

ACCT648 Applied Statistics for Data Analysis

Assignment 2 Answers

R Codes of Q1

```
#Assignment 2 Q1

dataset1 <- read.csv("Salary2021A.csv", stringsAsFactors = TRUE)

summary(dataset1)

salary <- dataset1$Salary

gpa <- dataset1$GPA

gg1.fit <- lm(salary~gpa)

plot(gpa, salary, main="Relationship between starting monthly salary and
      the grade point average", xlab="GPA", ylab="monthly salary")

abline(gg1.fit, lwd=3, col="red")

summary(gg1.fit)

#assumption check

plot(gpa, residuals(gg1.fit), main="Relationship between residuals and
      the GPA",
      xlab="GPA", ylab="Residuals")

library(fitdistrplus)

fnorm1 <- fitdist(residuals(gg1.fit), "norm")

summary(fnorm1)

plot(fnorm1)

#prediction and estimation

confint(gg1.fit, level=0.95)

predict(gg1.fit, data.frame(gpa=3.3), interval="confidence", level=0.95)

predict(gg1.fit, data.frame(gpa=3.2), interval="prediction", level=0.95)

gg2.fit <- lm(salary~gpa+l(gpa^2))

summary(gg2.fit)
```

R Output of Q1

```
> dataset1 <- read.csv("Salary2021A.csv", stringsAsFactors = TRUE)
> summary(dataset1)
      GPA      Salary
Min.   :2.210  Min.   :2390
1st Qu.:2.632  1st Qu.:2675
Median :3.245  Median :3130
Mean   :3.120  Mean   :3014
3rd Qu.:3.540  3rd Qu.:3360
Max.   :3.860  Max.   :3530
> salary <- dataset1$Salary
> gpa <- dataset1$GPA
> gg1.fit <- lm(salary~gpa)
> plot(gpa, salary, main="Relationship between starting monthly salary and
+      the grade point average", xlab="GPA", ylab="monthly salary")
> abline(gg1.fit, lwd=3, col="red")
> summary(gg1.fit)
```

```
Call:
lm(formula = salary ~ gpa)
```

```
Residuals:
      Min       1Q   Median       3Q      Max
-115.140  -50.590   -4.216   45.330  130.361
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    817.69      35.38   23.11  <2e-16 ***
gpa            703.84      11.19   62.88  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 61.4 on 116 degrees of freedom
Multiple R-squared:  0.9715, Adjusted R-squared:  0.9713
F-statistic: 3954 on 1 and 116 DF, p-value: < 2.2e-16
```

```
> #assumption check
> plot(gpa, residuals(gg1.fit), main="Relationship between residuals and
+      the GPA",
+      xlab="GPA", ylab="Residuals")
> library(fitdistrplus)
> fnorm1 <- fitdist(residuals(gg1.fit), "norm")
> summary(fnorm1)
Fitting of the distribution ' norm ' by maximum likelihood
Parameters :
      estimate Std. Error
mean -3.885781e-15   5.604109
sd    6.087618e+01   3.962703
Loglikelihood: -652.2781 AIC: 1308.556 BIC: 1314.098
Correlation matrix:
      mean sd
mean   1  0
sd     0  1

> plot(fnorm1)
> #prediction and estimation
> confint(gg1.fit, level=0.95)
              2.5 %    97.5 %
(Intercept) 747.6105 887.7597
gpa         681.6730 726.0145
```

```
> predict(gg1.fit, data.frame(gpa=3.3), interval="confidence", level=0.95)
      fit      lwr      upr
1 3140.37 3128.485 3152.254
> predict(gg1.fit, data.frame(gpa=3.2), interval="prediction", level=0.95)
      fit      lwr      upr
1 3069.985 2947.85 3192.12
>
> gg2.fit <- lm(salary~gpa+I(gpa^2))
> summary(gg2.fit)

Call:
lm(formula = salary ~ gpa + I(gpa^2))

Residuals:
    Min       1Q   Median       3Q      Max
-116.068  -49.960   -4.509   45.477  129.776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  784.456    261.117   3.004  0.00327 **
gpa          726.219    174.555   4.160 6.15e-05 ***
I(gpa^2)     -3.662     28.510  -0.128  0.89802
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61.66 on 115 degrees of freedom
Multiple R-squared:  0.9715, Adjusted R-squared:  0.971
F-statistic: 1960 on 2 and 115 DF, p-value: < 2.2e-16
```

