STAT 350

Assignment 4

Part A

From text page 254-255, 6.15 c, d, e, f, g, 6.16, 6.17. Page 257 6.25 and 6.26. Page 324 7.46. Page 394 9.11. Page 398 9.25.

6.15-17 I began with this SAS code:

```
data patsat;
infile '615.dat' firstobs=2;
input Satisf Age Severity Anxiety ;
proc glm  data=patsat;
 model Satisf = Age Severity Anxiety ;
 estimate '617a' Intercept 1 Age 35 Severity 45 Anxiety 2.2;
 output out=anovres r=resid p=fitted;
proc print data=anovres;
```

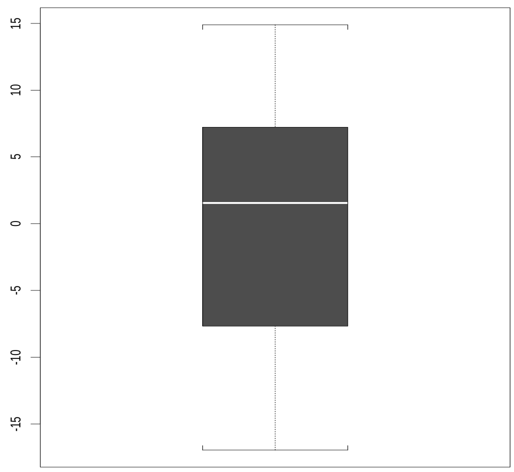
This code produces the anova table, t tests for individual coefficients and the estimates required for predicted values; it also prints out residuals for use later. The output shows:

6.15 c The fitted regression function is

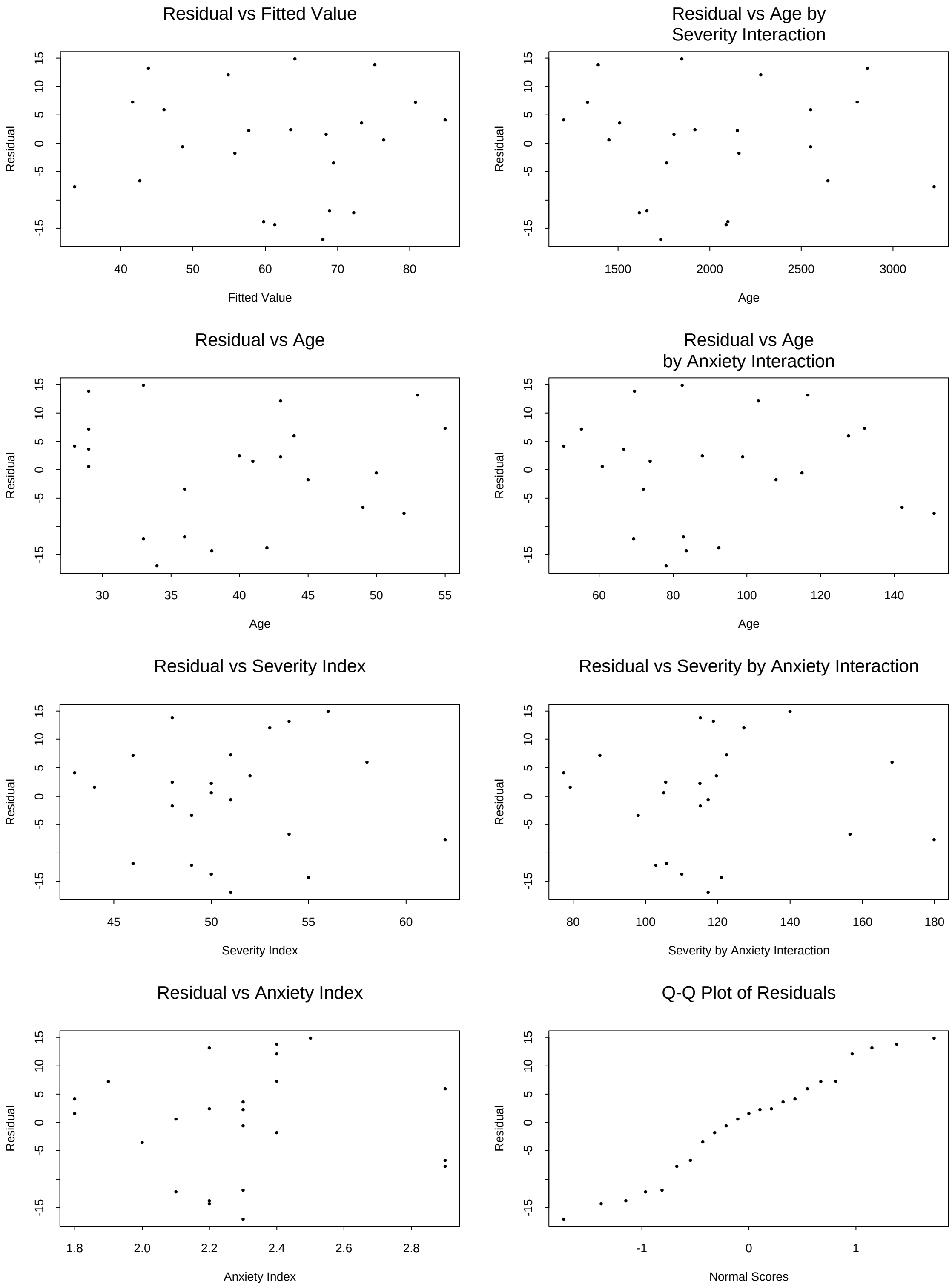
$$\hat{\mu} = 162.88 - 1.210X_1 - 0.666X_2 - 8.613X_3$$

