STAT 350

Assignment 5 Solutions

1. Suppose the two cars in our mileage problem are of the same make but that one vehicle was equipped with a special pollution control device. In 4 or 5 sentences comment on the experimental design as a method of determining whether or not the device reduces emissions and on what else you would want to find out from the experimenter to help interpret the results.

   The main difficulty with this design is that you only tried one car in each of the two conditions. As a result the observed differences could be due, for instance, to differences in the manufacturing process for the two cars. You would certainly want to know that you picked the car on which to have the pollution control device installed at random but you need more cars with and more cars without the device.

2. In this question you will derive some of the formulas for case deleted statistics. Suppose that X is an $n \times p$ design matrix. Let x_i^T be the i^{th} row of X (so that, x_i is a column vector or dimension p). Let $X_{(i)}$ be the design matrix with case i deleted and $Y_{(i)}$ the vector of $n-1$ responses with case i deleted.

   (a) Use partitioned matrices to show

   $$X_{(i)}^T Y_{(i)} = X^T Y - x_i Y_i \,.$$

   Write

   $$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

   and

   $$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

   Then

   $$X^T Y = [x_1 | x_2 | \cdots | x_n] \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{j=1}^{n} x_j Y_j$$

   Since $X_{(i)}$ and $Y_{(i)}$ simply omit the ith rows we see

   $$X_{(i)}^T Y_{(i)} = X^T Y - x_i Y_i \,.$$

(b) Show

$$X_{(i)}^T X_{(i)} = X^T X - x_i x_i^T \,.$$

We have

$$X^T X = [x_1|x_2|\cdots|x_n] \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \sum_{j=1}^{n} x_j x_j^T$$

Again omitting the ith term we find

$$X_{(i)}^T X_{(i)} = X^T X - x_i x_i^T \,.$$

(c) Suppose B is an invertible symmetric $p \times p$ matrix and v is a column vector of dimension p. Show, by direct multiplication, that $(B - vv^T)^{-1}$ is of the form

$$B^{-1} + rB^{-1}vv^T B^{-1}$$

and give a formula for the scalar r.

Simply multiply

$$(B - vv^T)(B^{-1} + rB^{-1}vv^T B^{-1}) = I + rvv^T B^{-1} - vv^T B^{-1} - rvv^T B^{-1}vv^T B^{-1}$$

The trick students often miss is that in the last term the quantity $v^T B^{-1} v$ is a scalar, that is, just a single number (check that it is a 1×1 matrix). It may help to give it a name like s to see that the last term is

$$rsvv^T B^{-1} \,.$$

Thus

$$(B - vv^T)(B^{-1} + rB^{-1}vv^T B^{-1}) = I + (r - 1 - rs)vv^T B^{-1}$$

which is the identity matrix if

$$r - rs - 1 = 0 \,.$$

Solving gives $r = 1/(1 - s) = 1/(1 - v^T B^{-1} v)$. Many students get a formula like

$$r = (1 - vv^T B^{-1})^{-1}$$

which is a $p \times p$ matrix not a scalar.

(d) Apply the previous part to show

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + r_i (X^T X)^{-1} x_i x_i^t (X^T X)^{-1}$$

and give a formula for r_i in terms of the leverage $h_{ii} = x_i^T (X^T X)^{-1} x_i$.

Put $B = X^T X$ and $v = x_i$. You get

$$r_i = 1/(1 - x_i^T (X^T X)^{-1} x_i) = 1/(1 - h_{ii})$$

and

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + r_i (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} \,.$$

(e) Show $1 + h_{ii}r_i = r_i$.

We have

$$(1 + r_i h_{ii}) = \frac{1 - h_{ii}}{1 - h_{ii}} + \frac{h_{ii}}{1 - h_{ii}} = \frac{1}{1 - h_{ii}} = r_i$$

(f) Show that

$$\begin{aligned} \hat{\beta}_{(i)} &= \hat{\beta} + r_i (X^T X)^{-1} x_i x_i^T \hat{\beta} - (1 + h_{ii} r_i)(X^T X)^{-1} x_i Y_i \\ &= \hat{\beta} - r_i (X^T X)^{-1} x_i \hat{\epsilon}_i \,. \end{aligned}$$

