**Homework Chapter 4**

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

> attach(hospital)

>

> # 4.2

>

> #part A

>

> lnTOTCHG<-log(TOTCHG)

> plot(lnTOTCHG~AGE)

>



> #many points in this scatterplot are far away from each other. It shows that Age and LNTOTCHG is not a close relationship.

>

> cor(AGE,lnTOTCHG)

[1] 0.06827977

>

> #The correlation between AGE and lnTOTCHG is low, thus it means that it doesn't have a strong relationship.

>

> model<-lm(lnTOTCHG~AGE)

> summary(model)

Call:

lm(formula = lnTOTCHG ~ AGE)

Residuals:

Min 1Q Median 3Q Max

-1.3510 -0.4396 -0.1833 0.2573 3.1520

Coefficients:

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

(Intercept) 7.510756 0.041180 182.388 <2e-16 \*\*\*

AGE 0.007308 0.004785 1.527 0.127

---

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

Residual standard error: 0.7428 on 498 degrees of freedom

Multiple R-squared: 0.004662, Adjusted R-squared: 0.002663

F-statistic: 2.333 on 1 and 498 DF, p-value: 0.1273

> #using alpha = 0.05, and the p-value is 0.1273, thus the p-value is greater than alpha, which means AGE is not a significant predictor of lnTOTCHG.

>

>

>

> #part B

>

> AGE0<-(AGE==0)

> AGE1<-(AGE!=0)

These are complete opposites, you only need to use one of these. Whichever one you don’t use will be the baseline reference value incorporated into the intercept. That is why you have NA’s in the output below.

> model2<-lm(lnTOTCHG~AGE+AGE0+AGE1)

> summary(model2)

Call:

lm(formula = lnTOTCHG ~ AGE + AGE0 + AGE1)

Residuals:

Min 1Q Median 3Q Max

-1.3936 -0.4044 -0.1507 0.2520 3.2988

Coefficients: (1 not defined because of singularities)

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

(Intercept) 8.38760 0.16866 49.729 < 2e-16 \*\*\*

AGE -0.05290 0.01218 -4.345 1.69e-05 \*\*\*

AGE0TRUE -0.92933 0.17364 -5.352 1.33e-07 \*\*\*

AGE1TRUE NA NA NA NA

---

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

Residual standard error: 0.723 on 497 degrees of freedom

Multiple R-squared: 0.0589, Adjusted R-squared: 0.05512

F-statistic: 15.55 on 2 and 497 DF, p-value: 2.807e-07

>

> #the p-value = 1.69e-05 which is less than alpha = 0.05, thus indicates that the AGE binary variable is statistically significant to predict lnTOTCHG.

The binary variable has a p-value of 1.33e-07.

> #part C

>

> FEMALE0<-(FEMALE==0)

> FEMALE1<-(FEMALE==1)

>

> model3<-lm(lnTOTCHG~AGE+FEMALE)

> summary(model3)

Call:

lm(formula = lnTOTCHG ~ AGE + FEMALE)

Residuals:

Min 1Q Median 3Q Max

-1.3664 -0.4082 -0.1804 0.2611 3.2174

Coefficients:

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

(Intercept) 7.607796 0.049934 152.358 < 2e-16 \*\*\*

AGE 0.011143 0.004871 2.287 0.022594 \*

FEMALE -0.227622 0.067659 -3.364 0.000827 \*\*\*

---

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

Residual standard error: 0.7352 on 497 degrees of freedom

Multiple R-squared: 0.02682, Adjusted R-squared: 0.02291

F-statistic: 6.85 on 2 and 497 DF, p-value: 0.001163

> model4<-(lm(lnTOTCHG~AGE\*FEMALE+FEMALE\*AGE))

> summary(model4)

Call:

lm(formula = lnTOTCHG ~ AGE \* FEMALE + FEMALE \* AGE)

Residuals:

Min 1Q Median 3Q Max

-1.5993 -0.3952 -0.1798 0.2745 3.3438

Coefficients:

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

(Intercept) 7.538970 0.053801 140.127 < 2e-16 \*\*\*

AGE 0.031254 0.007846 3.984 7.81e-05 \*\*\*

FEMALE -0.077180 0.081443 -0.948 0.34377

AGE:FEMALE -0.032345 0.009950 -3.251 0.00123 \*\*

---

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

Residual standard error: 0.7283 on 496 degrees of freedom

Multiple R-squared: 0.04713, Adjusted R-squared: 0.04136

F-statistic: 8.177 on 3 and 496 DF, p-value: 2.538e-05

>

> #the FEMALE p-value for model3 is 0.000827 which is very low compared to model4, which is 0.34377.

>

> #this indicates that the gender effect is significant in model3 but not in model4.

But, the interaction term that has gender as part of it, is significant.