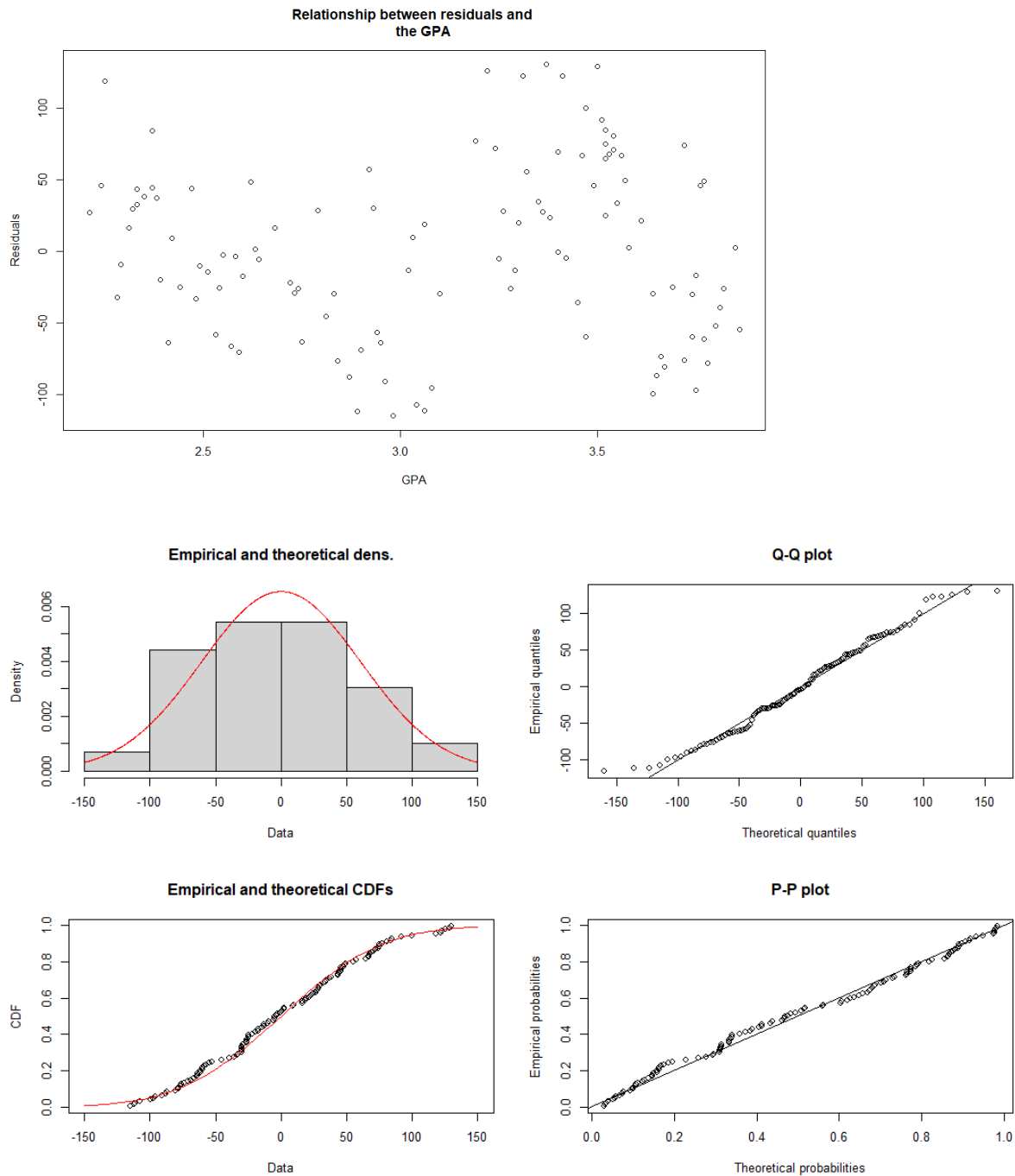
Answer of Q1



(a)

(b) $\text{Salary} = 817.69 + 703.84\text{GPA}$

(c) Based on the residual analysis, we have no reason to doubt the regression assumptions.



(d) Since $p\text{-value} < 2.2e-16 < 0.05$, reject null hypothesis. There is a significant linear relationship between the GPA and the starting monthly salary.

(e) 95% confidence interval estimate of the population slope is (681.67, 726.01)

(f) 95% confidence interval estimate of the mean of starting salary for graduates with GPA 3.3 is (\$3128.49, \$3152.25)

(g) 95% prediction interval of the starting salary for a graduate with GPA 3.2 is (\$2947.85, \$3192.12)

(h) No. Since the data range of GPA is between 2.21 and 3.86, it is not appropriate to predict salary for a graduate with GPA 2.01 which is less than 2.21.

(i) No. The second order factor of GPA is not significant as its p-value $0.8980 > 0.05$. (must have the first order and second order term in the second order factor model due to the hierarchical principle.)

R Codes of Q2

```
warehouse <- read.csv("Warehouse2021A.csv", stringsAsFactors = TRUE)
attach(warehouse)
summary(warehouse)
plot(Order, Cost, main="Relationship between distribution cost and
      number of orders", xlab="Number of orders",
      ylab="Distribution Cost in thousand dollars")
abline(regout1, lwd=3, col="red")
library(fitdistrplus)
regout1 <- lm(Cost~Order)
summary(regout1)
plot(Order, residuals(regout1), main="Model 1: Relationship between number of orders and
      Residuals", xlab="Number of orders", ylab="Residuals")
fnorm1 <- fitdist(residuals(regout1), distr="norm")
summary(fnorm1)
plot(fnorm1)

regout2 <- lm(Cost~Order+I(Order^2))
summary(regout2)
regout3 <- lm(Cost~Order+I(Order^2)+I(Order^3))
summary(regout3)
c(AIC(regout1), AIC(regout2), AIC(regout3))
plot(Order, residuals(regout2), main="Model 2: Relationship between number of orders and
      Residuals in model 2", xlab="number of orders", ylab="Residuals")
fnorm2 <- fitdist(residuals(regout2), distr="norm")
summary(fnorm2)
plot(fnorm2)
```