We have:

$$\begin{aligned} \hat{\beta}_{(i)} &= (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \\ &= \left[(X^T X)^{-1} + r_i (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} \right] \left[X^T Y - x_i Y_i \right] \\ &= \hat{\beta} + r_i (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} X^T Y \\ &\quad - (X^T X)^{-1} x_i Y_i - r_i (X^T X)^{-1} x_i x_i^T (X^T X)^{-1} x_i Y_i \\ &= \hat{\beta} + r_i (X^T X)^{-1} x_i x_i^T \hat{\beta} \\ &\quad - (X^T X)^{-1} x_i Y_i - r_i (X^T X)^{-1} x_i h_{ii} Y_i \\ &= \hat{\beta} + r_i (X^T X)^{-1} x_i x_i^T \hat{\beta} - (r_i h_{ii} + 1)(X^T X)^{-1} x_i Y_i \\ &= \hat{\beta} - r_i (X^T X)^{-1} x_i (Y_i - x_i^T \hat{\beta}) \\ &= \hat{\beta} - r_i (X^T X)^{-1} x_i \hat{\epsilon}_i \end{aligned}$$

(g) Deduce that $\hat{\mu}_{(i)} = \hat{\mu}_i - r_i h_{ii} \hat{\epsilon}_i$.

Multiply $\hat{\beta}_{(i)}$ by x_i to get

$$\hat{\mu}_{(i)} = \hat{\mu} - r_i x_i (X^T X)^{-1} x_i \hat{\epsilon}_i = \hat{\mu} - r_i h_{ii} \hat{\epsilon}_i$$

(h) Show that the i^{th} PRESS residual $Y_i - \hat{\mu}_{(i)}$ is given by

$$Y_i - \hat{\mu}_{(i)} = r_i \hat{\epsilon}_i$$

We have:

$$PRESS_i = Y_i - \hat{\mu}_{(i)} = Y_i - \hat{\mu}_i + r_i h_{ii} \hat{\epsilon}_i = (1 + r_i h_{ii}) \hat{\epsilon}_i = r_i \hat{\epsilon}_i$$

(i) Derive the formula for the i^{th} externally studentized (case deleted) residual.

The variance of the PRESS residual is $Var(r_i \hat{\epsilon}_i) = \sigma^2 r_i^2 (1 - h_{ii}) = \sigma^2 r_i$. The externally studentized residual is the PRESS residual divided by an estimate of its

standard error where σ^2 is estimated by $MSE_{(i)}$. Thus the externally studentized residual is

$$\frac{PRESS_i}{\sqrt{MSE_{(i)} r_i}} = \frac{\hat{\epsilon}_i \sqrt{r_i}}{\sqrt{MSE_{(i)}}} = \frac{\hat{\epsilon}_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

To get the simplest formula we must actually simplify $MSE_{(i)}$.

$$\begin{aligned} MSE_{(i)} &= \frac{\sum_{j\neq i}(Y_j - x_j^T \hat{\beta}_{(i)})^2}{(n-1)-p} \\ &= \frac{\sum_j (Y_j - x_j^T(\hat{\beta} - r_i (X^T X)^{-1} x_i \hat{\epsilon}_i))^2 - PRESS_i^2}{n-p-1} \\ &= \frac{\sum (Y_j - x_j^T \hat{\beta})^2}{n-p-1} + \frac{2 r_i \sum (Y_j - x_j^T \hat{\beta}) x_j^T (X^T X)^{-1} x_i}{n-p-1} \\ &\quad + \frac{r_i^2 x_i^T (X^T X)^{-1} \sum x_j x_j^T (X^T X)^{-1} x_i \hat{\epsilon}_i^2 - r_i^2 \hat{\epsilon}_i^2}{n-p-1} \\ &= \frac{ESS + 0 + r_i^2 x_i^t (X^T X)^{-1} x_i \hat{\epsilon}_i^2 - r_i^2 \hat{\epsilon}_i^2}{n-p-1} \end{aligned}$$

The middle term vanishes because

$$\sum (Y_j - x_j^T \hat{\beta}) x_j = X^T Y - X^T X \hat{\beta} = 0\,.$$

Use $r_i^2(1-h_{ii}) = r_i$ to get

$$MSE_{(i)} = \frac{ESS - r_i \hat{\epsilon}_i^2}{n-p-1}$$

and then you can deduce the final formula given in class.

3. Problem 22.11 parts a, b and c, 22.12 part c.

   For this question I used the SAS code

```
data knees;
  infile 'knees.dat' firstobs=2;
  input Days Fitness Patient Age ;
proc glm  data=knees;
  class Fitness ;
  model Days = Fitness|Age ;
proc glm  data=knees;
  class Fitness ;
  model Days = Fitness Age ;
  output out=anovres r=resid p=fitted;
proc rank data=anovres normal=blom out=ressc;
  var resid;
```

