M349R (Unique 54230)

**Instructor:** Gustavo Cepparo

Project 3

Juan Acosta

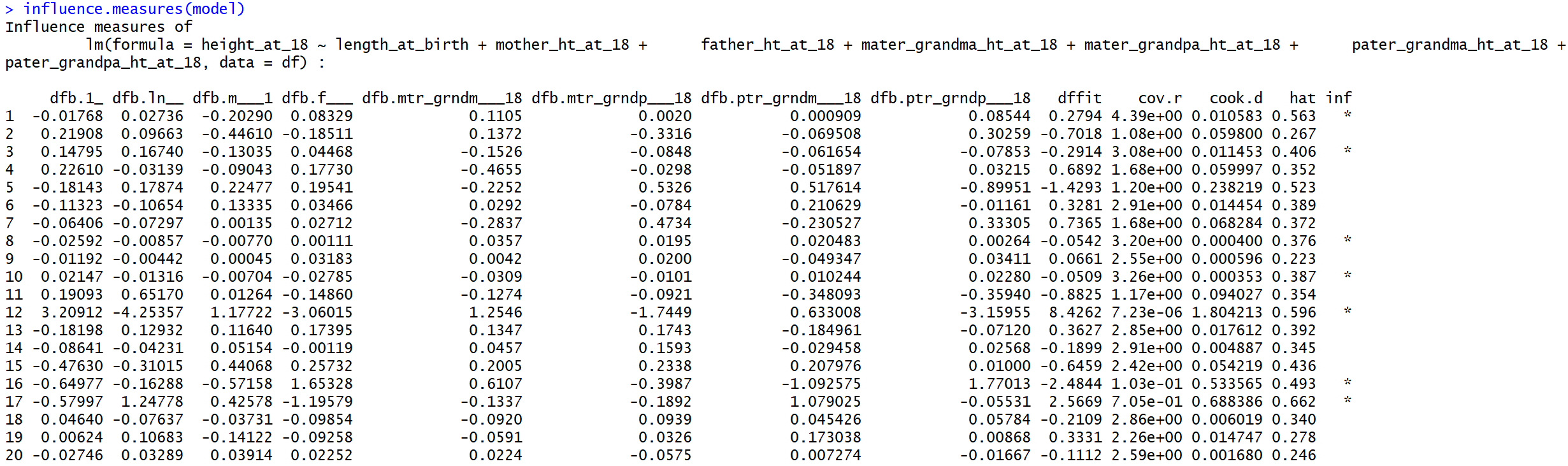
JA45384

Fall 2018

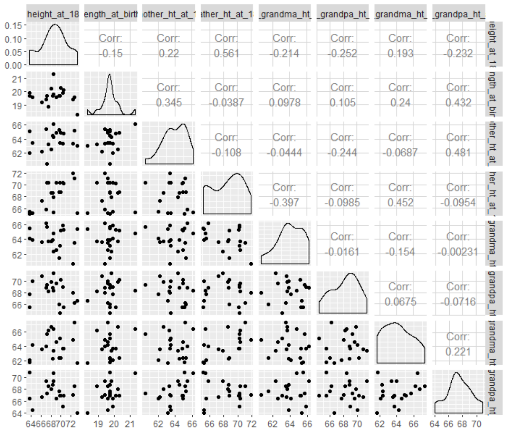
**Problem 1**

a) No points with high leverage

b) Obs. 11 has small length\_at\_birth, Obs. 16 large length\_at\_birth, and Obs. 15 small pater\_grandpa\_ht\_at\_18.

c) 

looking at difft we can see Obs. 1, 3, 8, 9, 12, 16, 17 are influential as they are flagged to be over the fence by r. These values could be treated differently by adding a dummy variable to create a subcategory or they could be dropped from the output.

d)

There does not seem to be a very big issue with multicollinearity yet there are some slightly high values between parents and grandparents

**Problem 2**

p(favorable|high) = 400 / 2000 = 0.2

p(unfavorable|high) = 1600 / 2000 = 0.8

p(favorable|low) = 250 / 1000 = 0.25

p(unfavorable|low) = 750 / 1000 = 0.75

odds ratio(favorable) = (250 \* 1600) / (400 \* 750) = 1.333

log(odds ratio) = 0.28768207245178085

sigma log(odds ratio) = sqrt((1/250) + (1/750) + (1/400) + (1/1600)) = 0.0919691977

