431 assignment1

Yiming Shen 20891774

28/09/2023

Question 1

```
library(ALSM)
```

```
## Loading required package: leaps

## Loading required package: SuppDists

## Loading required package: car

## Loading required package: carData
```

```
data(GroceryRetailer)
str(GroceryRetailer)
```

```
## 'data.frame':    52 obs. of  4 variables:
##  $ y : int  4264 4496 4317 4292 4945 4325 4110 4111 4161 4560 ...
##  $ x1: int  305657 328476 317164 366745 265518 301995 269334 267631 296350 277223 ...
##  $ x2: num  7.17 6.2 4.61 7.02 8.61 6.88 7.23 6.27 6.49 6.37 ...
##  $ x3: int  0 0 0 0 1 0 0 0 0 0 ...
```

(a)

```
model <- lm(y ~ x1 + x2 + x3, data = GroceryRetailer)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = GroceryRetailer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -264.05 -110.73  -22.52   79.29  295.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.150e+03  1.956e+02  21.220  < 2e-16 ***
## x1           7.871e-04  3.646e-04   2.159   0.0359 *
## x2          -1.317e+01  2.309e+01  -0.570   0.5712
## x3           6.236e+02  6.264e+01   9.954 2.94e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.3 on 48 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6689
## F-statistic: 35.34 on 3 and 48 DF,  p-value: 3.316e-12
```

```
confint(model)
```

```
##                     2.5 %       97.5 %
## (Intercept)  3.756677e+03 4.543098e+03
## x1           5.409544e-05 1.520065e-03
## x2          -5.959506e+01 3.326302e+01
## x3           4.976064e+02 7.495025e+02
```

Based on the ouput, we find the estimated regression equation is: y_hat = 4150 + 0.0007871*x1 + (-13.17)*x2 + 623.6*x3

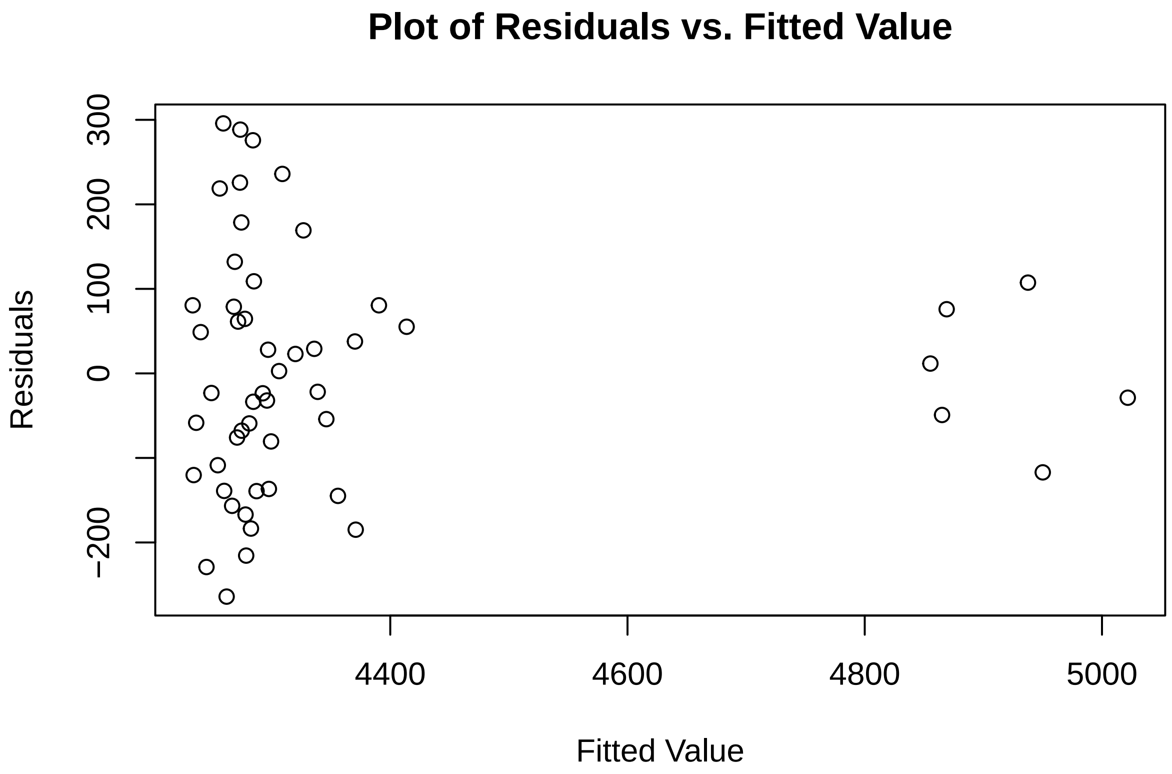
Interpretation: beta1_hat = 0.0007871: when other variables are kept unchanged, as the number of cases shipped increase one, the average of total labour hours is estimated to increase 0.0007871 hours. 95% C.I. for beta1_hat is [5.409544e-05, 0.001520065]

