> health<-read.table(choose.files(),header=TRUE,sep=",")

> #the file name that I use here is "HealthExp"

> health1<-subset(health,subset=GENDER==1)

> health2<-subset(health1,subset=EXPENDIP>0)

> attach(health2)

>

>

> #part a(i)

> summary(EXPENDIP)

Min. 1st Qu. Median Mean 3rd Qu. Max.

38.34 2986.00 5546.00 8247.00 9178.00 52400.00

> #now we can see that the median is 5546.00 and the mean is 8247.00

> #part a (ii)

>

> sd(EXPENDIP)

[1] 9482.407

> #The standard deviation that I get is 9482.407which is higher than the mean by 1235.407.

> #because Mean>median, so this means it is right skewness.

>

>

> #part b

> boxplot(EXPENDIP)

>

#right side of line is longer than left side, but many right side outliers , so graph is right skewed.

> hist(EXPENDIP)

>

> #The histogram is right skewed. Most of the data are on the left side and there are outliers on the right side.

> #The histogram is right skewed. Most of the data are on the left side and there are outliers on the right side.

> qqnorm(EXPENDIP)

> qqline(EXPENDIP)

>

> #The large number of points that deviate away from the line indicates that the data is not normally distributed.

> sqrtEXP<-sqrt(EXPENDIP)

> hist(sqrtEXP)

> 

> # The histogram seems still skewed to the right but it is better than the original histogram.

> qqnorm(sqrtEXP)

> qqline(sqrtEXP)

> 

># this qqplot is much more normally distributed than the original qqplot but it is still not normally distributed. There are also large numbers of points deviate away from the line.

> # part C

> lnEXP<-log(EXPENDIP)

>

> hist(lnEXP)

>



>#this histogram is no longer right skewed but slightly skewed to the left. It is also more likely normally distributed than the original data.

> qqnorm(lnEXP)

> qqline(lnEXP)



>#this qqplot is also no longer right skewed but slightly skewed to the left. It is also more likely normally distributed than the original data.

>#1.3

> autoIC <- read.table(choose.files(), header=TRUE, sep=",")

> #the file name that I use here is "AutoClaims"

> attach(autoIC)

> hist(PAID)

> 

>#this histogram is strongly right skewed.

> LNPAID<-log(PAID)

> hist(LNPAID)

> 

>#this histogram is now normally distributed.

> qqnorm(LNPAID)

> qqline(LNPAID)

>



>#this qqplot is now normally distributed, however there are some points that can be seen to be away from the line.

2.10

> nurse<-read.table(choose.files(),header=TRUE,sep=",")

> #this file name that I use is "WiscNursingHome"

>

> nurse1<-subset(nurse,URBAN==1)

> #for part i c

> nurse2<-subset(nurse1,CRYEAR==2000)

> attach(nurse2)

>

> LOGTPY<-log(TPY)

> cor(TPY,LOGTPY)

[1] 0.9475888

> #the correlation is high which is 0.9475888, which means it is a very strong,postive correlation

>

> #a ii

>

> cor(cbind(TPY,NUMBED,SQRFOOT),use="pairwise.complete.obs")

TPY NUMBED SQRFOOT

TPY 1.0000000 0.9740821 0.8102416

NUMBED 0.9740821 1.0000000 0.8070094

SQRFOOT 0.8102416 0.8070094 1.0000000

>

> #All the correlations is higher than 0.8070094 which is high.

> #Thus they are all strongly,positively correlated. The strongest correlation is between TPY and NUMBED.

>

>

> #a iii

>

> NUMBED10<-NUMBED/10

> cor(TPY,NUMBED10)

[1] 0.9740821

> #Now the correlation between NUMBED10 and TPY is still 0.9740821 which is very strong and positive correlation.

> #Although NUMBED scale is changed, but it does not affect the correlation result.

>

> #part B

>

> par(mfrow=c(1,2))

> plot(TPY,NUMBED)

> plot(TPY,SQRFOOT)



**(Both of these plots are backwards. You will note that in the problem it says that TPY should be the outcome variable in the regression. That means it should be the Y variable on the plots.)**

>#We can see that NUMBED and TPY has a very strong, positive correlation. Many points lie very close to a line.