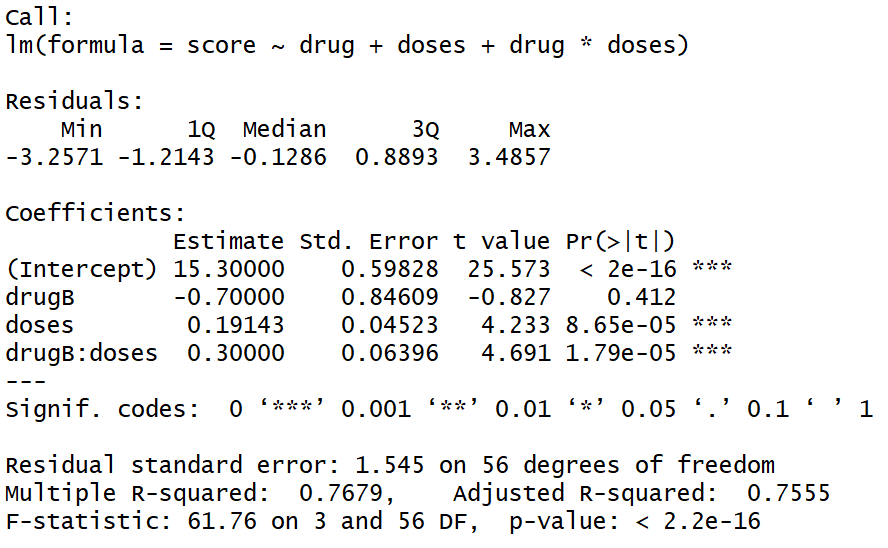
lower bound = exp(log(odds ratio)– z score \* sigma log(odds ratio)) = 1.1134

upper bound = exp(log(odds ratio) + z score \* sigma log(odds ratio)) = 1.5967

Calculating the log(odds ratio) to create a confidence interval we get the CI (1.1134,1.5967) and we can conclude that a favorable response is related to stress since 1 is not in the CI.

**Problem 3**

a)



Anxiety Score = 15.30000 - 0.70000\*x1 + 0.19143\*x2 + 0.3\*x1x2

(0.59828) (0.84609) (0.04523) (0.06396)

R2 = 0.769

RMSE = 1.555

β0 -> There are no rats with no drugs so the intercept is beyond our range of prediction

β1 -> controlling for drug dosage, the use of drug A decreases the anxiety score of