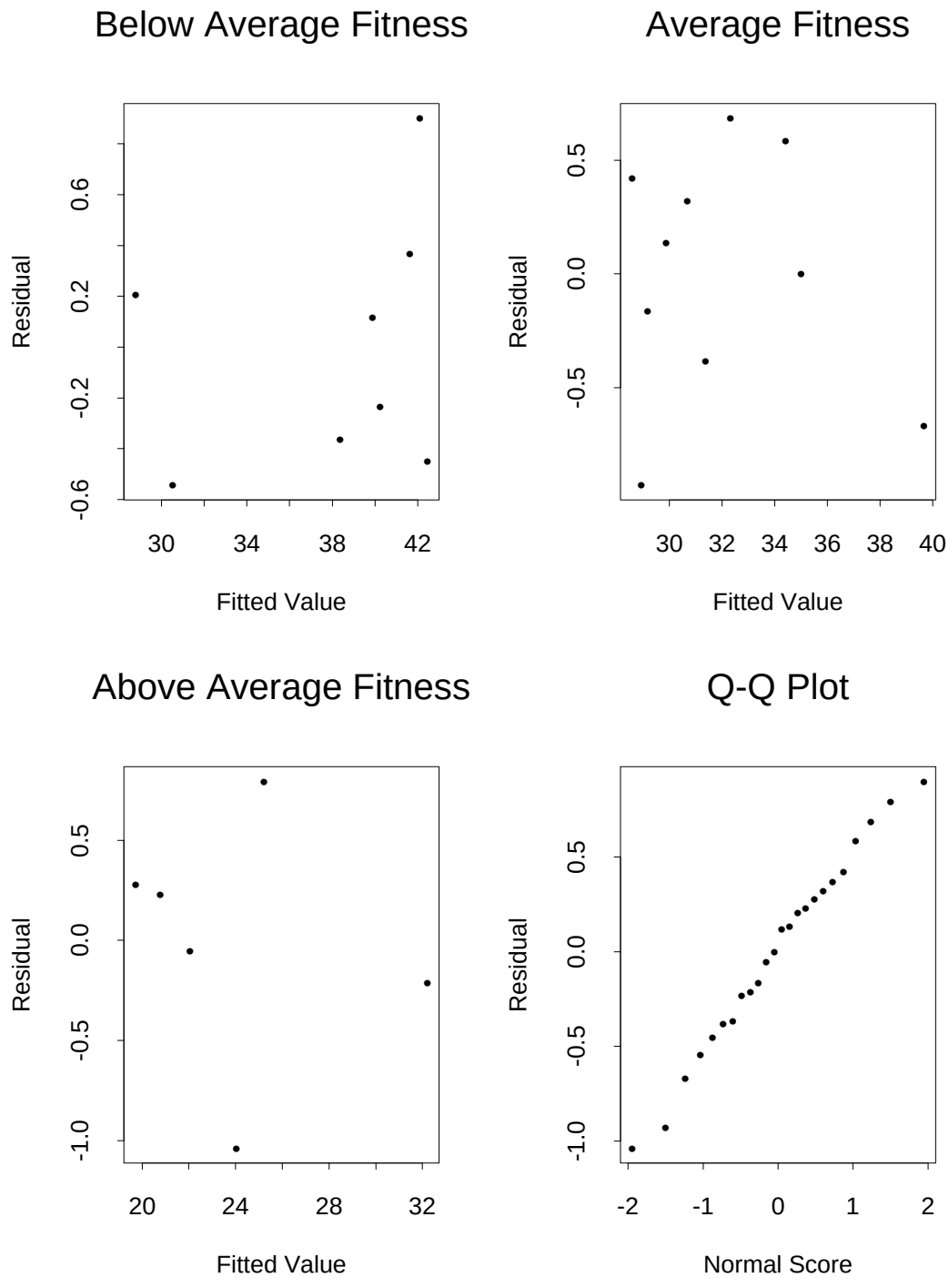
```
  ranks nscores;
proc corr data=ressc;
  var resid nscores;
proc print data=ressc;
  var fitted fitness age resid nscores;
run;
```