>#We can see that SQRFOOT and TPY also has strong, positive correlation. However it is not as strong as NUMBED and TPY as we can see more points spread away from a line.

> #part C

> model1<-lm(TPY~NUMBED)

> summary(model1)

Call:

lm(formula = TPY ~ NUMBED)

Residuals:

Min 1Q Median 3Q Max

-62.093 -1.625 1.255 5.312 37.729

Coefficients:

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

(Intercept) -0.58920 1.87025 -0.315 0.753

NUMBED 0.91495 0.01545 59.203 <2e-16 \*\*\*

---

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

Residual standard error: 11.71 on 189 degrees of freedom

Multiple R-squared: 0.9488, Adjusted R-squared: 0.9486

F-statistic: 3505 on 1 and 189 DF, p-value: < 2.2e-16

> #R^2 = 0.9488. It is very close to 1, this means it is very good.

> #Adjusted R^2 = 0.9486. It is also very close to 1, which means very good.

> #the t value of NUMBED is 59.203, which is very high (very good).

>

> #df=189

>

> qt(0.975,189)

[1] 1.972595

> #the t(NUMBED)> critical value, which is 1.9726, which means NUMBED is a significant predictor of TPY.

>

> #cii

>

>

> model2<-lm(TPY~SQRFOOT)

> summary(model2)

Call:

lm(formula = TPY ~ SQRFOOT)

Residuals:

Min 1Q Median 3Q Max

-114.174 -16.812 -3.116 17.958 93.400

Coefficients:

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

(Intercept) 35.86091 3.96667 9.041 <2e-16 \*\*\*

SQRFOOT 1.10180 0.05813 18.955 <2e-16 \*\*\*

---

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

Residual standard error: 30.32 on 188 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.6565, Adjusted R-squared: 0.6547

F-statistic: 359.3 on 1 and 188 DF, p-value: < 2.2e-16

> qt(0.975,188)

[1] 1.972663

>

> #R^2 = 0.6565. It is above 0.5 (moderate) , which means it is ok.

> #Adjusted R^2 = 0.6547.

> #the t value of SQRFOOT is 18.955 , which is very high (very good).

> #the t(SQRFOOT)> critical value, which is 1.972663, which means SQRFOOT is a significant predictor of TPY.

>#model (1) has higher R^2 value, which means it is a better model than model (2)

> #part c iii

>

> LOGNUMBED<-log(NUMBED)

> model3<-lm(LOGTPY~LOGNUMBED)

> summary(model3)

Call:

lm(formula = LOGTPY ~ LOGNUMBED)

Residuals:

Min 1Q Median 3Q Max

-0.87956 -0.01005 0.02547 0.07130 0.20984

Coefficients:

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

(Intercept) -0.10828 0.09039 -1.198 0.232

LOGNUMBED 1.00101 0.01971 50.778 <2e-16 \*\*\*

---

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

Residual standard error: 0.1375 on 189 degrees of freedom

Multiple R-squared: 0.9317, Adjusted R-squared: 0.9313

F-statistic: 2578 on 1 and 189 DF, p-value: < 2.2e-16

> #df = 189

>

> qt(0.975,189)

[1] 1.972595

> #R^2 = 0.9317. It is very close to 1, this means it is very good.

> #Adjusted R^2 = 0.9313. It is also very close to 1, which means very good.

> #the t value of LOGNUMBED is 50.778 , which is very high (very good).

> #the t(LOGNUMBED)> critical value, which is 1.972595, which means LOGNUMBED is a significant predictor of TPY.

> #part c iv

>

> LOGSQRFOOT<-log(SQRFOOT)

> model4<-lm(LOGTPY~LOGSQRFOOT)

> summary(model4)

Call:

lm(formula = LOGTPY ~ LOGSQRFOOT)

Residuals:

Min 1Q Median 3Q Max

-0.87110 -0.15067 0.02138 0.21316 0.55224

Coefficients:

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

(Intercept) 1.7360 0.1341 12.94 <2e-16 \*\*\*

LOGSQRFOOT 0.7067 0.0344 20.55 <2e-16 \*\*\*

---

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

Residual standard error: 0.2909 on 188 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.6919, Adjusted R-squared: 0.6902

F-statistic: 422.1 on 1 and 188 DF, p-value: < 2.2e-16

> #df=188

>

> qt(0.975,188)

[1] 1.972663

>

>

> #R^2 = 0.6919. It is above 0.5 (moderate) , which means it is ok.