beta2_hat = -13.17: when other variables are kept unchanged, as the indirect costs of the total labour hours as a percentage increase one unit, the average of total labour hours is estimated to decrease 13.17 hours. 95% C.I. for beta2_hat is [-59.59506, 33.26302]

beta3_hat = 623.6: when other variables are kept unchanged, compared with a week without holiday, the average of total labour hours in the week with a holiday is estimated to increase 623.6 hourss. 95% C.I. for beta3_hat is [497.6064, 749.5025]

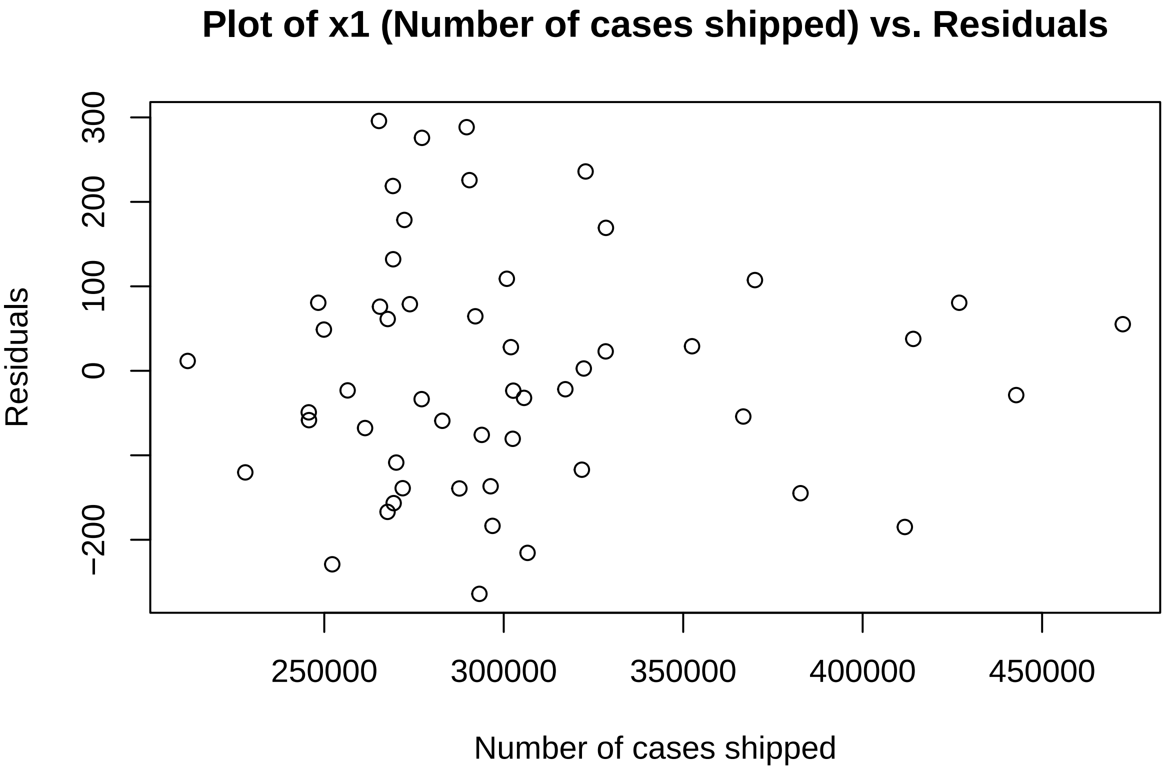
(b)

```
# method 1
plot(fitted(model), residuals(model), xlab="Fitted Value", ylab="Residuals",
     main = "Plot of Residuals vs. Fitted Value")
```