R Output of Q2

```
> summary(warehouse)

      Cost      Order
Min.   :151.0  Min.   : 789
1st Qu.:354.5  1st Qu.:1674
Median :469.0  Median :2087
Mean   :502.5  Mean   :2067
3rd Qu.:708.2  3rd Qu.:2719
Max.   :829.0  Max.   :2957

> plot(Order, Cost, main="Relationship between distribution cost and
+      number of orders", xlab="Number of orders",
+      ylab="Distribution Cost in thousand dollars")
> abline(regout1, lwd=3, col="red")
> library(fitdistrplus)
> regout1 <- lm(Cost~Order)
> summary(regout1)

Call:
lm(formula = Cost ~ Order)

Residuals:
    Min       1Q   Median       3Q      Max
-58.27 -21.41   0.69  20.96  60.32

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.416e+02  1.283e+01  -11.04 5.18e-15 ***
Order         3.116e-01  5.889e-03   52.91 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.17 on 50 degrees of freedom
Multiple R-squared:  0.9825, Adjusted R-squared:  0.9821
F-statistic: 2799 on 1 and 50 DF, p-value: < 2.2e-16

> plot(Order, residuals(regout1), main="Model 1: Relationship between
+      Residuals", xlab="Number of orders", ylab="Residuals")
```

```
> fnorm1 <- fitdist(residuals(regout1), distr="norm")
> summary(fnorm1)
Fitting of the distribution ' norm ' by maximum likelihood
Parameters :
      estimate Std. Error
mean -2.134777e-15   3.966894
sd    2.860568e+01   2.805018
Loglikelihood: -248.1723   AIC:  500.3446   BIC:  504.2471
Correlation matrix:
      mean sd
mean    1  0
sd      0  1

> plot(fnorm1)
> regout2 <- lm(Cost~Order+I(Order^2))
> summary(regout2)

Call:
lm(formula = Cost ~ Order + I(Order^2))

Residuals:
      Min       1Q   Median       3Q      Max
-30.7613  -8.8647  -0.3878  11.9820  28.4556

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.218e+01  1.788e+01   3.478  0.00107 **
Order        7.210e-02  1.982e-02   3.638  0.00066 ***
I(Order^2)   6.137e-05  5.022e-06  12.219 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.65 on 49 degrees of freedom
Multiple R-squared:  0.9957, Adjusted R-squared:  0.9955
F-statistic: 5626 on 2 and 49 DF, p-value: < 2.2e-16

> regout3 <- lm(Cost~Order+I(Order^2)+I(Order^3))
```



```
> summary(regout3)
```

Call:

```
lm(formula = Cost ~ Order + I(Order^2) + I(Order^3))
```

Residuals:

Min	1Q	Median	3Q	Max
-30.483	-7.540	-0.805	12.047	26.823

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.128e+01	6.211e+01	0.504	0.617
order	1.305e-01	1.140e-01	1.144	0.258
I(Order^2)	2.826e-05	6.387e-05	0.443	0.660
I(Order^3)	5.779e-09	1.112e-08	0.520	0.605

Residual standard error: 14.76 on 48 degrees of freedom

Multiple R-squared: 0.9957, Adjusted R-squared: 0.9954

F-statistic: 3695 on 3 and 48 DF, p-value: < 2.2e-16

```
> c(AIC(regout1), AIC(regout2), AIC(regout3))
```

```
[1] 502.3446 431.6471 433.3551
```

```
> plot(Order, residuals(regout2), main="Model 2: Relationship between  
number of orders and
```

```
+ Residuals in model 2", xlab="number of orders", ylab="Residuals")
```

```
> fnorm2 <- fitdist(residuals(regout2), distr="norm")
```

```
> summary(fnorm2)
```

Fitting of the distribution ' norm ' by maximum likelihood

Parameters :

