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

> attach(life)

>

>

> # 5.3

>

>

>

> #part A

>

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

> plot(GDP,LIFEEXP)

> plot(PRIVATEHEALTH,LIFEEXP)



**> # in the plot of LIFEEXP vs GDP, we can see that most points are all close to 0 and just a few points have high GDP values.**

**> # in the plot of LIFEEXP vs PRIVATEHEALTH, we only have a few large outliers on the right but a better distribution of points than LIFEEXP vs GDP.**

> lnGDP<-log(GDP)

> lnHEALTH<-log(PRIVATEHEALTH)

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

> plot(lnGDP,LIFEEXP)

> plot(lnHEALTH,LIFEEXP)

>



**> # from the plot, we can see that both of the plots have better spread and it seems that there is no obvious outliers.**

> # part B

>

> # a new CSV file is created without the sections of RESEARCHERS, SMOKING and FEMALE-BOSS variables.

>

> life2<-read.csv(file.choose(),quote="",header=TRUE)

> life2b=life2[complete.cases(life2),]

>

> attach(life2b)

> lnGDP<-log(GDP)

> lnHEALTH<-log(PRIVATEHEALTH)

> model1<-lm(LIFEEXP ~ REGION + ILLITERATE + POP + FERTILITY + lnHEALTH + PUBLICEDUCATION + HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP)

REGION needs to be entered as a factor since it is a categorical variable.

> step(model1)

Start: AIC=338.59

LIFEEXP ~ REGION + ILLITERATE + POP + FERTILITY + lnHEALTH +

PUBLICEDUCATION + HEALTHEXPEND + BIRTHATTEND + PHYSICIAN +

lnGDP

Df Sum of Sq RSS AIC

- lnHEALTH 1 7.92 2605.0 336.88

- lnGDP 1 14.58 2611.6 337.12

- BIRTHATTEND 1 15.71 2612.7 337.17

- POP 1 17.35 2614.4 337.23

<none> 2597.0 338.59

- ILLITERATE 1 68.45 2665.5 339.08

- PUBLICEDUCATION 1 151.65 2748.7 342.03

- PHYSICIAN 1 280.64 2877.7 346.44

- FERTILITY 1 516.42 3113.5 354.00

- HEALTHEXPEND 1 674.87 3271.9 358.76

- REGION 1 1270.55 3867.6 374.82

Step: AIC=336.88

LIFEEXP ~ REGION + ILLITERATE + POP + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- lnGDP 1 13.61 2618.6 335.38

- POP 1 19.62 2624.6 335.60

- BIRTHATTEND 1 20.28 2625.2 335.62

<none> 2605.0 336.88

- ILLITERATE 1 65.32 2670.3 337.26

- PUBLICEDUCATION 1 158.47 2763.4 340.55

- PHYSICIAN 1 275.16 2880.1 344.52

- FERTILITY 1 508.57 3113.5 352.00

- HEALTHEXPEND 1 685.26 3290.2 357.30

- REGION 1 1309.56 3914.5 373.98

Step: AIC=335.38

LIFEEXP ~ REGION + ILLITERATE + POP + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN

Df Sum of Sq RSS AIC

- POP 1 11.87 2630.4 333.81

- BIRTHATTEND 1 14.95 2633.5 333.93

<none> 2618.6 335.38

- ILLITERATE 1 69.27 2687.8 335.89

- PUBLICEDUCATION 1 182.92 2801.5 339.86

- PHYSICIAN 1 280.15 2898.7 343.14

- FERTILITY 1 651.18 3269.8 354.70

- HEALTHEXPEND 1 1070.39 3689.0 366.28

- REGION 1 1359.71 3978.3 373.53

Step: AIC=333.81

LIFEEXP ~ REGION + ILLITERATE + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN

Df Sum of Sq RSS AIC

- BIRTHATTEND 1 27.83 2658.3 332.82

<none> 2630.4 333.81

- ILLITERATE 1 67.66 2698.1 334.25

- PUBLICEDUCATION 1 174.24 2804.7 337.97

- PHYSICIAN 1 289.13 2919.6 341.82

- FERTILITY 1 665.15 3295.6 353.46

- HEALTHEXPEND 1 1063.65 3694.1 364.41

- REGION 1 1348.06 3978.5 371.53

Step: AIC=332.82

LIFEEXP ~ REGION + ILLITERATE + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + PHYSICIAN