the rat by 0.7

β2 -> controlling for the type of drug, every milligram increase of drug

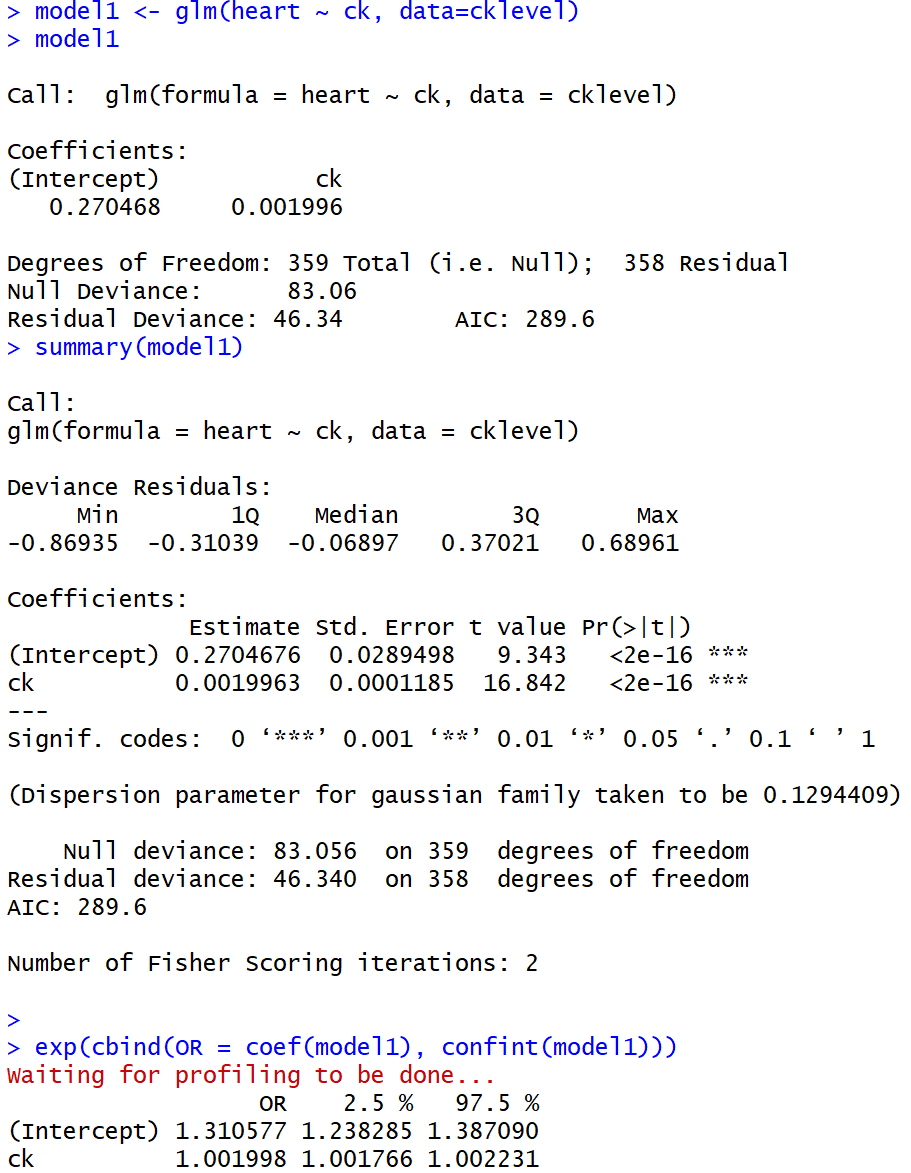
increases the anxiety score by 0.19143

β3 -> controlling for dosage, every milligram increase of drug A increases the anxiety score

by 0.3 more than a milligram increase of drug B

b) looking at the p-value of β3 we can see that 0.0000179 < 0.05 Therefore, is sufficiently different than 0. Which implies that x1 and x2 produce different changes in anxiety score of the rats

**Problem 4**

For created dataset cklevel with ck measurement of creatinine kinase of patient suspected of having a heart attack and heart a boolean classifier that determines if the patient indeed had a heart attack (1) or not (0)

The model:

heart = β0 + β1\*ck

Predicted coefficients with std err in parenthesis:

β0 = 0.2704676 (0.0289498)

β1 = 0.0019963 (0.0001185)

We observe that = β1 is statistically significant.

We can determine that for every one unit change in ck, the log odds of suffering a heart attack (versus not having a heart attack) increases by 0.0019963

**Problem 5**

Part 1

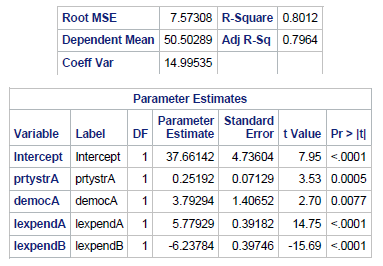
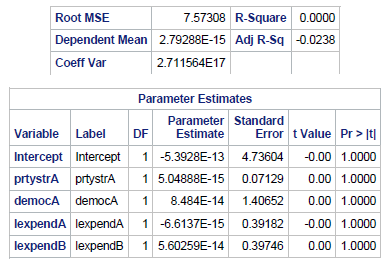
Following properties of variance, we know that for Var(u|inc, price, educ, female) = σ2inc2 if we divide u by sqrt(inc2) = inc we will get Var(u/inc| inc, price, educ, female) = σ2. Therefore,