	estimate	Std. Error
mean	4.895590e-17	1.971845
sd	1.421918e+01	1.394305

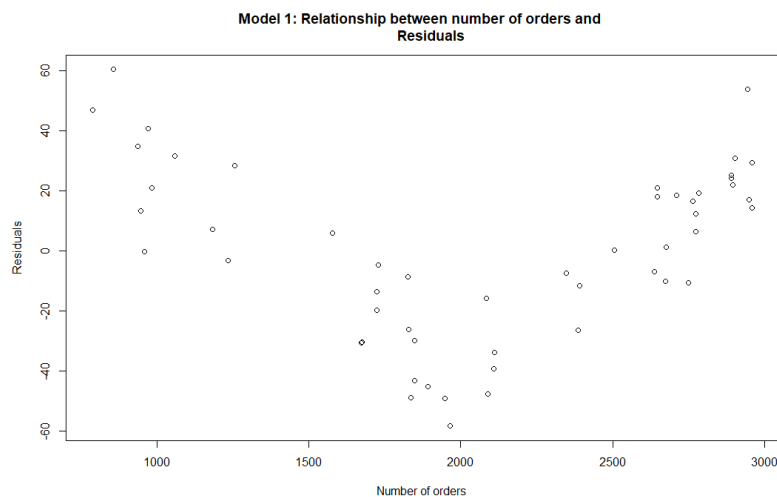
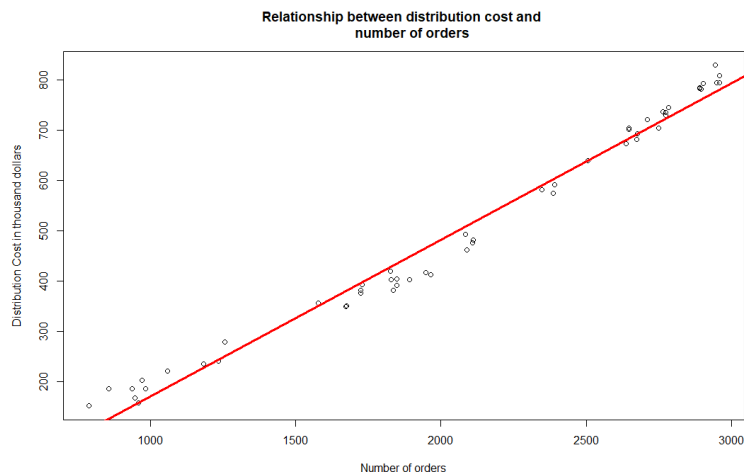
Loglikelihood: -211.8236 AIC: 427.6471 BIC: 431.5496

Correlation matrix:

	mean	sd
mean	1	0
sd	0	1

```