> # 4.3

>

> read.table(file.choose(),header=TRUE,sep",""

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

> nursing1<-subset(nursing,subset=CRYEAR==2001)

> attach(nursing1)

>

> lnTPY<-log(TPY)

>

>

> #part E

>

> LOGNUMBED<-log(NUMBED)

> LOGSQRFOOT<-log(SQRFOOT)

>

> model1<-lm(lnTPY~LOGNUMBED+LOGSQRFOOT+URBAN+PRO+TAXEXEMPT+SELFFUNDINS+MCERT)

> summary(model1)

Call:

lm(formula = lnTPY ~ LOGNUMBED + LOGSQRFOOT + URBAN + PRO + TAXEXEMPT +

SELFFUNDINS + MCERT)

Residuals:

Min 1Q Median 3Q Max

-0.86506 -0.02214 0.01842 0.04944 0.27161

Coefficients:

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

(Intercept) -0.1682300 0.0522137 -3.222 0.0014 \*\*

LOGNUMBED 1.0012880 0.0184534 54.260 <2e-16 \*\*\*

LOGSQRFOOT 0.0210698 0.0152666 1.380 0.1684

URBAN -0.0150685 0.0104113 -1.447 0.1487

PRO -0.0003839 0.0178122 -0.022 0.9828

TAXEXEMPT 0.0189684 0.0176291 1.076 0.2827

SELFFUNDINS 0.0033153 0.0105713 0.314 0.7540

MCERT -0.0063803 0.0180890 -0.353 0.7245

---

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

Residual standard error: 0.09306 on 342 degrees of freedom

(5 observations deleted due to missingness)

Multiple R-squared: 0.9673, Adjusted R-squared: 0.9666

F-statistic: 1443 on 7 and 342 DF, p-value: < 2.2e-16

>

> #the R^2 value is 0.9673, which is high. (means very good)

> #the variable that is statistically significant is LOGNUMBED with p-value of <2e-16,which is so small.

>

> #part E (i)

>

> cor(cbind(lnTPY,LOGSQRFOOT),use="pairwise.complete.obs")

lnTPY LOGSQRFOOT

lnTPY 1.0000000 0.8150901

LOGSQRFOOT 0.8150901 1.0000000

>

> model2<-lm(lnTPY~LOGNUMBED+URBAN+PRO+TAXEXEMPT+SELFFUNDINS+MCERT)

> e2<-residuals(model2)

>

> model3<-lm(LOGSQRFOOT~LOGNUMBED+URBAN+PRO+TAXEXEMPT+SELFFUNDINS+MCERT)

> e3<-residuals(model3)

>

> cor(cbind(e2,e3),use="pairwise.complete.obs")

e2 e3

e2 1.00000000 0.07217879

e3 0.07217879 1.00000000

Warning message:

In cbind(e2, e3) :

number of rows of result is not a multiple of vector length (arg 2)

>

> #from the data, we can see that basic correlation = 0.8150901

> #from the data, we can see that partial correlation = 0.07217879

> #the partial correlation is smaller because it does not include the effects from other variables on lnTPY.

>

>

> #part E (ii)

>

> #the F score evaluates the model as a whole while the t scores evaluate every variables' marginal effect over what other x variables does not provide.

> #since the value of LOGNUMBED is very significant, it makes the model to be significant by itself.

>

>

> # 4.5

>

> #part C

>

> #H0 : B2 = B3 = B9 = 0

> # H1 : at least one of B2, B3 or B9 not equals to zero.

>

> lott<-read.table(file.choose(),header=TRUE,sep=",")

> attach(lott)

>

>

> model1<-lm(SALES~PERPERHH+MEDSCHYR+MEDHVL+PRCRENT+PRC55P+HHMEDAGE+MEDINC+POP)

>

> summary(model1)

Call:

lm(formula = SALES ~ PERPERHH + MEDSCHYR + MEDHVL + PRCRENT +

PRC55P + HHMEDAGE + MEDINC + POP)

Residuals:

Min 1Q Median 3Q Max

-6825.1 -1910.8 50.1 1099.8 14012.8

Coefficients:

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

(Intercept) 4.698e+04 3.267e+04 1.438 0.1581

PERPERHH -2.820e+03 4.034e+03 -0.699 0.4885

MEDSCHYR -3.762e+03 1.661e+03 -2.265 0.0289 \*

MEDHVL 1.504e+02 6.599e+01 2.279 0.0279 \*

PRCRENT -5.890e+01 9.393e+01 -0.627 0.5341

PRC55P -2.391e+02 3.391e+02 -0.705 0.4849

HHMEDAGE 3.080e+02 5.307e+02 0.580 0.5649

MEDINC -8.733e+01 1.342e+02 -0.651 0.5189

POP 6.645e-01 6.693e-02 9.929 1.81e-12 \*\*\*

---

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

Residual standard error: 3580 on 41 degrees of freedom

Multiple R-squared: 0.8367, Adjusted R-squared: 0.8048

F-statistic: 26.25 on 8 and 41 DF, p-value: 8.407e-14

> model2<-lm(SALES~MEDSCHYR+MEDHVL+POP)

> summary(model2)

Call:

lm(formula = SALES ~ MEDSCHYR + MEDHVL + POP)

Residuals:

Min 1Q Median 3Q Max

-6588.6 -1727.7 31.3 1365.2 14246.1

Coefficients:

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

(Intercept) 4.180e+04 1.583e+04 2.641 0.01125 \*

MEDSCHYR -3.830e+03 1.355e+03 -2.826 0.00695 \*\*

MEDHVL 1.275e+02 3.862e+01 3.302 0.00186 \*\*

POP 6.483e-01 5.592e-02 11.594 3e-15 \*\*\*

---

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

Residual standard error: 3445 on 46 degrees of freedom

Multiple R-squared: 0.8303, Adjusted R-squared: 0.8193

F-statistic: 75.05 on 3 and 46 DF, p-value: < 2.2e-16

>

> #alpha = 0.0

> #alpha = 0.05

> #all of the p-values in the observations are less than 0.05, so we can reject H0 and accept H1.