getting the output

```
                        General Linear Models Procedure
Dependent Variable: DAYS
                          Sum of        Mean
Source        DF      Squares      Square F Value Pr > F
Model          5 1082.0560870 216.4112174  655.36 0.0001
Error         18    5.9439130   0.3302174
Correctd Totl 23 1088.0000000
        R-Square         C.V.      Root MSE  DAYS Mean
        0.994537     1.795767     0.5746454  32.000000
Source      DF  Type III SS   Mean Square F Value Pr > F
FITNESS      2   5.44989183    2.72494592    8.25 0.0029
AGE          1 369.44147783  369.44147783 1118.78 0.0001
AGE*FITNESS  2   0.22183487    0.11091744    0.34 0.7191

                        General Linear Models Procedure
Dependent Variable: DAYS
                          Sum of        Mean
Source        DF      Squares      Square F Value Pr > F
Model          3 1081.8342521 360.6114174 1169.72 0.0001
Error         20    6.1657479   0.3082874
Correctd Totl 23 1088.0000000
    R-Square         C.V.      Root MSE      DAYS Mean
    0.994333     1.735114     0.5552363      32.000000
Source  DF  Type III SS   Mean Square F Value   Pr > F
FITNESS  2 246.08370505  123.04185252  399.11   0.0001
AGE      1 409.83425209  409.83425209 1329.39   0.0001

                   Correlation Analysis
          Pearson Correlation Coefficients

                     RESID
       NSCORES      0.99488

 OBS     FITTED    FITNESS     AGE      RESID      NSCORES
```

1	28.7930	1	18.3	0.20697	0.26136
2	42.4503	1	30.0	-0.45028	-0.87524
3	38.3648	1	26.5	-0.36478	-0.60318
4	40.2324	1	28.1	-0.23244	-0.48332
5	42.1001	1	29.7	0.89991	1.94690
6	39.8822	1	27.8	0.11775	0.05171
7	30.5440	1	19.8	-0.54396	-1.03865
8	41.6332	1	29.3	0.36682	0.73241
9	29.8639	2	20.8	0.13613	0.15568
10	34.9999	2	25.2	0.00007	-0.05171
11	39.6691	2	29.2	-0.66907	-1.23590
12	28.9300	2	20.0	-0.93004	-1.49843
13	30.6810	2	21.5	0.31903	0.60318
14	31.3813	2	22.1	-0.38134	-0.73241
15	28.5799	2	19.7	0.42015	0.87524
16	34.4163	2	24.7	0.58372	1.03865
17	29.1635	2	20.2	-0.16349	-0.26136
18	32.3152	2	22.9	0.68483	1.23590
19	25.2062	3	22.7	0.79380	1.49843
20	32.2099	3	28.7	-0.20991	-0.37006
21	20.7705	3	18.9	0.22949	0.37006
22	19.7199	3	18.0	0.28005	0.48332
23	24.0389	3	21.7	-1.03891	-1.94690
24	22.0545	3	20.0	-0.05452	-0.15568

For part a of 22.11 the residuals are printed out above. For part b the plots desired are:



The conclusion is that neither the residual plots nor the Q-Q plot show any major problems.

For part c, the generalized model is

$$Y_{ij} = \mu + \tau_i + \gamma_i(x_{ij} - \bar{x}_{..}) + \epsilon_{ij}$$

and the null hypothesis is that all the γ_i are equal. This is tested by comparing the two model statements `model days = fitness | age` *and* `model days = fitness age`*, doing an extra sum of squares F test. The resulting F statistic is the Type III sum of squares for fitness∗age giving $F = 0.34$ and $P = 0.7191$. The null hypothesis is accepted.*

For 22.12 part c the F statistic is obtained from the Type III sum of squares for fitness for the `model days = fitness age` *statement. This has $F = 399.11$ and $P = 0.0001$ There is clearly an effect of the variable Fitness.*

The `estimate` *statements permit us to compare the three levels of fitness. SAS prints out estimates of the differences in the intercepts. The relevant output is*

```
                                T for H0: Pr>|T| Std Err of
Parameter             Estimate  Param=0           Estimate
High v Low Fitness -8.72289277 -26.20   0.0001 0.33296397
High v Med Fitness -6.87551411 -23.84   0.0001 0.28837673
Med  v Low Fitness -1.84737866  -6.44   0.0001 0.28694289
```

Notice that the fit group appears to recuperate about 9 days faster than the unfit group!

4. Problem 24.12, 24.13 parts c,d,e,f only and 24.14. Note that 24.12 has only parts d and e in the text.

   Note: I should not have asked for part g.

   Here is code for the analysis of variance.

```
options pagesize=60 linesize=80;
data electron;
 infile 'anova.dat' firstobs=2;
 input Time Sex Sequence Exper Replic;
proc glm  data=electron;
 class Sex Sequence Exper ;
 model time = Sex|Sequence|Exper ;
 means sex sequence exper sex*sequence*exper;
 estimate 'sexdif' sex 1 -1;
 estimate 'seq12dif' sequence 1 -1 0;
 estimate 'seq13dif' sequence 1 0 -1;
 estimate 'seq23dif' sequence 0 1 -1;
 estimate 'expdif' exper 1 -1;
 output out=anovres r=resid;
proc rank data=anovres normal=blom out=ressc;
 var resid;
```

```
 ranks nscores;
proc corr data=ressc;
 var resid nscores;
run;
```