> plot(fnorm2)
```

Answer of Q2



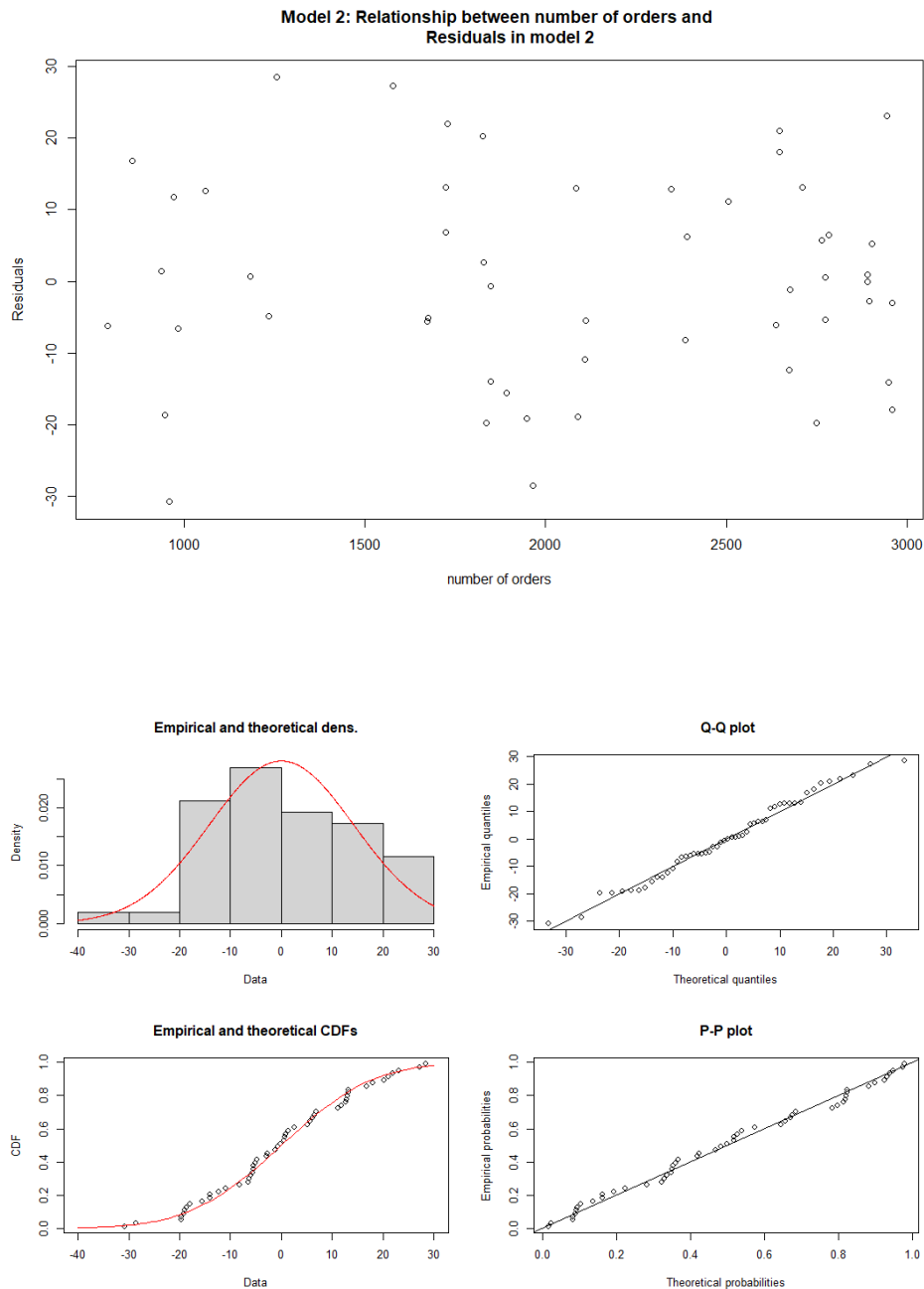
The scatter plot supports the linear relationship but the residuals plot does not as it has the V-shape pattern in the plot. It implies the higher order regression model may be more appropriate.