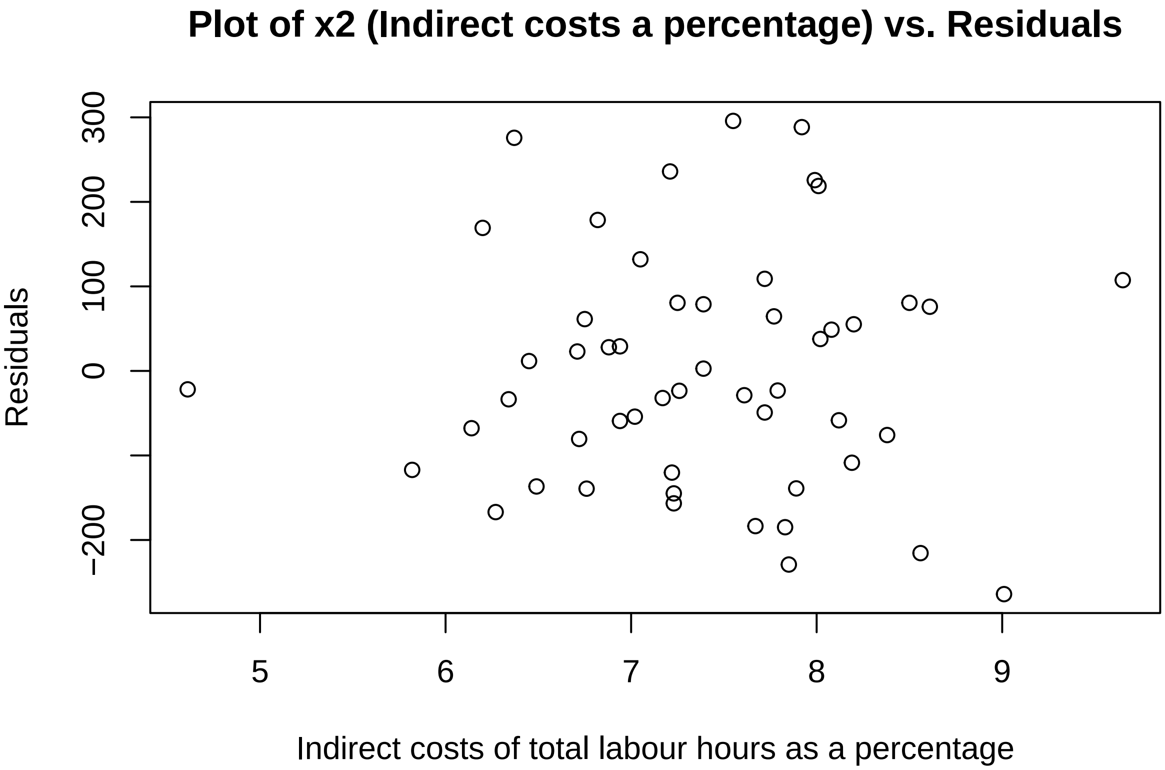


Based on the plot, we find that almost all points are spread within a constant band, except a few outliers, which shows the constant variance of model assumptions.