Df Sum of Sq RSS AIC

- ILLITERATE 1 42.15 2700.4 332.33

<none> 2658.3 332.82

- PUBLICEDUCATION 1 177.74 2836.0 337.04

- PHYSICIAN 1 327.85 2986.1 341.99

- FERTILITY 1 884.01 3542.3 358.38

- HEALTHEXPEND 1 1142.98 3801.2 365.16

- REGION 1 1545.30 4203.6 374.82

Step: AIC=332.33

LIFEEXP ~ REGION + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

PHYSICIAN

Df Sum of Sq RSS AIC

<none> 2700.4 332.33

- PUBLICEDUCATION 1 176.57 2877.0 336.41

- PHYSICIAN 1 295.44 2995.9 340.30

- FERTILITY 1 1047.68 3748.1 361.81

- HEALTHEXPEND 1 1112.50 3812.9 363.45

- REGION 1 1507.82 4208.2 372.92

Call:

lm(formula = LIFEEXP ~ REGION + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + PHYSICIAN)

Coefficients:

(Intercept) REGION FERTILITY PUBLICEDUCATION

83.362469 -1.919506 -3.130739 -0.714716

HEALTHEXPEND PHYSICIAN

0.003977 0.020627

> model2<-lm(LIFEEXP ~ ILLITERATE + POP + FERTILITY + lnHEALTH + PUBLICEDUCATION + HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP)

>

> step(model2)

Start: AIC=374.82

LIFEEXP ~ ILLITERATE + POP + FERTILITY + lnHEALTH + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- POP 1 2.52 3870.1 372.88

- lnHEALTH 1 46.92 3914.5 373.98

- PHYSICIAN 1 56.99 3924.6 374.22

- lnGDP 1 67.47 3935.1 374.48

<none> 3867.6 374.82

- ILLITERATE 1 83.70 3951.3 374.88

- PUBLICEDUCATION 1 101.83 3969.4 375.31

- HEALTHEXPEND 1 157.18 4024.8 376.64

- BIRTHATTEND 1 190.04 4057.6 377.42

- FERTILITY 1 449.17 4316.8 383.37

Step: AIC=372.88

LIFEEXP ~ ILLITERATE + FERTILITY + lnHEALTH + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- lnHEALTH 1 49.01 3919.1 372.09

- PHYSICIAN 1 60.16 3930.3 372.36

- lnGDP 1 66.16 3936.3 372.51

<none> 3870.1 372.88

- ILLITERATE 1 83.60 3953.7 372.93

- PUBLICEDUCATION 1 100.29 3970.4 373.34

- HEALTHEXPEND 1 159.51 4029.6 374.76

- BIRTHATTEND 1 208.93 4079.0 375.93

- FERTILITY 1 449.92 4320.0 381.44

Step: AIC=372.09

LIFEEXP ~ ILLITERATE + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- PHYSICIAN 1 50.11 3969.2 371.31

- lnGDP 1 59.38 3978.5 371.53

- ILLITERATE 1 74.75 3993.9 371.90

<none> 3919.1 372.09

- PUBLICEDUCATION 1 111.13 4030.3 372.77

- HEALTHEXPEND 1 162.31 4081.4 373.99

- BIRTHATTEND 1 272.38 4191.5 376.54

- FERTILITY 1 414.94 4334.1 379.75

Step: AIC=371.31

LIFEEXP ~ ILLITERATE + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND + lnGDP

Df Sum of Sq RSS AIC

- lnGDP 1 56.46 4025.7 370.67

- ILLITERATE 1 64.12 4033.4 370.85

<none> 3969.2 371.31

- PUBLICEDUCATION 1 119.30 4088.5 372.15

- HEALTHEXPEND 1 222.82 4192.0 374.55

- BIRTHATTEND 1 295.62 4264.9 376.21

- FERTILITY 1 555.71 4524.9 381.89

Step: AIC=370.67

LIFEEXP ~ ILLITERATE + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND

Df Sum of Sq RSS AIC

- ILLITERATE 1 73.63 4099.3 370.41

<none> 4025.7 370.67

- PUBLICEDUCATION 1 163.83 4189.5 372.50

- BIRTHATTEND 1 248.32 4274.0 374.41

- HEALTHEXPEND 1 504.00 4529.7 379.99

- FERTILITY 1 909.55 4935.2 388.22

Step: AIC=370.41

LIFEEXP ~ FERTILITY + PUBLICEDUCATION + HEALTHEXPEND + BIRTHATTEND

Df Sum of Sq RSS AIC

<none> 4099.3 370.41

- PUBLICEDUCATION 1 163.96 4263.3 372.17

- BIRTHATTEND 1 175.27 4274.6 372.43

- HEALTHEXPEND 1 487.16 4586.5 379.19

- FERTILITY 1 843.92 4943.2 386.38

Call:

lm(formula = LIFEEXP ~ FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND)