We try the second order and third order regression models.

Based on the adjusted R-square or/and AIC, the second order regression model is a better model as it has higher adjusted R-squared 0.9955 and lower AIC 431.6471 (while the adjusted R-square and AIC of the first order regression model are 0.9821 and 502.34, respectively and the adjusted R-square and AIC of the third order regression model are 0.9954 and 433.36, respectively

(It is acceptable if you only consider the second order regression model.)

The residual analysis of the second order regression model does not show any doubt about the regression assumptions:



Therefore, we suggest the predicting model as:

$$\text{Cost} = 62.18 + 0.0721 \text{ Order} + 6.137\text{e-}5 \text{ Order}^2$$

R Codes of Q3

#Assignment 2: Q3

```
mac <- read.csv("McDonald2021A.csv", stringsAsFactors = TRUE)
```

```
summary(mac)
```

```
attach(mac)
```

```
drive <- mac$Time[1:82]
```

```
counter <- mac$Time[83:164]
```

```
# drive <- mac$Time[Type=="Drive"]
```

```
# counter <- mac$Time[Type=="Counter"]
```

```
res1 <- t.test(drive, counter, alternative="two.sided",
```

```
          mu=0, var.equal=TRUE, conf.level=0.95 )
```

```
res1
```

```
reg1 <- lm(Time ~ Type, data=mac)
```

```
summary(reg1)
```

```
confint(reg1, level=0.95)
```

R Output of Q3

```
> mac <- read.csv("McDonald2021A.csv", stringsAsFactors = TRUE)
```

```
> summary(mac)
```

Time	Type
Min. :0.000	Counter:82
1st Qu.:1.800	Drive :82
Median :3.100	
Mean :3.068	
3rd Qu.:4.300	
Max. :6.400	

```
> drive <- mac$Time[1:82]
```

```
> counter <- mac$Time[83:164]
```

```
> res1 <- t.test(drive, counter, alternative="two.sided",  
+                 mu=0, var.equal=TRUE, conf.level=0.95 )
```

```
> res1
```