COMMENTS

- *You use the* `class` *statement to let SAS know that the factors are categorical.*
- *The* `model` *statement requests that all two way and the three way interaction terms be fitted.*
- *The* `means` *statement computes means and standard deviations for the groups defined by the levels of sex, of sequence, of exper and of all the 3 way combinations.*
- *The* `estimate` *statements compute the differences in means between the two sexes, between the 1st and second sequence, the first and third sequence and between the two levels of experience. This hinges on the use of a balanced design.*
- *The* `proc rank` *statements compute approximations to the expected values of normal order statistics needed for the correlation test in 23.12b.*
- *The correlation is actually computed by* `proc corr`*.*

```
                    General Linear Models Procedure
                        Class Level Information

                       Class    Levels    Values

                       SEX           2    1 2

                       SEQUENCE      3    1 2 3

                       EXPER         2    1 2


                    Number of observations in data set = 60


                    General Linear Models Procedure

Dependent Variable: TIME
                                     Sum of            Mean
Source                   DF          Squares          Square    F Value     Pr > F

Model                    11     973645.93333     88513.26667     103.16     0.0001
```

Error	48	41186.00000	858.04167	
Corrected Total	59	1014831.93333		

R-Square	C.V.	Root MSE	TIME Mean
0.959416	2.760738	29.292348	1061.0333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEX	1	540360.60000	540360.60000	629.76	0.0001
SEQUENCE	2	49319.63333	24659.81667	28.74	0.0001
SEX*SEQUENCE	2	542.50000	271.25000	0.32	0.7305
EXPER	1	382401.66667	382401.66667	445.67	0.0001
SEX*EXPER	1	91.26667	91.26667	0.11	0.7457
SEQUENCE*EXPER	2	911.23333	455.61667	0.53	0.5914
SEX*SEQUENCE*EXPER	2	19.03333	9.51667	0.01	0.9890

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	540360.60000	540360.60000	629.76	0.0001
SEQUENCE	2	49319.63333	24659.81667	28.74	0.0001
SEX*SEQUENCE	2	542.50000	271.25000	0.32	0.7305
EXPER	1	382401.66667	382401.66667	445.67	0.0001
SEX*EXPER	1	91.26667	91.26667	0.11	0.7457
SEQUENCE*EXPER	2	911.23333	455.61667	0.53	0.5914
SEX*SEQUENCE*EXPER	2	19.03333	9.51667	0.01	0.9890

Level of SEX	N	TIME Mean	TIME SD
1	30	1155.93333	92.2458983
2	30	966.13333	88.6102405

Level of SEQUENCE	N	TIME Mean	TIME SD
1	20	1044.15000	129.834419
2	20	1101.40000	128.286850
3	20	1037.55000	132.294916

Level of ----------TIME----------

EXPER	N	Mean	SD
1	30	1140.86667	104.700800
2	30	981.20000	104.142608

Level of SEX	Level of SEQUENCE	Level of EXPER	TIME N	TIME Mean	TIME SD
1	1	1	5	1218.60000	32.6389338
1	1	2	5	1051.00000	41.5752330
1	2	1	5	1274.20000	32.2676308
1	2	2	5	1122.40000	24.7951608
1	3	1	5	1218.20000	24.2837394
1	3	2	5	1051.20000	23.0911238
2	1	1	5	1036.40000	34.2169549
2	1	2	5	870.60000	16.7868997
2	2	1	5	1077.40000	37.0985175
2	2	2	5	931.60000	22.4677547
2	3	1	5	1020.40000	12.8569048
2	3	2	5	860.40000	34.9971427

Dependent Variable: TIME

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
sexdif	189.800000	25.10	0.0001	7.56325180
seq12dif	-57.250000	-6.18	0.0001	9.26305385
seq13dif	6.600000	0.71	0.4796	9.26305385
seq23dif	63.850000	6.89	0.0001	9.26305385
expdif	159.666667	21.11	0.0001	7.56325180

Correlation Analysis

2 'VAR' Variables: RESID NSCORES

Simple Statistics

Variable	N	Mean	Std Dev	Sum
RESID	60	0	26.420973	0
NSCORES	60	0	0.983909	0

Simple Statistics

Variable	Minimum	Maximum	Label
RESID	-55.000000	50.600000	
NSCORES	-2.312559	2.312559	RANK FOR VARIABLE RESID

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 60

	RESID	NSCORES
RESID	1.00000	0.99162
	0.0	0.0001
NSCORES	0.99162	1.00000
RANK FOR VARIABLE RESID	0.0001	0.0

You can read the type III sums of squares table to do F tests without doing multiple runs because each effect has a sum of squares which is unaffected by the presence of the others. The conclusion is that the three way interaction is insignificant and all three two way interactions are insignificant. All three main effects are significant; none can be eliminated.