Coefficients:

(Intercept) FERTILITY PUBLICEDUCATION HEALTHEXPEND

70.982327 -3.325297 -0.688028 0.002363

BIRTHATTEND

0.097656

> summary(step(model1))

Start: AIC=338.59

LIFEEXP ~ REGION + ILLITERATE + POP + FERTILITY + lnHEALTH +

PUBLICEDUCATION + HEALTHEXPEND + BIRTHATTEND + PHYSICIAN +

lnGDP

Df Sum of Sq RSS AIC

- lnHEALTH 1 7.92 2605.0 336.88

- lnGDP 1 14.58 2611.6 337.12

- BIRTHATTEND 1 15.71 2612.7 337.17

- POP 1 17.35 2614.4 337.23

<none> 2597.0 338.59

- ILLITERATE 1 68.45 2665.5 339.08

- PUBLICEDUCATION 1 151.65 2748.7 342.03

- PHYSICIAN 1 280.64 2877.7 346.44

- FERTILITY 1 516.42 3113.5 354.00

- HEALTHEXPEND 1 674.87 3271.9 358.76

- REGION 1 1270.55 3867.6 374.82

Step: AIC=336.88

LIFEEXP ~ REGION + ILLITERATE + POP + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- lnGDP 1 13.61 2618.6 335.38

- POP 1 19.62 2624.6 335.60

- BIRTHATTEND 1 20.28 2625.2 335.62

<none> 2605.0 336.88

- ILLITERATE 1 65.32 2670.3 337.26

- PUBLICEDUCATION 1 158.47 2763.4 340.55

- PHYSICIAN 1 275.16 2880.1 344.52

- FERTILITY 1 508.57 3113.5 352.00

- HEALTHEXPEND 1 685.26 3290.2 357.30

- REGION 1 1309.56 3914.5 373.98

Step: AIC=335.38

LIFEEXP ~ REGION + ILLITERATE + POP + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN

Df Sum of Sq RSS AIC

- POP 1 11.87 2630.4 333.81

- BIRTHATTEND 1 14.95 2633.5 333.93

<none> 2618.6 335.38

- ILLITERATE 1 69.27 2687.8 335.89

- PUBLICEDUCATION 1 182.92 2801.5 339.86

- PHYSICIAN 1 280.15 2898.7 343.14

- FERTILITY 1 651.18 3269.8 354.70

- HEALTHEXPEND 1 1070.39 3689.0 366.28

- REGION 1 1359.71 3978.3 373.53

Step: AIC=333.81

LIFEEXP ~ REGION + ILLITERATE + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN

Df Sum of Sq RSS AIC

- BIRTHATTEND 1 27.83 2658.3 332.82

<none> 2630.4 333.81

- ILLITERATE 1 67.66 2698.1 334.25

- PUBLICEDUCATION 1 174.24 2804.7 337.97

- PHYSICIAN 1 289.13 2919.6 341.82

- FERTILITY 1 665.15 3295.6 353.46

- HEALTHEXPEND 1 1063.65 3694.1 364.41

- REGION 1 1348.06 3978.5 371.53

Step: AIC=332.82

LIFEEXP ~ REGION + ILLITERATE + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + PHYSICIAN

Df Sum of Sq RSS AIC

- ILLITERATE 1 42.15 2700.4 332.33

<none> 2658.3 332.82

- PUBLICEDUCATION 1 177.74 2836.0 337.04

- PHYSICIAN 1 327.85 2986.1 341.99

- FERTILITY 1 884.01 3542.3 358.38

- HEALTHEXPEND 1 1142.98 3801.2 365.16

- REGION 1 1545.30 4203.6 374.82

Step: AIC=332.33

LIFEEXP ~ REGION + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

PHYSICIAN

Df Sum of Sq RSS AIC

<none> 2700.4 332.33

- PUBLICEDUCATION 1 176.57 2877.0 336.41

- PHYSICIAN 1 295.44 2995.9 340.30

- FERTILITY 1 1047.68 3748.1 361.81

- HEALTHEXPEND 1 1112.50 3812.9 363.45

- REGION 1 1507.82 4208.2 372.92

Call:

lm(formula = LIFEEXP ~ REGION + FERTILITY + PUBLICEDUCATION +

HEALTHEXPEND + PHYSICIAN)