Two Sample t-test

data: drive and counter

t = -2.069, df = 162, p-value = 0.04013

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.93192746 -0.02173107

```

sample estimates:
mean of x mean of y
  2.829268  3.306098

>
> reg1 <- lm(Time ~ Type, data=mac)
> summary(reg1)

Call:
lm(formula = Time ~ Type, data = mac)

Residuals:
    Min       1Q   Median       3Q      Max
-2.8293 -1.2293 -0.1061  1.3707  3.0939

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.3061     0.1630  20.288  <2e-16 ***
TypeDrive     -0.4768     0.2305  -2.069   0.0401 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.476 on 162 degrees of freedom
Multiple R-squared:  0.02574, Adjusted R-squared:  0.01973
F-statistic: 4.281 on 1 and 162 DF, p-value: 0.04013

> confint(reg1, level=0.95)
              2.5 %      97.5 %
(Intercept)  2.9842945  3.62790058
TypeDrive    -0.9319275 -0.02173107

```

Answer of Q3

- (a) Since the p-value is $0.04013 < 0.05$, we can reject the null hypothesis. The data provide evidence that the average waiting time differs for customers in these two location experiences. In fact, the drive-through experience has shorter waiting time than the inside counter experience. (We can use the simple linear regression approach or the two sample t-test to get the p-value 0.04013. You just need to show the result based on one of two approaches.)
- (b) $\text{Time} = 3.3061 - 0.4768\text{Type}(\text{Drive})$
- (c) 95% confidence interval estimate of the population slope is $(-0.9319, -0.0217)$ which represent the average amount of waiting time for drive-through experience reducing by $(0.0217, 0.9319)$, compared with inside counter experience.

R Codes of Q4

Assignment 2 Q4

```
abc <- read.csv("ABC2021A.csv", stringsAsFactors = TRUE)
```

```
summary(abc)
```

```
attach(abc)
```

```
lm.moving1 <- lm(Hours~., data=abc)
```

```
summary(lm.moving1)
```

```
lm.moving2 <- lm(Hours~Size+Elevator+Large, data=abc)
```

```
summary(lm.moving2)
```

```
lm.moving3 <- lm(Hours~Size+Large*Elevator, data=abc)
```

```
summary(lm.moving3)
```

```
c(AIC(lm.moving1), AIC(lm.moving2), AIC(lm.moving3))
```

```
predict(lm.moving3, data.frame(Size=510, Large=1, Elevator="Yes", Rain="Yes"),  
        interval="prediction")
```

```
predict(lm.moving3, data.frame(Size=450, Large=2, Elevator="No", Rain="No"),  
        interval="confidence")
```

```
b0 <- lm.moving3$coefficients[1]+lm.moving3$coefficients[4]
```

```
b1 <- lm.moving3$coefficients[2]
```

```
b2 <- lm.moving3$coefficients[3]+lm.moving3$coefficients[5]
```

```
b0
```

```
b1
```

```
b2
```