In the question on Bonferroni confidence intervals the 5 quantities to be estimated are each estimated by a difference of two averages so that the variance of the estimated difference is of the form

$$\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})$$

You work out standard t type confidence intervals by estimating the means as usual (see the `means` *statment) and replacing σ^2 by the MSE in the formula for the variance. The Bonferroni method just replaces the α of 0.01 by $\alpha/5 = 0.02$ for 5 confidence intervals. The t critical value is $t_{0.02/2,48} = 2.41$. For the sex difference for instance the average time for men is 1155.933 seconds while for women it is 966.1333. The difference is 189.8 and this is computed by the means statement whose output is*

Parameter	Estimate	T for H0: Parameter=0	Pr>|T|	Std Error of Estimate
sexdif	189.800000	25.10	0.0001	7.56325180
seq12dif	-57.250000	-6.18	0.0001	9.26305385
seq13dif	6.600000	0.71	0.4796	9.26305385
seq23dif	63.850000	6.89	0.0001	9.26305385
expdif	159.666667	21.11	0.0001	7.56325180

Notice the standard errors. The mean squared error in the model is 858.06. There are 30 men and 30 women so the variance of the difference between men's average and

women's average is 858.06(1/30+1/30) = 57.204. The standard error of the estimate of $\mu_{1\cdot\cdot} - \mu_{2\cdot\cdot}$ *is then* $\sqrt{57.204} = 7.563$ *as in the output. The desired confidence intervals are all and estimate from the column labelled* `Estimate` *plus or minus 2.41 times a standard error from the last column.*

The needed mean for $\hat{\mu}_{231}$ *is produced by* `means`*. The key point is that the standard error to attach to the mean is* $\sqrt{MSE/5}$ *which is based on 48 degrees of freedom, not on 4.*

5. Question 10.7 and 10.11 parts a, b, d, and f.

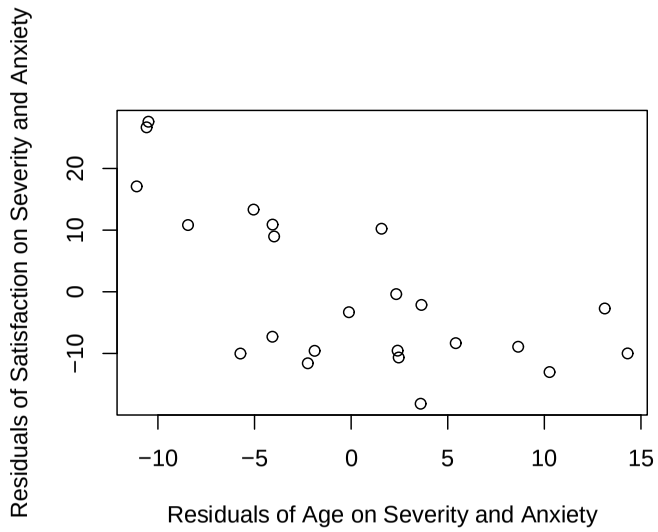
10.7 a *To prepare added variable plots we need to:*

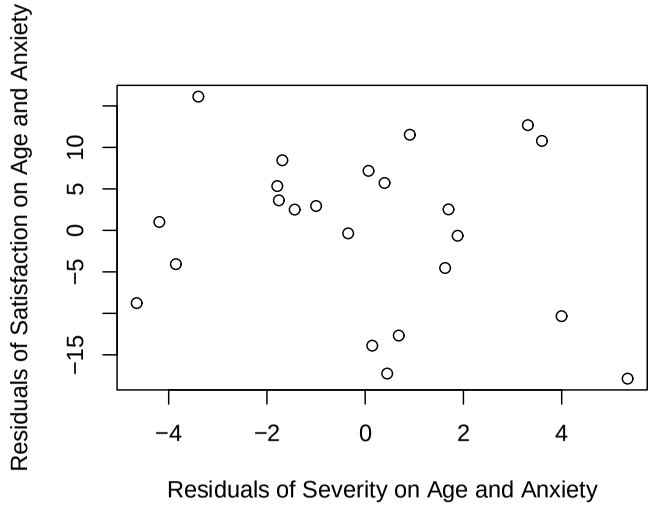
- *take each of the three dependent variables in turn out of the model*
- *regress the removed variable on the other two predictors and get residuals.*
- *regress the response variable on the same two predictors.*
- *plot the residuals against each other.*