> # this indicates that MEDSCHYR, MEDHVL and POP are jointly significant variables for predicting sales.

You need to do the partial F test to determine whether terms are jointly significant. See my solution for this.

**> # Question 4**

>

> #part A

>

> life<-read.csv(file.choose(),header=TRUE,sep=",")

> attach(life)

>

> plot(PHYSICIAN,LIFEEXP)

>



> # We can see from the plot that the variables have a strong relationship.

>

> cor(cbind(PHYSICIAN,LIFEEXP),use="pairwise.complete.obs")

PHYSICIAN LIFEEXP

PHYSICIAN 1.0000000 0.6330605

LIFEEXP 0.6330605 1.0000000

>

> #the correlation of LIFEEXP and PHYSICIAN is more than 0.5, which is more than average.

>

> model1<-lm(LIFEEXP~PHYSICIAN)

> summary(model1)

Call:

lm(formula = LIFEEXP ~ PHYSICIAN)

Residuals:

Min 1Q Median 3Q Max

-19.655 -6.560 2.749 6.959 13.517

Coefficients:

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

(Intercept) 59.545693 0.929939 64.03 <2e-16 \*\*\*

PHYSICIAN 0.050744 0.004625 10.97 <2e-16 \*\*\*

---

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

Residual standard error: 8.621 on 180 degrees of freedom

(3 observations deleted due to missingness)

Multiple R-squared: 0.4008, Adjusted R-squared: 0.3974

F-statistic: 120.4 on 1 and 180 DF, p-value: < 2.2e-16

>

> # the p-value for PHYSICIAN is <2.2e-16 which is less than alpha = 0.05.

> # the variable PHYSICIAN is a significant predictor for LIFEEXP.

>

>

> # part B (i)

>

> model2<-lm(LIFEEXP~factor(REGION))

> summary(model2)

Call:

lm(formula = LIFEEXP ~ factor(REGION))

Residuals:

Min 1Q Median 3Q Max

-22.4476 -2.6448 0.3552 3.5045 20.4045

Coefficients:

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

(Intercept) 69.5476 1.4345 48.483 < 2e-16 \*\*\*

factor(REGION)2 -0.4346 1.9841 -0.219 0.8269

factor(REGION)3 2.5948 1.8350 1.414 0.1591

factor(REGION)4 -6.1698 2.6190 -2.356 0.0196 \*

factor(REGION)5 -2.1476 4.0573 -0.529 0.5972

factor(REGION)6 -17.2522 1.7435 -9.895 < 2e-16 \*\*\*

factor(REGION)7 1.7972 1.8836 0.954 0.3413

factor(REGION)8 10.0611 1.9841 5.071 9.92e-07 \*\*\*

---

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

Residual standard error: 6.574 on 177 degrees of freedom

Multiple R-squared: 0.6615, Adjusted R-squared: 0.6481

F-statistic: 49.42 on 7 and 177 DF, p-value: < 2.2e-16

>

> #the R^2 value for model1 is 0.4008 while the R^2 value for model2 is 0.6615.

> #model2 is a more preferable model because of its higher R^2 value.

>

>

> # part B (ii)

>

> newdata<-data.frame(REGION=2)

> predict(model2,newdata,interval="confidence",level=0.95)

fit lwr upr

1 69.11304 66.40804 71.81805

>

> #we are 95% confident that life expectancy in East Asia and the Pacific area will fall between 66.40804 and 71.818.

>

> #part B (iii)

>

This is a categorical variable, so you must put this in as a factor. This has caused you to not have the correct output below.

> model3<-lm(LIFEEXP~REGION)

> summary(model3)

Call:

lm(formula = LIFEEXP ~ REGION)

Residuals:

Min 1Q Median 3Q Max

-26.312 -7.712 3.740 7.740 15.853

Coefficients:

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

(Intercept) 67.9075 1.8146 37.424 <2e-16 \*\*\*

REGION -0.1826 0.3454 -0.529 0.598

---

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

Residual standard error: 11.1 on 183 degrees of freedom

Multiple R-squared: 0.001525, Adjusted R-squared: -0.003931

F-statistic: 0.2795 on 1 and 183 DF, p-value: 0.5977

>

> qt(0.025,183)

[1] -1.973012

>

> # H0 : B1 = 0

> # H1 : B1 not equals to zero.

>

> #t-statistic = |-0.529|

> #t-critical = |-1.973|

> # p-value = 0.5977 which is more than alpha = 0.05.

> # we accept H0 and reject the alternative. So, REGION is not a significant determinant of LIFEEXP.

>