(\hat{\beta}|X)$$

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

$$= Var\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right)$$

$$= \frac{Var(\sum_{i=1}^{n} x_i y_i)}{Var(\sum_{i=1}^{n} x_i^2)}$$

$$= \frac{\sum_{i=1}^{n} x_i Var(y_i)}{(\sum_{i=1}^{n} x_i^2)^2}$$

From assumption (1), $Y_i = \beta x_i + e_i (i = 1, 2, ..., n)$, then $Var(y_i) = \sigma^2 x_i$

$$= \frac{\sum_{i=1}^{n} x_i * \sigma^2 x_i}{(\sum_{i=1}^{n} x_i^2)^2}$$

$$= \sigma^2 \frac{\sum_{i=1}^{n} x_i * x_i}{(\sum_{i=1}^{n} x_i^2)^2}$$

$$= \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x_i^2)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

(iii) Under assumption (4): The errors are normally distributed with a mean of 0 and variance σ^2 (especially when the sample size is small), then $\hat{\beta}|X$ is normally distributed.

5.

Looking at the plots for Model 1 and Model 2, the lines represent the least squares regression lines. In Model 1, the coordinated pair points are closer to the least squares regression line. In Model 2, the coordinated pair points are far away from the least squares regression line. This means that RSS for Model1 is less than the RSS for Model 2. The SSreg for Model 1 must be greater than the SSreg for Model 2 because most of the sum of squares of Model 1 is explained by the variance.

Therefore, (d) is correct: RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

6.

(a)

$$(y_i - \hat{y}_i)$$

$$= (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

Assuming that $\beta_1 \neq 0$, then $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i$ and $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

$$= (y_i - \bar{y}) - (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})$$

$$= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$$

(b)

From the question,

$$(y_i - \hat{y}_i)$$

$$= (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

$$= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$$

then

$$(\hat{y}_i - \bar{y}) = \hat{\beta}_1 (x_i - \bar{x})$$

(c)

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Since $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\hat{x}_i$ and $(\hat{y}_i - \bar{y}) = \hat{\beta}_1(x_i - \bar{x})$

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

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

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

$$= \hat{\beta}_1 \sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - \hat{\beta}_0 x_i - \hat{\beta}_0\bar{x} - \hat{\beta}_1 x_i^2 - \hat{\beta}_1 x_i\bar{x})$$

$$= \hat{\beta}_1\bigg(\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i) - \sum_{i=1}^{n}(\hat{\beta}_0 x_i - \hat{\beta}_0\bar{x}) - \sum_{i=1}^{n}(\hat{\beta}_1 x_i^2 - \hat{\beta}_1 x_i\bar{x})\bigg)$$

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

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

Since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, then

$$= \hat{\beta}_1\bigg(\sum_{i=1}^{n} y_i(x_i - \bar{x}) - 0 - \hat{\beta}_1\sum_{i=1}^{n} x_i(x_i - \bar{x})\bigg)$$

And since $\sum_{i=1}^{n} x_i(x_i - \bar{x}) = SXX$ and $\sum_{i=1}^{n} y_i(x_i - \bar{x}) = SXy$, then

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

And given in the question, $\hat{\beta}_1 = \dfrac{SXY}{SXX}$, then

$$= \hat{\beta}_1\bigg(SXY - \frac{SXY}{SXX}SXX\bigg)$$

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

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

$$= 0$$