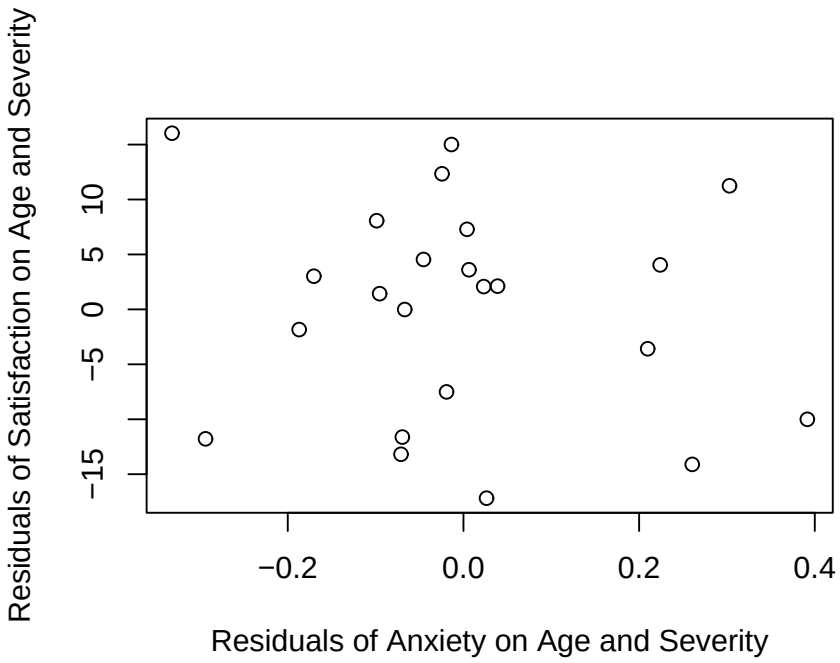
R-code

```
d=read.table("615.dat",header=T)
attach(d)
f1=lm(Satisfaction~Severity+Anxiety)
f1m=lm(Age~Severity+Anxiety)
y = residuals(f1)
x = residuals(f1m)
postscript("AgeAdded.ps",horizontal=F,height=4,width=5)
plot(x,y,xlab="Residuals of Age on Severity and Anxiety",
ylab="Residuals of Satisfaction on Severity and Anxiety")
dev.off()
f2=lm(Satisfaction~Age+Anxiety)
f2m=lm(Severity~Age+Anxiety)
y = residuals(f2)
x = residuals(f2m)
postscript("SeverityAdded.ps",horizontal=F,height=4,width=5)
plot(x,y,xlab="Residuals of Severity on Age and Anxiety",
ylab="Residuals of Satisfaction on Age and Anxiety")
dev.off()
f3=lm(Satisfaction~Age+Severity)
f3m=lm(Anxiety~Age+Severity)
y = residuals(f3)
x = residuals(f3m)
postscript("AnxietyAdded.ps",horizontal=F,height=4,width=5)
plot(x,y,xlab="Residuals of Anxiety on Age and Severity",
ylab="Residuals of Satisfaction on Age and Severity")
dev.off()
```

The plots are







I see a clear trend in the added variable plot for age and nothing much in the other two. I don't see any obvious curvature so the linear models we are using seem ok.

10.11 *SAS CODE*

```
options pagesize=60 linesize=80;
data patsat;
 infile '615.dat' firstobs=2;
 input Satisf Age Severity Anxiety ;
proc reg  data=patsat;
 model Satisf = Age Severity Anxiety /XPX I;
 output out=anovres r=resid p=fitted
    h=hat dffits=dffits cookd=cookd
    rstudent=rstudent press=press;
proc print data=anovres;
```

The output shows that

10.11 a *The largest externally studentized residual is for observation 14 at -1.81. This should be compared to the value* $t_{0.05/23,19} = 3.2$ *roughly. (I used the 0.9975 column; you really want 0.9978 so my critical point is a bit too small.) There are no surprising* Y *outliers.*

10.11 b *The largest leverage is 0.34 (for observation 9) which should, according to the text (p 377) be compared to 2(4)/23=.35 or so. This is not too large for such a small*

data set but it would probably warrant a quick look at this point and at case 15 whose leverage is 0.31.