Beer/inc = β0/inc + β1 + β2price/inc + β3educ/inc + β4female/inc + u/inc

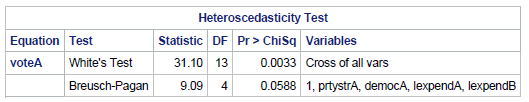
Part 2

i) R2 = 0 because the residuals are the part of the variability of the data that is not represented by these variables

initial regression residuals regression

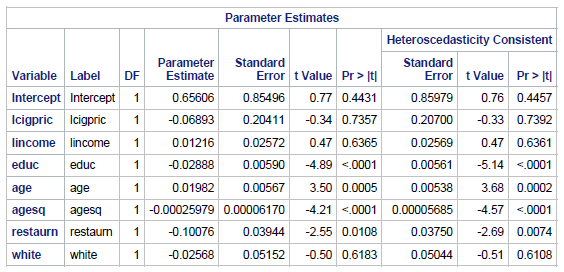
 

ii)



Homogeneity of variance of residuals has been met (Homoskedasticity) since 0.058> 0.05

Part 3

i) The robust standard errors are slightly larger

Robust Standard Errors

ii)if educ = 4, holding other factors fixed, the log odd of smoking will decrease approximately on the average by 0.02888\*4 = 0.11552

iii) 0.00025979\*age2 will surpass age\*0.01982 at age = 991000/25979 = 38 Therefore, approximately on the average after age 38, the odds of log smoking decreases as age increases.

iv) If a person lives in a state with regulations regarding smoking in restaurants, then this person will decrease approximately on the average their log odds of smoking by 10% holding everything else fixed

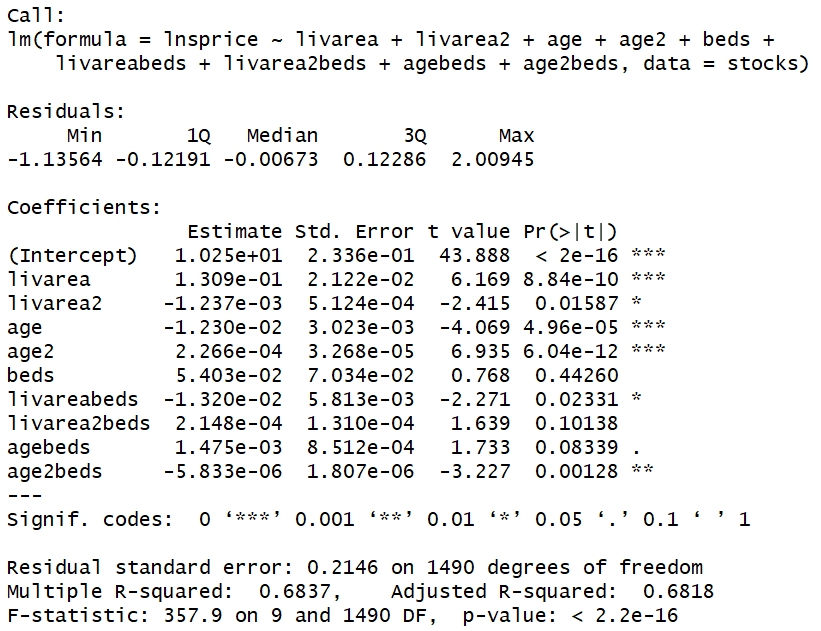
v)Calculating:

0.65606-0.06893\*log(67.44)+0.01216\*log(6500)-0.02888\*16+0.01982\*77-0.00025979\*77\*77-0.10076\*0-0.02568\*0 = 0.10012282889

The log odds of this person being a smoker is 10.012282889% according to our model.

This makes sense, as the person is above 38, they’re well educated, live in a state with regulations regarding smoking in restaurants and is not white.

**Problem 6**

a)

Ln(sprice) = 10.25 + 0.1309\*livarea - 0.001237\*livarea2 – 0.0123\*age + 0.0002266\*age2 + 0.05403\*beds – 0.0132\*livarea\*beds + 0.0002148\*livarea2\*beds + 0.001475\*age\*beds – 0.000005833age2\*beds

For 2 bedroom houses:

* The direct effect of 2 beds on ln(sprice) will on average increase by 0.10806
* The effect of livarea on ln(sprice) will decrease on average by 0.0264 and the effect of livarea2 on ln(sprice) will increase on average by 0.0004296.
* The effect of age on ln(sprice) will increase on average by 0.002950 and the effect of age2 on ln(sprice) will decrease on average by 0.0000011666.