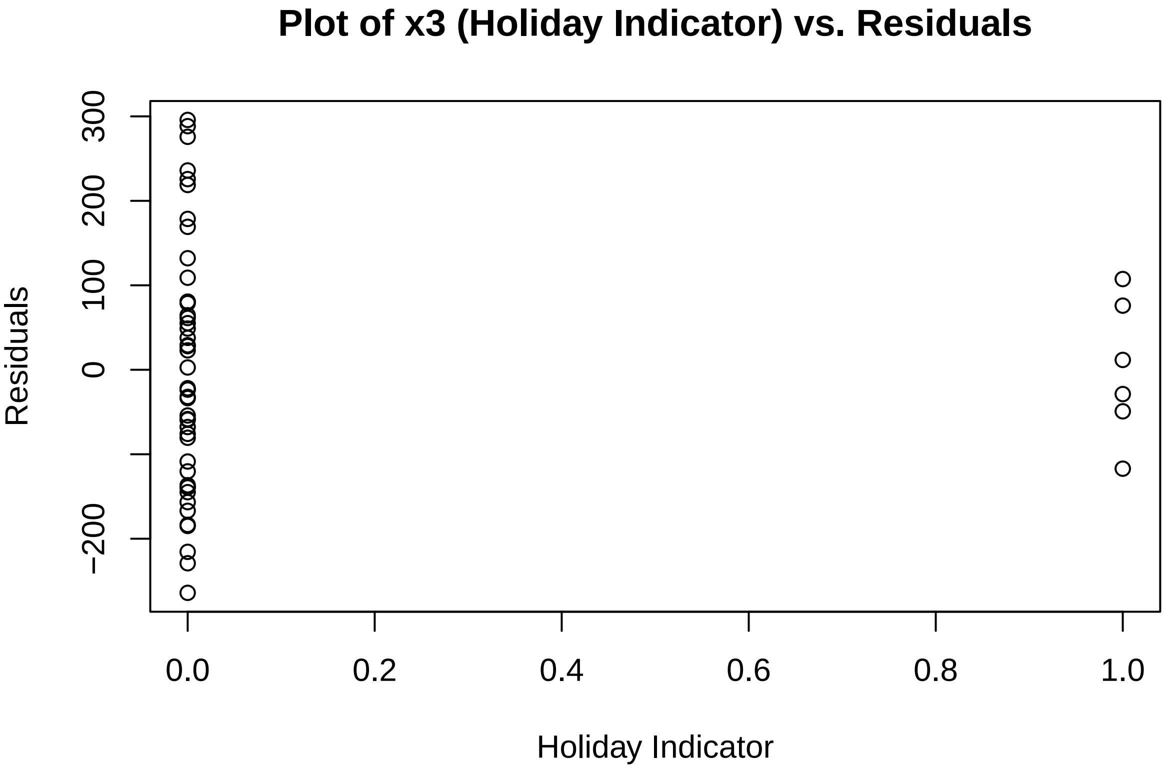
```
# method 2
plot(GroceryRetailer$x1, residuals(model), xlab="Number of cases shipped",
     ylab="Residuals",
     main="Plot of x1 (Number of cases shipped) vs. Residuals")
```



```
plot(GroceryRetailer$x2, residuals(model),
     xlab="Indirect costs of total labour hours as a percentage",
     ylab="Residuals",
     main="Plot of x2 (Indirect costs a percentage) vs. Residuals")
```

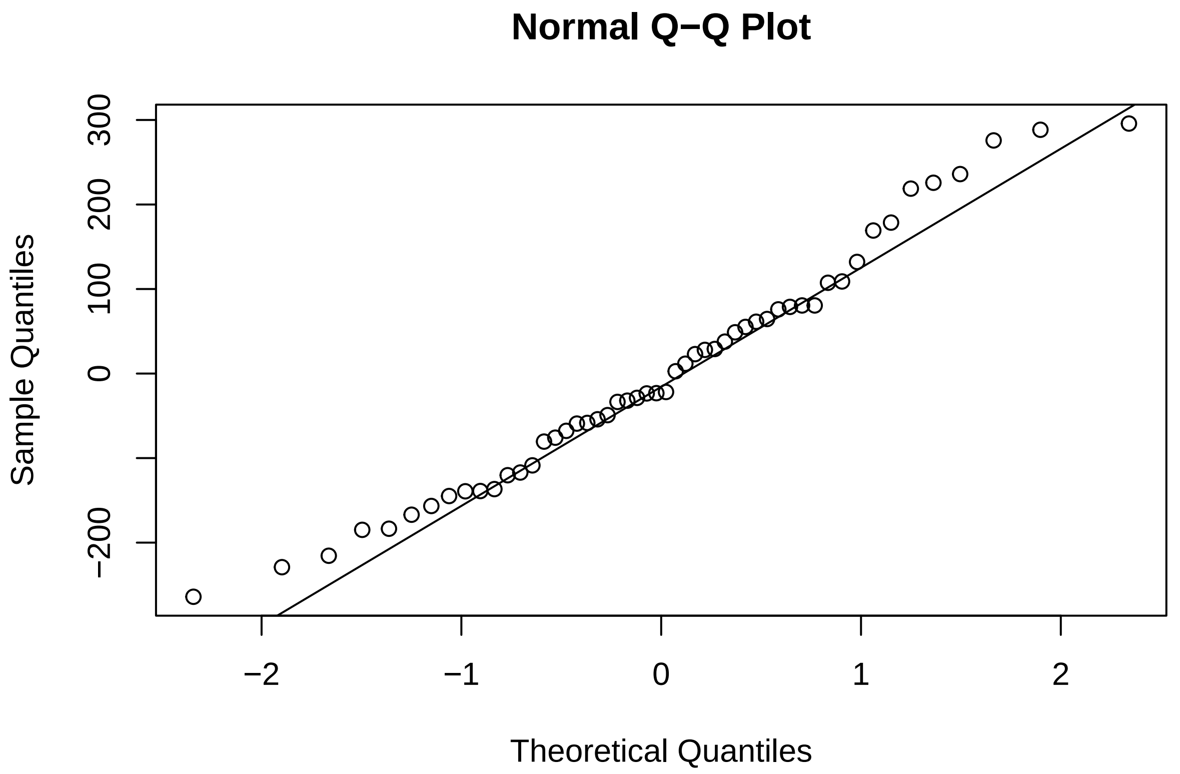


```
plot(GroceryRetailer$x3, residuals(model), xlab="Holiday Indicator",
     ylab="Residuals",
     main="Plot of x3 (Holiday Indicator) vs. Residuals")
```