The question asks for an interpretation of b_2. Mathematically it means that holding Age and Anxiety constant an increase of 1 unit in severity of disease is associated with an average decrease of 0.666 units in Satisfaction. The book wants you, however, to think about the **real world** interpretation. Patients with more severe illnesses are less satisfied with the hospital, after adjusting for Age and Anxiety level. Whether the amount less is a lot or a little depends on the units in which severity and satisfaction are measured and since these are indices we cannot really tell.

6.15 d Here is a box plot of the (raw) residuals from Splus; I see no problem with outliers.



6.15 e Here is a set of plots from Splus



There seems to be no particular problem in any of the plots. The plots show no need for inclusion of the interactions terms and no sign of non-normality.

6.15 f You need replicate observations to compute a pure error sum of squares and you don't have any such. Sometimes people try a clustering technique to split the data set into groups of 'near replicates' and then treating these groups as groups of replicates but the technique doesn't work all that well.

6.15 g You have to look in the text for this one. The test regresses squared residuals on the covariates and computes a χ^2 statistic which looks a lot like an F test (because it was intended to be analogous to such an F test) except for the numerator not being divided by degrees of freedom and the denominator being somewhat different; see page 115 and page 239. I used the code above to print out a data set which includes the needed residuals. I saved the results in a file, deleting all the extra output, and then ran this SAS code:

```
options pagesize=60 linesize=80;
data patsatr;
 infile '615res.dat' firstobs=2;
 input Obs Satisf Age Severity Anxiety Resid Fitted;
 rsq=Resid**2;
proc glm  data=patsatr;
  model rsq = Age Severity Anxiety ;
run;
```

You take the Model Sum of Squares from the output which is 24518 and the Error Sum of Squares from the original output which is 2011.6 and compute

$$\chi^2 = [24518/2]/[2011.6/23]^2 = 1.60$$

From table B 3 we see the P-value is between 0.1 and 0.9 (Splus gives a P value of 0.65) so that there is no evidence of heteroscedasticity related to the values of the covariates.

6.16 a The overall F statistic is 13.01 with a P-value of 0.0001 so the hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$ is rejected at the level 0.1 and, indeed, at any level down to 0.0001. The test implies that at least one of the three coefficients is not 0.

