**19.45/20**

**QUESTION 1**

> inso<-read.csv(file.choose(),header=TRUE)

> attach(inso)

> #For question 1

> #part a

> summary(inso)

ASSET87 ASSET88 ANN\_PREM CUR\_LIAB

Min. : 4.117 Min. : 3.57 Min. : -4.165 Min. : 0.003

1st Qu.: 31.664 1st Qu.: 27.98 1st Qu.: 16.930 1st Qu.: 8.081

Median : 91.785 Median : 75.94 Median : 62.372 Median : 45.704

Mean : 929.321 Mean : 808.40 Mean : 473.601 Mean : 376.396

3rd Qu.: 365.889 3rd Qu.: 321.02 3rd Qu.: 187.043 3rd Qu.: 163.883

Max. :27101.160 Max. :23679.05 Max. :13533.750 Max. :8669.124

SURPLUS PREM\_LEV STATE INDUSTRY

Min. : 1.525 Min. :-1.926 IL :14 Auto Liab.& Physical :33

1st Qu.: 7.882 1st Qu.: 1.321 MN : 9 Commercial Multiperil: 5

Median : 32.693 Median : 1.889 OH : 8 Homeowners : 4

Mean : 356.446 Mean : 2.117 WI : 8 Homeowners : 3

3rd Qu.: 95.643 3rd Qu.: 2.638 IA : 6 Medical Malpractice : 3

Max. :13985.982 Max. : 9.926 MO : 5 Worker's Compensation: 3

(Other):14 (Other) :13

> #From the summary, the most noticeable thing is that most of the variables are strongly skewed to the right with mean greater than the median.

> #This means that they are not normally distributed.

> #Since PREM\_LEV has mean that is close to its median which is 2.117, it means that PREM\_LEV variable is approximately normally distributed.

>

>

>

>

> #part b

>

> cor(inso[,c("ASSET87","ASSET88","ANN\_PREM","CUR\_LIAB","SURPLUS","PREM\_LEV")])

ASSET87 ASSET88 ANN\_PREM CUR\_LIAB SURPLUS PREM\_LEV

ASSET87 1.00000000 0.9998315 0.99476808 0.97012393 0.9822031 -0.09968432

ASSET88 0.99983154 1.0000000 0.99422568 0.96937021 0.9838436 -0.10000547

ANN\_PREM 0.99476808 0.9942257 1.00000000 0.94605412 0.9841001 -0.09131432

CUR\_LIAB 0.97012393 0.9693702 0.94605412 1.00000000 0.9146067 -0.08940129

SURPLUS 0.98220314 0.9838436 0.98410014 0.91460674 1.0000000 -0.10386182

PREM\_LEV -0.09968432 -0.1000055 -0.09131432 -0.08940129 -0.1038618 1.00000000

>

> #The two variables that are more strongly correlated with PREM\_LEV are SURPLUS with negative correlation of -0.10386182.

> #The variable that has second highest correlation with PREM\_LEV is ASSET88 with correlation of -0.1000055

>

>

> #part C

>

> scatterplot = data.frame(ASSET87,ASSET88,ANN\_PREM,CUR\_LIAB,SURPLUS,PREM\_LEV)

> pairs(scatterplot,upper.panel=NULL)



> #From the scatterplot, we can say that none of the data have a strong linear relationship and it seems to have a few outliers in each plot.

> #PREM\_LEV seems to have a negative linear relationship with every x-variables in which some of the plots distributed on scale of 0 with other x-variables.

>

> #part D

>

>

> model1=lm(PREM\_LEV~ASSET87+ASSET88+ANN\_PREM+CUR\_LIAB+SURPLUS)

> summary(model1)

Call:

lm(formula = PREM\_LEV ~ ASSET87 + ASSET88 + ANN\_PREM + CUR\_LIAB +

SURPLUS)

Residuals:

Min 1Q Median 3Q Max

-3.8644 -0.6534 -0.1179 0.3614 7.6202

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.060806 0.219115 9.405 2.86e-13 \*\*\*

ASSET87 -0.003330 0.004148 -0.803 0.4254

ASSET88 -0.004311 0.005669 -0.761 0.4500

ANN\_PREM 0.005870 0.002870 2.046 0.0453 \*

CUR\_LIAB 0.006487 0.003489 1.859 0.0681 .

SURPLUS 0.003966 0.002563 1.547 0.1272

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.546 on 58 degrees of freedom

Multiple R-squared: 0.07761, Adjusted R-squared: -0.001908

F-statistic: 0.976 on 5 and 58 DF, p-value: 0.4402

> #value of R^2 = 0.07761

> #value of R^2 = 0.07761 , a low value is not good as it does not explain variations that is present in PREM\_LEV.

> #value of R^2 adjusted = -0.001908, (a negative, low value is not good either)

> #p-value = 0.4402 which is very high (very bad)

>

>

> #part E

>

> dim(inso)

[1] 64 8

> qt(0.025,64)

[1] -1.99773

>

>

> qt(0.025,58)

[1] -2.001717

>

>

> #the t test statistic value is |-0.803| which is less than critical t value which is |-.2001717| . We accept H0 and reject H1. ASSET87 is not a significant predictor for PREM\_LEV.