10.11 c *You are supposed to compute* $x^T(X^TX)^{-1}x$ *when* $x^T = [1, 30, 58, 2]$. *I printed out the entries in* $(X^TX)^{-1}$ *(using the* `I` *option on the* `model` *statement. I used S to compute the desired leverage, getting 0.87 which is an unusually large leverage; I conclude that this would be a substantial extrapolation.*

10.11 d *The relevant lines of output are*

```
                  S                                                   R
                  E  A                                                S
     S            V  N          F                                     T          D
     A            E  X          I           R          C              P          U          F
     T            R  I          T           E          O              R          D          F
O    I  A         I  E          T           S          O         H    E          E          I
B    S  G         T  T          E           I          K         A    S          N          T
S    F  E         Y  Y          D           D          D         T    S          T          S

 14 51 34 51 2.3 67.9539 -16.9539 0.05661 0.07185 -18.2663 -1.80980 -0.50353
```

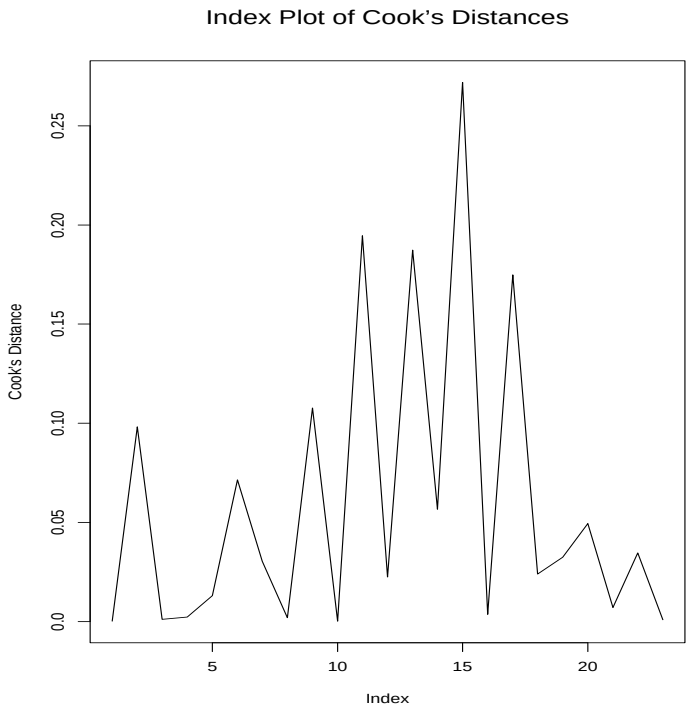
The values of DFFITS and COOKD are not too large; see the notes on Diagnostics for guidelines. I conclude this data point is ok.

10.11 e *I did this partly in S. You need to compute all the fitted values with case 14 removed; the easiest way is to delete case 14 from the data file and rerun. You get the predicted value for case 14 by subtracting the PRESS residual for case 14 from the true Y for case 14. I got*

$$\sum |\hat{\mu}_j - \hat{\mu}_{j(i)}|/|\hat{\mu}_j| = 1.4\%$$

which seems pretty minor.

10.11 f *You are to plot* D_i *against* i. *The result is*



Case 15 looks a bit surprising and should probably be investigated.

6. In this problem you will prove that

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty$$

is a density.

(a) Let $I = \int_{-\infty}^{\infty} \phi(x)dx$. Show that

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x)\phi(y)\, dx\, dy.$$

HINT: What is $\int_{-\infty}^{\infty} \phi(y)dy$ in terms of I.
We have

$$\begin{aligned} I^2 &= I \int \phi(y)dy \qquad \textit{by change of variables} \\ &= \int I\phi(y)dy \\ &= \int\int \phi(x)dx\phi(y)dy \\ &= \int\int \phi(x)\phi(y)dxdy \\ &= J \end{aligned}$$

(b) Now if

$$J = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dx \, dy$$

do the double integral J in polar co-ordinates ($x = r\cos\theta$, $y = r\sin\theta$) to show $J = 1$.

When you do an integral in polar co-ordinates you have to: replace each x in the integrand with $r\cos\theta$ and each y with $r\sin\theta$, replace $dxdy$ with $|Jacobian|drd\theta$, and find the set of r, θ values which correspond to the set of x, y values over which we are integrating. The Jacobian is the absolute value of the determinant filled up with derivatives of (x, y) with respect to r and θ. This 2 by 2 matrix has determinant r. The value of r, being a distance from the origin is in the range 0 *to* ∞ *while the angle in the plane is measured over any interval of length* 2π *such as* $[0, 2\pi)$*. This makes*

$$J = \int_0^{\infty} \int_0^{2\pi} \frac{1}{2\pi} \exp(-r^2/2) r dr d\theta \, .$$

The θ integral gives2π leaving

$$J = \int_0^{\infty} r \exp(-r^2/2) \, dr = 1 \, .$$

(c) Deduce that ϕ is a density.

All you have to do is prove that $\phi \geq 0$ and $\int \phi = 1$. But ϕ is clearly positive. Thus $I > 0$ and since $I^2 = J = 1$ we have $I = 1$.