Residuals:

Min 1Q Median 3Q Max

-17.2699 -2.4995 0.2889 3.8926 10.1838

Coefficients:

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

(Intercept) 83.3624695 2.8533753 29.215 < 2e-16 \*\*\*

REGION -1.9195060 0.2707756 -7.089 2.94e-10 \*\*\*

FERTILITY -3.1307395 0.5298170 -5.909 6.05e-08 \*\*\*

PUBLICEDUCATION -0.7147163 0.2946219 -2.426 0.0173 \*

HEALTHEXPEND 0.0039771 0.0006531 6.089 2.74e-08 \*\*\*

PHYSICIAN 0.0206267 0.0065734 3.138 0.0023 \*\*

---

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

Residual standard error: 5.478 on 90 degrees of freedom

Multiple R-squared: 0.7678, Adjusted R-squared: 0.7549

F-statistic: 59.53 on 5 and 90 DF, p-value: < 2.2e-16

> summary(step(model2))

Start: AIC=374.82

LIFEEXP ~ ILLITERATE + POP + FERTILITY + lnHEALTH + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- POP 1 2.52 3870.1 372.88

- lnHEALTH 1 46.92 3914.5 373.98

- PHYSICIAN 1 56.99 3924.6 374.22

- lnGDP 1 67.47 3935.1 374.48

<none> 3867.6 374.82

- ILLITERATE 1 83.70 3951.3 374.88

- PUBLICEDUCATION 1 101.83 3969.4 375.31

- HEALTHEXPEND 1 157.18 4024.8 376.64

- BIRTHATTEND 1 190.04 4057.6 377.42

- FERTILITY 1 449.17 4316.8 383.37

Step: AIC=372.88

LIFEEXP ~ ILLITERATE + FERTILITY + lnHEALTH + PUBLICEDUCATION +

HEALTHEXPEND + BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- lnHEALTH 1 49.01 3919.1 372.09

- PHYSICIAN 1 60.16 3930.3 372.36

- lnGDP 1 66.16 3936.3 372.51

<none> 3870.1 372.88

- ILLITERATE 1 83.60 3953.7 372.93

- PUBLICEDUCATION 1 100.29 3970.4 373.34

- HEALTHEXPEND 1 159.51 4029.6 374.76

- BIRTHATTEND 1 208.93 4079.0 375.93

- FERTILITY 1 449.92 4320.0 381.44

Step: AIC=372.09

LIFEEXP ~ ILLITERATE + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND + PHYSICIAN + lnGDP

Df Sum of Sq RSS AIC

- PHYSICIAN 1 50.11 3969.2 371.31

- lnGDP 1 59.38 3978.5 371.53

- ILLITERATE 1 74.75 3993.9 371.90

<none> 3919.1 372.09

- PUBLICEDUCATION 1 111.13 4030.3 372.77

- HEALTHEXPEND 1 162.31 4081.4 373.99

- BIRTHATTEND 1 272.38 4191.5 376.54

- FERTILITY 1 414.94 4334.1 379.75

Step: AIC=371.31

LIFEEXP ~ ILLITERATE + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND + lnGDP

Df Sum of Sq RSS AIC

- lnGDP 1 56.46 4025.7 370.67

- ILLITERATE 1 64.12 4033.4 370.85

<none> 3969.2 371.31

- PUBLICEDUCATION 1 119.30 4088.5 372.15

- HEALTHEXPEND 1 222.82 4192.0 374.55

- BIRTHATTEND 1 295.62 4264.9 376.21

- FERTILITY 1 555.71 4524.9 381.89

Step: AIC=370.67

LIFEEXP ~ ILLITERATE + FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND

Df Sum of Sq RSS AIC

- ILLITERATE 1 73.63 4099.3 370.41

<none> 4025.7 370.67

- PUBLICEDUCATION 1 163.83 4189.5 372.50

- BIRTHATTEND 1 248.32 4274.0 374.41

- HEALTHEXPEND 1 504.00 4529.7 379.99

- FERTILITY 1 909.55 4935.2 388.22

Step: AIC=370.41

LIFEEXP ~ FERTILITY + PUBLICEDUCATION + HEALTHEXPEND + BIRTHATTEND

Df Sum of Sq RSS AIC

<none> 4099.3 370.41

- PUBLICEDUCATION 1 163.96 4263.3 372.17

- BIRTHATTEND 1 175.27 4274.6 372.43

- HEALTHEXPEND 1 487.16 4586.5 379.19

- FERTILITY 1 843.92 4943.2 386.38

Call:

lm(formula = LIFEEXP ~ FERTILITY + PUBLICEDUCATION + HEALTHEXPEND +

BIRTHATTEND)