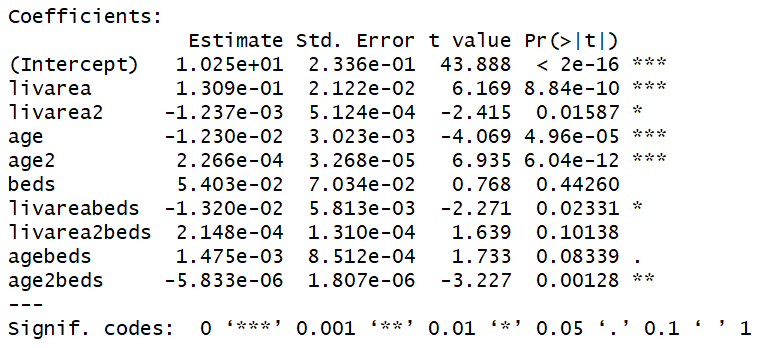
For 3 bedroom houses:

* The direct effect of 2 beds on ln(sprice) will on average increase by 0.16209
* The effect of livarea on ln(sprice) will decrease on average by 0.0396 and the effect of livarea2 on ln(sprice) will increase on average by 0.0006444.
* The effect of age on ln(sprice) will increase on average by 0.004425 and the effect of age2 on ln(sprice) will decrease on average by 0.0000017499.

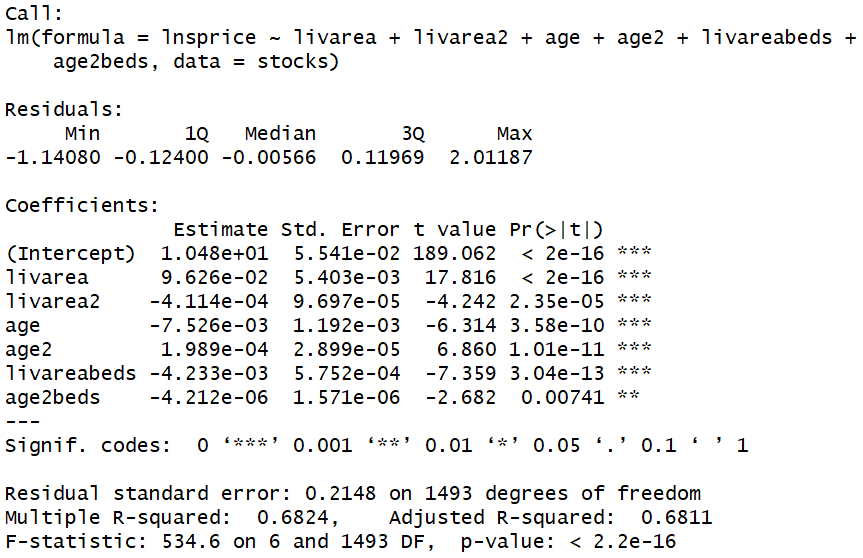
For 4 bedroom houses:

* The direct effect of 2 beds on ln(sprice) will on average increase by 0.21612
* The effect of livarea on ln(sprice) will decrease on average by 0.0528 and the effect of livarea2 on ln(sprice) will increase on average by 0.0008592.
* The effect of age on ln(sprice) will increase on average by 0.005900 and the effect of age2 on ln(sprice) will decrease on average by 0.0000023332.

From our regression results we can see the tests for the null hypotheses:

Since we are testing at 0.05 we fail to reject the null in all coefficients without a star and with a dot. Furthermore, we can see that these coefficients are also statistically insignificant because they are not greater than twice the std deviation.

Beds, livarea2beds, agebeds



For 2 bedroom houses:

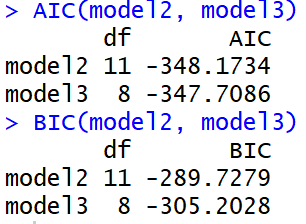
* The effect of livarea on ln(sprice) will decrease on average by 0.008466.
* The effect of age2 on ln(sprice) will decrease on average by 0.000008424.