> #Adjusted R^2 = 0.6902

> #the t value of SQRFOOT is 20.55 , which is very high (very good).

> #the t(LOGSQRFOOT)> critical value, which is 1.972663, which means LOGSQRFOOT is a significant predictor of TPY.

> #2.20

>

> nurse<-read.table(choose.files(),header=TRUE,sep=",")

> nurse1<-subset(nurse,URBAN==1)

> nurse2<-subset(nurse1,CRYEAR==2001)

> attach(nurse2)

>

>

> #part a

>

> LOGTPY<-log(TPY)

> LOGNUMBED<-log(NUMBED)

> summary(LOGTPY)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.939 4.078 4.510 4.469 4.801 6.088

> summary(LOGNUMBED)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.045 4.171 4.590 4.564 4.905 6.125

> cor(LOGTPY,LOGNUMBED)

[1] 0.9809583

>

> plot(LOGTPY,LOGNUMBED)

>



Again, the TPY variable should be on the Y axis.

>

> #part b

>

> model1<lm(LOGTPY~LOGNUMBED)

Error: object 'model1' not found

> model1<-lm(LOGTPY~LOGNUMBED)

> summary(model1)

Call:

lm(formula = LOGTPY ~ LOGNUMBED)

Residuals:

Min 1Q Median 3Q Max

-0.86702 -0.01550 0.02279 0.05618 0.12358

Coefficients:

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

(Intercept) -0.18225 0.06795 -2.682 0.00798 \*\*

LOGNUMBED 1.01918 0.01480 68.884 < 2e-16 \*\*\*

---

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

Residual standard error: 0.105 on 186 degrees of freedom

Multiple R-squared: 0.9623, Adjusted R-squared: 0.9621

F-statistic: 4745 on 1 and 186 DF, p-value: < 2.2e-16

> #coeffficient for LOGNUMBED is 1.01918

> #R^2 = 0.9623. It is very close to 1, this means it is very good.

> #Adjusted R^2 = 0.9621. It is also very close to 1, which means very good.

> #the t value of LOGNUMBED is 68.884 , which is very high (very good).

> #the t(LOGNUMBED)> critical value, which is 1.9728, which means LOGNUMBED is a significant predictor of TPY.

> qt(0.975,186)

[1] 1.9728

>

>c i

>#the t value for test statistic is 68.884 which is greater than t critical value of 1.9728, therefore we Accept H1 and reject H0.

>#the p-value, <2.2e-16 is lesser than α=0.05, so we accept H1 and reject H0.

> #c ii

>

> tratio<-(1.01918-1)/0.01480

> tratio

[1] 1.295946

> qt(0.975,186)

[1] 1.9728

>

> #t test statistic is 1.295946, which is less than 1.9728, so we accept H0 and reject H1.

> pvalueii<-(1-pt(tratio,186))\*2

> pvalueii

[1] 0.1965999

>

> #the p-value is 0.1965999 which is greater than alpha = 0.05, so we accept H0 and reject H1

>

> #c iii

> tratio<-(1.01918-1)/0.01480

> tratio

[1] 1.295946

>

> qt(0.95,186)

[1] 1.653087

> #t test statistic = 1.295946 which is lesser than 1.653087, so we accept H0 and reject H1.

>

> pvalueiii<-(1-pt(tratio,186))

> pvalueiii

[1] 0.09829996

> #p-value is 0.0982996, which is greater than alpha = 0.05. so we accpet H0 and reject H1.

>

>

> #e i

> log100<-log(100)

> NewData<-data.frame(LOGNUMBED=log100)

> predict(model1,NewData, interval="prediction",level=.95)

fit lwr upr

1 4.511247 4.303561 4.718934

>

> #e iii

>

> #the 95% prediction interval is ( 4.303561 , 4.718934 )

>

> #e iv

>

>#e^ 4.511247 = 91.0353

>#(e^4.303561,e^, 4.718934) = (73.9627,112.0487)