Residuals:

Min 1Q Median 3Q Max

-22.404 -3.045 1.398 4.236 14.244

Coefficients:

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

(Intercept) 70.9823272 6.3128175 11.244 < 2e-16 \*\*\*

FERTILITY -3.3252974 0.7682738 -4.328 3.85e-05 \*\*\*

PUBLICEDUCATION -0.6880283 0.3606427 -1.908 0.05957 .

HEALTHEXPEND 0.0023633 0.0007186 3.289 0.00143 \*\*

BIRTHATTEND 0.0976560 0.0495091 1.972 0.05159 .

---

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

Residual standard error: 6.712 on 91 degrees of freedom

Multiple R-squared: 0.6475, Adjusted R-squared: 0.6321

F-statistic: 41.8 on 4 and 91 DF, p-value: < 2.2e-16

**> # the second model which does not include REGION in the model has higher AIC of 370.41 than the first model, with AIC of 332.33.**

So, does that make that model better or worse?

> # part C (i)

>

> attach(life2)

The following objects are masked from life2b:

BIRTHATTEND, COUNTRY, FERTILITY, GDP, HEALTHEXPEND, ILLITERATE,

LIFEEXP, PHYSICIAN, POP, PRIVATEHEALTH, PUBLICEDUCATION, REGION

> lnGDP<-log(GDP)

> lnHEALTH<-log(PRIVATEHEALTH)

> model3<-lm(LIFEEXP ~ FERTILITY + PUBLICEDUCATION + lnHEALTH)

>

> rstandard<-rstandard(model3)

> leverages<-hatvalues(model3)

>

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

> hist(rstandard)

> hist(leverages)



You need to comment on these plots. Are they skewed? Are there outliers or high leverage points here? Discuss this and the appropriate cutoffs.

QUESTION 5.4

> term<-read.csv(file.choose(),quote="",header=TRUE)

> term1<-subset(term,FACE>0)

> attach(term1)

>

> # 5.4

>

> dim(term1)

[1] 275 18

>

> lnFACE<-log(FACE)

> lnINCOME<-log(INCOME)

>

> model1<-lm(lnFACE ~ lnINCOME + EDUCATION + NUMHH + factor(MARSTAT) + AGE + GENDER)

> summary(model1)

Call:

lm(formula = lnFACE ~ lnINCOME + EDUCATION + NUMHH + factor(MARSTAT) +

AGE + GENDER)

Residuals:

Min 1Q Median 3Q Max

-5.8093 -0.8354 0.1154 0.8171 4.5824

Coefficients:

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

(Intercept) 2.886672 0.930791 3.101 0.00213 \*\*

lnINCOME 0.436053 0.079038 5.517 8.15e-08 \*\*\*

EDUCATION 0.213627 0.038708 5.519 8.07e-08 \*\*\*

NUMHH 0.234231 0.074229 3.156 0.00179 \*\*

factor(MARSTAT)1 0.179107 0.339458 0.528 0.59820

factor(MARSTAT)2 -0.672994 0.575681 -1.169 0.24343

AGE -0.005916 0.008036 -0.736 0.46228

GENDER 0.714108 0.372637 1.916 0.05639 .

---

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

Residual standard error: 1.507 on 267 degrees of freedom

Multiple R-squared: 0.368, Adjusted R-squared: 0.3515

F-statistic: 22.21 on 7 and 267 DF, p-value: < 2.2e-16

> # part a)

> install.packages("car")

--- Please select a CRAN mirror for use in this session ---

trying URL 'http://cran.case.edu/bin/macosx/contrib/3.1/car\_2.0-20.tgz'

Content type 'application/x-gzip' length 1327641 bytes (1.3 Mb)

opened URL

==================================================

downloaded 1.3 Mb

The downloaded binary packages are in

/var/folders/2z/843dcpq55wl\_xy1f1m7fxy4c0000gn/T//Rtmpb0kgE8/downloaded\_packages