For 3 bedroom houses:

* The effect of livarea on ln(sprice) will decrease on average by 0.012699.
* The effect of age2 on ln(sprice) will decrease on average by 0.000012636.

For 4 bedroom houses:

* The effect of livarea on ln(sprice) will decrease on average by 0.016932.
* The effect of age2 on ln(sprice) will decrease on average by 0.000016848.



AICmodel2<AICmodel3

Therefore, according to AIC the model from part (a) is preferred

SCmodel2>SCmodel3

Therefore, according to SC the model from part (c) is preferred

**Problem 7**

Solve problems 5.15 and 5.16 (page 274 and 276 from our text) by using copied and pasted output from SAS and conclusions for each part (5.15 a to e and 5.16 a to f)

5.15

a) No, there are some very large numbers and some very small ones. Obs. 3, 14, 16 have very small magnitude. Obs. 1, 2, 5 are slightly small. Obs. 12 is very large and Obs. 11 is large

b)



RStudent = -2.2241

Degrees of freedom = n – k – 2 = 10

Although it is within the confidence interval of the t test, it is barely on the edge, so we could consider it to be slightly influential.

c) Belsley, Kuh and Welsch (1980) suggest that points with a hat diagonal greater than 2p/n be considered high leverage points. Therefore, 2\*3/16 = 0.375, which means that 14, 15, and 16 are leverage point and should be treated differently.

d) Yes, because by removing hospital 14 all β’s changed and since hospital 14’s shift was opposite to 16’s then all β‘s shifted towards 16, reducing Cook’s Distance.

e) Yes, most Cook’s distances in the original output were orders of magnitude smaller than the D value of 17, but the cooks distances are not as variable in the smaller output.

5.16

a) The value of β4 is clearly statistically significant and represents that on the average being a big hospital will be more influential than almost anything else.

b) Yes, once the outliers are removed the data is homogenized and 14 is noticeably better at labor management than any other large hospital

c) Hospital 15. Yes, most Cook’s distances in the original output were orders of magnitude smaller than the D value of 17, but the cooks distances are not as variable in the lower dimensional output.

d) 17,030-15,175 = 1855 <- Which isn’t a surprise.

16,886 – 14,906 = 1980

17,618 – 14,511 = 3107

e) Figure 5.40 (b)

f) 5.16 captures more of the leverage of large hospitals than any other model and can better evaluate the efficiency of hospitals while controlling for scale. Also, few outliers.

Code

**Problem 1**

df <- p1data

names(df) <- c("height\_at\_18", "length\_at\_birth",

"mother\_ht\_at\_18", "father\_ht\_at\_18",

"mater\_grandma\_ht\_at\_18",

"mater\_grandpa\_ht\_at\_18",

"pater\_grandma\_ht\_at\_18",

"pater\_grandpa\_ht\_at\_18")

model <- lm(height\_at\_18 ~ length\_at\_birth + mother\_ht\_at\_18 +

father\_ht\_at\_18 + mater\_grandma\_ht\_at\_18 +

mater\_grandpa\_ht\_at\_18 + pater\_grandma\_ht\_at\_18 +

pater\_grandpa\_ht\_at\_18, data=df)

model

summary(model)

influence.measures(model)

cat("Mean of hat/leverage values: ",mean(hatvalues(model)))

plot(height\_at\_18 ~ length\_at\_birth +

mother\_ht\_at\_18 +

father\_ht\_at\_18 +

mater\_grandma\_ht\_at\_18 +

mater\_grandpa\_ht\_at\_18 +

pater\_grandma\_ht\_at\_18 +

pater\_grandpa\_ht\_at\_18,

data=df,cex=sqrt(cooks.distance(model)))

plot(fitted(model),resid(model))

plot(fitted(model),rstudent(model))

ggpairs(df)

**Problem 3**

drug <- c("A","A","A","A","A","A","A","A","A","A","A","A","A","A","A","A","A","A",

"A","A","A","A","A","A","A","A","A","A","A","A", "B", "B","B","B","B","B","B",

"B","B","B","B","B","B","B","B","B","B","B","B","B","B","B","B","B","B","B","B",

"B","B","B")

doses <- c(5,5,5,5,5,5,5,5,5,5,10,10,10,10,10,10,10,10,10,10,20,20,20,20,20,20,20,20,20,