>

What are the hypothesis statements?

>

> #part F

>

What are the hypothesis statements?

> # the p-value = 0.0453 which is less than alpha = 0.05. We reject H0 and accept H1. ANN\_PREM is a significant predictor of PREM\_LEV.

>

> #part G

>

> #for the case where every other variables had values of 0, PREM\_LEV will have a value of 2.0608

>

>

> #part H

>

> plot(ANN\_PREM,PREM\_LEV)

>



> cor(ANN\_PREM,PREM\_LEV)

[1] -0.09131432

>

> #Correlation between ANN\_PREM and PREM\_LEV is not so strong with negative correlation of 0.09131432.

> #As expected, the slope of the plot corresponds to the correlation as the plots shows negative linear relationship. Since the plot only consider two variables, it does not affect other variables and therefore, the result can be expected.

>

> #part I

>

> model2<-lm(PREM\_LEV~ASSET87+ASSET88+CUR\_LIAB+SURPLUS)

> model3<-lm(ANN\_PREM~ASSET87+ASSET88+CUR\_LIAB+SURPLUS)

> co2<-residuals(model2)

> co3<-residuals(model3)

> plot(co2,co3)

> cor(co2,co3)

[1] 0.2594159



> #We can see that the plot has coincide and has positive slopes. They also have positive correlation of 0.2594159.

>

**QUESTION 2**

> #Question number 2

>

> hospital<-read.csv(file.choose(),header=TRUE)

> attach(hospital)

>

>

> #part A

>

> summary(hospital)

AGE FEMALE LOS RACE TOTCHG

Min. : 0.00 Min. :0.0000 Min. : 0.000 Min. :1.000 Min. : 532

1st Qu.:12.00 1st Qu.:0.0000 1st Qu.: 1.000 1st Qu.:1.000 1st Qu.: 1158

Median :14.00 Median :1.0000 Median : 2.000 Median :1.000 Median : 1805

Mean :13.11 Mean :0.6186 Mean : 2.649 Mean :1.082 Mean : 3805

3rd Qu.:16.00 3rd Qu.:1.0000 3rd Qu.: 3.000 3rd Qu.:1.000 3rd Qu.: 3755

Max. :17.00 Max. :1.0000 Max. :41.000 Max. :6.000 Max. :48388

APRDRG

Min. : 21.0

1st Qu.:456.5

Median :753.0

Mean :607.6

3rd Qu.:754.0

Max. :952.0

>

> #from the summary, we can see that TOTCHG has mean (3805) more than the median (1805). It is skewed to the right.

>

> #part B

>

> model1<-lm(TOTCHG~AGE+LOS)

> summary(model1)

Call:

lm(formula = TOTCHG ~ AGE + LOS)

Residuals:

Min 1Q Median 3Q Max

-7577 -1958 -1443 151 41668

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3675.69 1075.78 3.417 0.000774 \*\*\*

AGE -153.22 74.00 -2.070 0.039756 \*

LOS 807.06 91.49 8.821 7.08e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4470 on 191 degrees of freedom

Multiple R-squared: 0.3142, Adjusted R-squared: 0.307

F-statistic: 43.76 on 2 and 191 DF, p-value: 2.271e-16

>

> #assuming alpha = 0.01, therefore LOS is significant because its p-value = 0 which is less than alpha.

> # AGE is not significant because its p-value = 0.0397 is greater than alpha.

>

>

> #part C

>

> dim(hospital)

[1] 194 6

> qt(0.995,191)

[1] 2.601814

>

> #from part B, we can see that the cost increases by 807.06 per day (from standard deviation), holding AGE constant.

>

> #part D

>

> boxplot(TOTCHG~FEMALE)

>



> #both plots are not normally distributed, both are skewed to the right.

> #there are a few outliers for Female.

> #The female group (number 1) has smaller median, which means they have lower expenses.

>

> #part E

>

> boxplot(TOTCHG~RACE)

>



> #RACE 1 is more skewed to the right.

> #It seems like RACE 1 has more obvious outliers.

>

>

> #part F

>

> LOGTOTCHG<-log(TOTCHG)

> model2<-lm(LOGTOTCHG~AGE+LOS)

> summary(model2)

Call:

lm(formula = LOGTOTCHG ~ AGE + LOS)

Residuals:

Min 1Q Median 3Q Max

-2.9329 -0.5628 -0.1086 0.4509 2.7144

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.99132 0.19789 40.383 < 2e-16 \*\*\*

AGE -0.04768 0.01361 -3.502 0.000574 \*\*\*

LOS 0.12739 0.01683 7.570 1.55e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8223 on 191 degrees of freedom

Multiple R-squared: 0.2866, Adjusted R-squared: 0.2792

F-statistic: 38.37 on 2 and 191 DF, p-value: 9.819e-15

>

> #R^2 value for the original data is higher than the log version. (The higher, the better)

> #R^2 adjusted for the original version is higher (better) than the log version. (.3070 vs. 0.2792)

> #The residual standard error for the log version is lower than the original version. (The log version is better)

> #The F p-value smaller for the original version than for the log version. (the smaller, the better)