> help(vif)

No documentation for ‘vif’ in specified packages and libraries:

you could try ‘??vif’

> library(car)

> help(vif)

starting httpd help server ... done

>

>

> vif(model1)

GVIF Df GVIF^(1/(2\*Df))

lnINCOME 1.264796 1 1.124631

EDUCATION 1.175433 1 1.084174

NUMHH 1.482237 1 1.217472

factor(MARSTAT) 2.367621 2 1.240446

AGE 1.178852 1 1.085749

GENDER 1.959108 1 1.399681

>

> part A (i)

**> # The idea of collinearity is that it occurs when one explanatory variable is, or nearly is, a linear combination of the other explanatory variables.**

**> # The jth variance inflation factor is V IFj =1/(1-R²j) where R² is the coefficient of determination from a regression using xj as the "response" and the other x's (x1; x2;…; xj¡1; xj+1;…; xl) as the explanatory variables. It measures the linear relationship among the explanatory variables.**

> #part a (iv)

>

> cor(cbind(lnINCOME, EDUCATION, NUMHH, MARSTAT, AGE, GENDER), use="pairwise.complete.obs")

lnINCOME EDUCATION NUMHH MARSTAT AGE

lnINCOME 1.00000000 0.34270358 0.1793354 0.17516891 0.045229132

EDUCATION 0.34270358 1.00000000 -0.0635292 -0.03823413 0.091636944

NUMHH 0.17933542 -0.06352920 1.0000000 0.41698130 -0.318283557

MARSTAT 0.17516891 -0.03823413 0.4169813 1.00000000 -0.049174005

AGE 0.04522913 0.09163694 -0.3182836 -0.04917401 1.000000000

GENDER 0.22923044 -0.04449813 0.2969410 0.63691359 0.005044888

GENDER

lnINCOME 0.229230443

EDUCATION -0.044498132

NUMHH 0.296941007

MARSTAT 0.636913587

AGE 0.005044888

GENDER 1.000000000

>

**> # From the data, it seems like collinearity is not really of an issue for this model. This is because we can see that correlation between those variables are pretty small except between MARSTAT and GENDER.**

What about the VIF values and the cutoff for this?

> # part b (ii)

> summary(hatvalues(model1))

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

0.005291 0.011970 0.019000 0.029090 0.037540 0.167800

>

> leverages <- hatvalues(model1)

> leverages[order(leverages)]