R Output of Q4

```
> abc <- read.csv("ABC2021A.csv", stringsAsFactors = TRUE)
> summary(abc)
      Hours      Size      Large      Elevator      Rain
Min.   :13.50  Min.   :223.0  Min.   :0.000  No :30      No :49
1st Qu.:21.68  1st Qu.:385.0  1st Qu.:2.000  Yes:54     Yes:35
Median :27.10  Median :559.5  Median :3.000
Mean   :27.61  Mean   :566.3  Mean   :2.821
3rd Qu.:33.55  3rd Qu.:757.8  3rd Qu.:4.000
Max.   :45.10  Max.   :872.0  Max.   :5.000
> attach(abc)
> lm.moving1 <- lm(Hours~., data=abc)
> summary(lm.moving1)
```

```
Call:
lm(formula = Hours ~ ., data = abc)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-5.9162 -2.1467 -0.2479  2.1116  6.7038
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.42461    1.78607   6.397 1.04e-08 ***
Size         0.02369    0.00267   8.873 1.72e-13 ***
Large        1.87411    0.36905   5.078 2.48e-06 ***
ElevatorYes -4.30901    0.82715  -5.209 1.47e-06 ***
RainYes      0.60243    0.74889   0.804  0.424
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.95 on 79 degrees of freedom
Multiple R-squared:  0.883,    Adjusted R-squared:  0.8771
F-statistic: 149 on 4 and 79 DF, p-value: < 2.2e-16
```

```
> lm.moving2 <- lm(Hours~Size+Elevator+Large, data=abc)
> summary(lm.moving2)
```

```
Call:
lm(formula = Hours ~ Size + Elevator + Large, data = abc)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-6.115 -1.969 -0.271  2.058  6.356
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.168618    1.524582   7.982 8.79e-12 ***
Size         0.023210    0.002596   8.941 1.15e-13 ***
ElevatorYes -4.382621    0.820259  -5.343 8.40e-07 ***
Large        1.813037    0.360358   5.031 2.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.943 on 80 degrees of freedom
Multiple R-squared:  0.882,    Adjusted R-squared:  0.8776
F-statistic: 199.4 on 3 and 80 DF, p-value: < 2.2e-16
```

```
> lm.moving3 <- lm(Hours~Size+Large*Elevator, data=abc)
> summary(lm.moving3)
```

Call:

```
lm(formula = Hours ~ Size + Large * Elevator, data = abc)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-5.8784	-1.7742	-0.5381	1.8188	6.0665

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	8.944252	1.803628	4.959	3.97e-06	***
Size	0.022640	0.002481	9.125	5.54e-14	***
Large	2.841765	0.483951	5.872	9.68e-08	***
ElevatorYes	0.549831	1.812288	0.303	0.76239	
Large:ElevatorYes	-1.619772	0.536939	-3.017	0.00344	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.804 on 79 degrees of freedom

Multiple R-squared: 0.8942, Adjusted R-squared: 0.8889

F-statistic: 167 on 4 and 79 DF, p-value: < 2.2e-16

```
>
> c(AIC(lm.moving1), AIC(lm.moving2), AIC(lm.moving3))
[1] 426.9423 425.6276 418.4692
>
> predict(lm.moving3, data.frame(Size=510, Large=1, Elevator="Yes", Rain="
Yes"),
+         interval="prediction")
      fit      lwr      upr
1 22.26222 16.50163 28.02282
> predict(lm.moving3, data.frame(Size=450, Large=2, Elevator="No", Rain="N
o"),
+         interval="confidence")
      fit      lwr      upr
1 24.81556 23.08824 26.54288

> b0 <- lm.moving3$coefficients[1]+lm.moving3$coefficients[4]
> b1 <- lm.moving3$coefficients[2]
> b2 <- lm.moving3$coefficients[3]+lm.moving3$coefficients[5]
> b0
(Intercept)
9.494083
> b1
      Size
0.02263951
> b2
      Large
1.221992
```


Answer of Q4

- (a) M1: Hours = $11.42 + 0.02369\text{Size} + 1.874\text{Large} - 4.309\text{Elevator(Yes)} + 0.6024\text{Rain(Yes)}$
- (b) M2: Hours = $12.17 + 0.02321\text{Size} + 1.813\text{Large} - 4.3826\text{Elevator(Yes)}$
- (c) M3: Hours = $8.944 + 0.02264\text{Size} + 2.8418\text{Large} + 0.5498\text{Elevator(Yes)} - 1.6198\text{Large} * \text{Elevator(Yes)}$
- (d) The AIC of models M1, M2 and M3 are 426.94, 425.63 and 418.47, respectively. Since M3 has the lowest AIC, M3 is the best model.
- (e) 95% prediction interval is (16.50, 28.02) hours.
- (f) 95% confidence interval is (23.09, 26.54) hours.
- (g) Hours = $9.4941 + 0.02264\text{Size} + 1.2220\text{Large}$