Based on the plot, we find that there exists some linear relationship between x1 and residuals and same for x2 and x3, which shows the linearity of model assumptions.

```
# method 3
qqnorm(residuals(model))
qqline(residuals(model))
```



Based on the QQplot, we find that all points are distributed along the line, which shows the normality of model assumptions.

In conclusion, based on 3 methods, we find that all assumptions appear to be met.

(c)

```
mean1 <- mean(GroceryRetailer$x1)
mean2 <- mean(GroceryRetailer$x2)
# since no holiday, so x3=0
predict(model,newdata = data.frame(x1=mean1,x2=mean2,x3=0),
        interval = "confidence", level = 0.95)
```

```
##       fit      lwr      upr
## 1 4291.09 4248.576 4333.603
```

Based on the output, we find that the estimate of the mean total labour hours is 4291.09 hours. And the corresponding 95% C.I. is [4248.576, 4333.603]

(d)

```
model2 <- lm(y ~ x1*x3 + x2*x3, data = GroceryRetailer)
# ANOVA test
anova(model,model2)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x1 * x3 + x2 * x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     48 985530
## 2     46 951129  2     34400 0.8319 0.4417
```

Based on the output, we find the p-value of ANOVA test is 0.4417 Since the significance level is 0.05 p-value > 0.05, we do not reject null hypothesis, which shows that there is no association.

Question 2

(c)

```
# when 1000 sample size , n=20, pie = 0.5
result <- c()
for (i in 1:1000){
n <- 20
y <- rbinom(n, size=1, prob=0.5)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

```
## [1] 0.044
```

```
# when n = 100
result <- c()
for (i in 1:1000){
n <- 100
y <- rbinom(n, size=1, prob=0.5)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

```
## [1] 0.06
```

```
# when n = 1000
result <- c()
for (i in 1:1000){
n <- 1000
y <- rbinom(n, size=1, prob=0.5)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

```
## [1] 0.055
```

Comment: After simulating at n=20;100;1000, I find that with the increase of size n (from n=20 to n=100 to n=1000), the estimate of probability of rejecting the null hypothesis get closer to level 0.05

(d)

```
# when 1000 sample size , n=20, pie = 0.6
result <- c()
for (i in 1:1000){
n <- 20
y <- rbinom(n, size=1, prob=0.6)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                 sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

```
## [1] 0.146
```

```
# when n = 100, pie = 0.6
result <- c()
for (i in 1:1000){
n <- 100
y <- rbinom(n, size=1, prob=0.6)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                 sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

```
## [1] 0.53
```

```
# when n=1000, pie = 0.6
result <- c()
for (i in 1:1000){
n <- 1000
y <- rbinom(n, size=1, prob=0.6)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                 sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

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

```
# when 1000 sample size , n=20, pie = 0.8
result <- c()
for (i in 1:1000){
n <- 20
y <- rbinom(n, size=1, prob=0.8)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                 sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

```
## [1] NA
```

```
# when n = 100, pie = 0.8
result <- c()
for (i in 1:1000){
n <- 100
y <- rbinom(n, size=1, prob=0.8)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                 sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
```

```r
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

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

```r
# when n=1000, pie = 0.8
result <- c()
for (i in 1:1000){
n <- 1000
y <- rbinom(n, size=1, prob=0.8)
likelihood_ratio_stat <- (-2)*(sum(y)*log(0.5)+(n-sum(y))*log(1-0.5)-
                                 sum(y)*log(mean(y))-(n-sum(y))*log(1-mean(y)))
p_value <- 1 - pchisq(likelihood_ratio_stat,1)
result[i] <- p_value < 0.05
}
mean(result)
```

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

Table and Summary: please see the attach separate PDF