136 114 210 47 58 171

0.005290575 0.005994762 0.006274923 0.006510173 0.006600594 0.006801594

66 40 248 138 33 218

0.006915791 0.007188567 0.007252409 0.007326728 0.007637554 0.007665381

28 107 259 253 167 44

0.007742257 0.007753066 0.007777673 0.007796084 0.008324810 0.008411188

163 88 189 82 139 115

0.008496096 0.008553840 0.008876447 0.009003933 0.009045762 0.009068844

182 84 57 76 166 201

0.009172999 0.009218534 0.009245125 0.009425485 0.009451695 0.009566183

274 96 61 156 148 30

0.009574020 0.009647785 0.009802844 0.009858126 0.009887088 0.009920076

54 50 216 191 165 250

0.009999938 0.010051284 0.010055297 0.010081034 0.010340991 0.010344528

42 199 25 261 39 272

0.010362289 0.010365960 0.010416433 0.010523131 0.010533992 0.010560113

97 118 245 113 192 180

0.010647363 0.010749932 0.010814478 0.010848737 0.010852216 0.010962396

83 12 51 27 135 98

0.010981068 0.011087252 0.011134320 0.011291994 0.011324143 0.011333839

252 228 46 49 75 85

0.011411148 0.011417513 0.011536152 0.011536424 0.011540224 0.011549587

74 226 124 31 36 183

0.011770950 0.011935240 0.011966686 0.011977499 0.012047253 0.012150251

188 32 181 164 15 108

0.012189493 0.012207954 0.012286435 0.012320163 0.012341616 0.012537687

78 24 258 3 212 168

0.012620731 0.012624097 0.012724192 0.012813703 0.012867394 0.012977821

20 229 101 79 94 69

0.012979530 0.013008712 0.013018721 0.013058484 0.013210564 0.013523033

63 239 178 220 200 175

0.013583025 0.013802433 0.013841335 0.013872197 0.013943155 0.013952734

4 207 35 256 140 149

0.014039862 0.014091334 0.014110968 0.014185221 0.014313121 0.014321151

194 238 104 106 72 60

0.014347352 0.014428077 0.014458341 0.014585516 0.014634049 0.014655444

67 100 99 116 29 68

0.014660295 0.014968485 0.014977478 0.014984875 0.015091312 0.015158390

196 172 117 105 243 264

0.015192427 0.015204408 0.015212424 0.015323630 0.015524537 0.015582089

38 145 273 73 267 155

0.015799820 0.015877526 0.016218146 0.016259277 0.016285957 0.016654296

190 43 213 266 48 62

0.016952346 0.017212996 0.017265162 0.017283059 0.017304302 0.017961466

240 184 6 128 1 144

0.018139940 0.018152381 0.018157188 0.018669367 0.018856224 0.019004577

176 126 90 195 157 255

0.019290140 0.019300988 0.019407702 0.019544169 0.019569918 0.019638790

125 111 8 64 186 59

0.019651371 0.019688268 0.019701622 0.019874746 0.020137926 0.020407398

246 7 34 86 251 151

0.021019118 0.021278713 0.021379163 0.021518992 0.021812226 0.022112911

119 150 103 263 225 224

0.022218523 0.022336150 0.022468274 0.022606012 0.022648141 0.023350978

193 71 152 91 81 242

0.023608436 0.023665588 0.024230096 0.024623713 0.025121385 0.025132970

221 244 2 198 10 160

0.025705841 0.025814885 0.025945670 0.025995202 0.026913416 0.027236021

241 121 23 159 137 133

0.027988122 0.028396354 0.028939990 0.028945511 0.029111995 0.030062177

214 45 185 93 110 204

0.030177913 0.030525939 0.030703810 0.030772524 0.030773569 0.030975393

269 89 146 70 132 202

0.031367479 0.032252211 0.032280915 0.032434812 0.032533104 0.032686202

231 260 14 129 158 22

0.033218569 0.033409320 0.033604808 0.033697181 0.033762523 0.034016328

262 131 270 142 80 215

0.034198725 0.034914849 0.034983298 0.035086962 0.035325534 0.036057693

122 18 169 147 109 112

0.036406043 0.037312480 0.037758354 0.037950686 0.038138481 0.038402634

141 236 55 16 53 211

0.039070279 0.039614084 0.040317895 0.040679495 0.040696991 0.040764440

37 187 154 9 219 234

0.040808265 0.041186618 0.041599812 0.041882685 0.041944345 0.042401283

173 227 17 130 209 143

0.043302549 0.043384211 0.043455428 0.044768512 0.044868043 0.045483740

247 237 208 265 203 162

0.045539552 0.046066158 0.046248725 0.046452659 0.046788028 0.047118120

56 65 179 170 235 19

0.048167632 0.048269429 0.048596257 0.048966391 0.049218949 0.049494416

123 161 271 134 174 87

0.052659249 0.052979962 0.053334294 0.053903603 0.054623421 0.057148635

205 120 222 249 77 26

0.057437548 0.057523631 0.060168312 0.062950866 0.063125486 0.064578871

92 268 275 233 153 254

0.065261697 0.067669926 0.072892307 0.080250995 0.082667205 0.082894255

95 257 177 102 21 5

0.088010198 0.088434927 0.098717948 0.102035233 0.103045555 0.103532735

127 223 217 230 206 41

0.104525380 0.106877539 0.107489507 0.109069445 0.111462817 0.111842692

13 52 11 197 232

0.112523397 0.114603319 0.115993668 0.118728702 0.167763045

>

>

**> # From the summary statistics, we can see the maximum value is 0.167800.**

**> # an observation is considered to be high leverage if it exceeds 0.087273 (for this model)**

**> # from these observations, we can see that there are 17 observations that exceed the high leverage value : 95, 257, 177, 102, 21, 5, 127, 223, 217,230, 206, 41, 13, 52, 11, 197, and 232.**

> #part C (i)

>

> plot(residuals(model1) ~ fitted.values(model1), main="Residuals vs Fitted values")

>



**> # This plot can be used to detect heteroscedasticity graphically. As we can see, the plot shows heteroscedasticity because there is a large amount of spread for values in middle of x-axis and small amount of spread for values at the extreme.**

> #part C (ii)

>

> qqnorm(residuals(model1))

> qqline(residuals(model1))

>



**> # As we can see from the plot, it can be interpreted that the normality approximate assumption is violated due to the problems with the large and small residuals.**

> # part C (iii)

>

> plot(residuals(model1)~leverages, main="Residuals vs Leverages")

>



**> # This plot can help us to identify unusual points both in x and y space. This plot shows several problems with the outliers and also with the leverage. Many of the points exceeds the cut off point of 0.087273.**

**CHAPTER 6**

> job<-read.csv(file.choose(),quote="",header=TRUE)

> attach(job)

>

> # Chapter 6

>

> model1<-lm(GENENG~NRaw+ERaw+ORaw+ARaw+CRaw+MEANING+SAFETY+AVAILABLE+FIT)

You should have also used POSITION as an explanatory variable.