20,5,5,5,5,5,5,5,5,5,5,10,10,10,10,10,10,10,10,10,10,20,20,20,20,20,20,20,20,

20,20)

score <- c(15, 16, 16, 15, 18, 16, 13, 17, 19, 15, 18, 16, 17, 15, 18, 19, 19, 18, 20, 16,

20, 17, 19, 18, 21, 21, 18, 20, 19, 17, 16, 15, 17, 15, 18, 18, 17, 17, 15, 16,

19, 18, 21, 20, 22, 21, 23, 22, 20, 19, 24, 23, 25, 24, 23, 22, 25, 26, 25, 24)

model <- lm(score ~ drug + doses + drug\*doses)

summary(model)

**Problem 4**

model <- lm(heart ~ ck, data=cklevel)

model

summary(model)

exp(cbind(OR = coef(model), confint(model)))

model1 <- glm(heart ~ ck, data=cklevel)

model1

summary(model1)

exp(cbind(OR = coef(model1), confint(model1)))

**Problem 5**

**proc** **print** data=VOTE;

**run**;

**proc** **reg** data=VOTE;

model voteA=prtystrA democA lexpendA lexpendB;

output out=b r=residual;

**run**;

**proc** **print** data=b;

**run**;

**proc** **reg** data=b;

model residual=prtystrA democA lexpendA lexpendB;

output out=b r=residual;

**run**;

**proc** **model** data=VOTE;

parms b0 b1 b2 b3 b4;

voteA=b0+b1\*prtystrA+b2\*democA+b3\*lexpendA+b4\*lexpendB;

fit voteA/white breusch=(**1** prtystrA democA lexpendA lexpendB) hccme=**0**;

**run**;

**proc** **print** data=smoke;

**run**;

**data** smoke;

set smoke;

smokes = ifn(cigs>**0**,**1**,**0**);

**run**;

**proc** **print** data=smoke;

**run**;

**proc** **robustreg** data=smoke;

model smokes=lcigpric lincome educ age agesq restaurn white;

**run**;

**proc** **reg** data=smoke;

model smokes=lcigpric lincome educ age agesq restaurn white /acov;

ods output ACovEst = estcov;

ods output ParameterEstimates=pest;

**run**;

**quit**;

**data** temp\_dm;

set estcov;

drop model dependent;

array a(**5**) intercept lcigpric lincome educ age agesq restaurn white;

array b(**5**) std1-std5;

b(\_n\_) = sqrt((**395**/**390**)\*a(\_n\_));

std = max(of std1-std5);

keep variable std;

**run**;

**proc** **sql**;

select pest.variable, estimate, stderr, tvalue, probt, std as robust\_stderr,

estimate/robust\_stderr as tvalue\_rb,

(**1** - probt(abs(estimate/robust\_stderr), **394**))\***2** as probt\_rb

from pest, temp\_dm

where pest.variable=temp\_dm.variable;

**quit**;

**Problem 6**

stocks <- stockton4

stocks$lnsprice <- log(stocks$sprice)

stocks$livarea2 <- stocks$livarea\*\*2

stocks$age2 <- stocks$age\*\*2

stocks$livareabeds <- stocks$livarea\*stocks$beds

stocks$livarea2beds <- stocks$livarea2\*stocks$beds

stocks$agebeds <- stocks$age\*stocks$beds

stocks$age2beds <- stocks$age2\*stocks$livarea

model1 <- lm(lnsprice ~ livarea + livarea\*\*2 + age + age\*\*2 + beds +

livarea\*beds + (livarea\*\*2)\*beds + age\*beds + (age\*\*2)\*beds,

data = stocks)

summary(model1)

model2 <- lm(lnsprice ~ livarea + livarea2 + age + age2 + beds +

livareabeds + livarea2beds + agebeds + age2beds,

data = stocks)

summary(model2)

model3 <- lm(lnsprice ~ livarea + livarea2 + age + age2 +

livareabeds + age2beds,

data = stocks)

summary(model3)

AIC(model2, model3)

BIC(model2, model3)