> summary(model1)

Call:

lm(formula = GENENG ~ NRaw + ERaw + ORaw + ARaw + CRaw + MEANING +

SAFETY + AVAILABLE + FIT)

Residuals:

Min 1Q Median 3Q Max

-3.4239 -0.2686 0.0406 0.3173 2.1170

Coefficients:

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

(Intercept) 1.640886 0.249181 6.585 5.83e-11 \*\*\*

NRaw -0.007643 0.003155 -2.422 0.0155 \*

ERaw 0.013761 0.002524 5.452 5.62e-08 \*\*\*

ORaw 0.006522 0.003199 2.038 0.0417 \*

ARaw 0.002284 0.002774 0.823 0.4104

CRaw 0.017452 0.002257 7.734 1.67e-14 \*\*\*

MEANING 0.156899 0.014107 11.122 < 2e-16 \*\*\*

SAFETY 0.041401 0.014556 2.844 0.0045 \*\*

AVAILABLE 0.138290 0.022841 6.055 1.69e-09 \*\*\*

FIT 0.187404 0.013486 13.896 < 2e-16 \*\*\*

---

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

Residual standard error: 0.5117 on 1943 degrees of freedom

(12 observations deleted due to missingness)

Multiple R-squared: 0.3819, Adjusted R-squared: 0.379

F-statistic: 133.4 on 9 and 1943 DF, p-value: < 2.2e-16

> library(abind)

Error in library(abind) : there is no package called ‘abind’

>

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

> plot(SAFETY,GENENG)

> plot(ARaw,GENENG)

Why are you only looking at plots of these 2 X variables?



Comment on these plots.

You need to also take a look at the histograms of the explanatory variables to see whether they are skewed to see if they need a log transformation.

> lnGENENG<-log(GENENG)

>

> model2<-lm(lnGENENG~NRaw+ERaw+ORaw+ARaw+CRaw+MEANING+SAFETY+AVAILABLE+FIT)

> summary(model2)

Call:

lm(formula = lnGENENG ~ NRaw + ERaw + ORaw + ARaw + CRaw + MEANING +

SAFETY + AVAILABLE + FIT)

Residuals:

Min 1Q Median 3Q Max

-1.21462 -0.04505 0.00936 0.05913 0.38070

Coefficients:

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

(Intercept) 0.9589120 0.0495139 19.367 < 2e-16 \*\*\*

NRaw -0.0012365 0.0006270 -1.972 0.04874 \*

ERaw 0.0025378 0.0005016 5.060 4.59e-07 \*\*\*

ORaw 0.0009666 0.0006358 1.520 0.12857

ARaw 0.0004536 0.0005512 0.823 0.41060

CRaw 0.0030273 0.0004484 6.751 1.93e-11 \*\*\*

MEANING 0.0316808 0.0028031 11.302 < 2e-16 \*\*\*

SAFETY 0.0081012 0.0028923 2.801 0.00515 \*\*

AVAILABLE 0.0259884 0.0045386 5.726 1.19e-08 \*\*\*

FIT 0.0374373 0.0026797 13.971 < 2e-16 \*\*\*

---

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

Residual standard error: 0.1017 on 1943 degrees of freedom

(12 observations deleted due to missingness)

Multiple R-squared: 0.3716, Adjusted R-squared: 0.3687

F-statistic: 127.7 on 9 and 1943 DF, p-value: < 2.2e-16

Do you think this log transformation has worked?

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

> plot(SAFETY,GENENG)

> plot(ARaw,GENENG)

>



> # the multiple R squared value is lower in the first model but higher in the second model where log GENENG transformation is used. What about the F p-value and s values? So, which model do you think is best?

You need to look at a graph of the standardized residuals and leverage points to determine if there are any outliers and high leverage points that need to be removed.