6.16 b The text intended a joint interval using the Bonferroni procedure: estimate plus or minus $t_{0.05/3,19}$ times estimated standard errors. The estimates and estimated standard errors are in the SAS output

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	162.8758987	6.32	0.0001	25.77565190
AGE	-1.2103182	-4.01	0.0007	0.30145159
SEVERITY	-0.6659056	-0.81	0.4274	0.82099695
ANXIETY	-8.6130315	-0.70	0.4902	12.24125126

The required t critical value is 2.29; you would need to interpolate in the tables page 1337 between 0.98 and 0.985 since the lower tail area you actually want is 1-0.05/3=0.98333. Go 2/3 of the way from 2.205 to 2.346. I actually used Splus.

6.16 c From the output the value of R^2 is 0.67. We sometimes describe this as meaning that 2/3 of the variance in patient satisfaction is accounted for by these three covariates. This is a fairly high but not wonderful multiple correlation.
6.17 a The output of the estimate statement is

```
                              T for H0:    Pr > |T|   Std Error of
Parameter       Estimate     Parameter=0                Estimate
617a          71.6003409           16.11     0.0001     4.44322423
```

so that the estimate is $71.6 \pm 1.729(4.44)$. If you want to predict an individual observation, though, as in b) you have to take a standard error of the form $\sqrt{4.44^2 + \hat{\sigma}^2} = \sqrt{19.71 + 105.87} = 11$. Notice that the prediction interval is much wider. For a new individual with covariate values 35, 45 and 2.2, there is roughly a 90% chance that the satisfaction level will be in the range $71.6 \pm 1.73(11)$.
9.11 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
9.11 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.
9.11 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.
9.11 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.
9.11 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
```

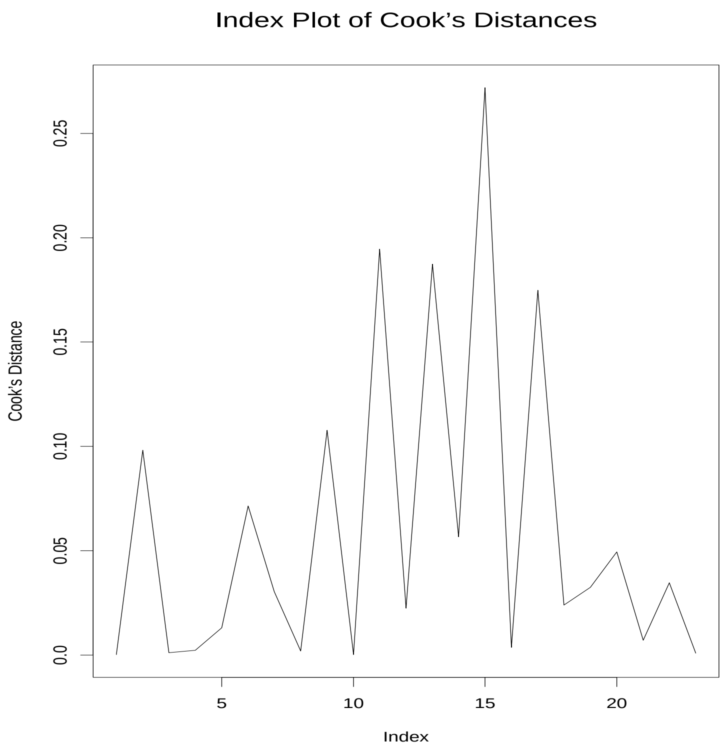
The values of DFFITS and COOKD are not too large; see Lecture 21 for guidelines. I conclude this data point is ok.

9.11 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.

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



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

6.25 You are to choose the βs to minimize

$$\sum (Y_i - 4X_{i,2} - \beta_0 - \beta_1 X_{i,1} - \beta_3 X_{i,3})^2$$

which simply is ordinary least squares with response variable

$$Y_i - 4X_{i,2}$$

and 3 parameters to be adjusted. In SAS you would create a variable having $Y_i - 4X_{i,2}$ and regress it on $X_{i,1}$ and $X_{i,3}$.

6.26 The multiple correlation coefficient is

$$R^2 = 1 - \frac{ESS}{TotalSS} = \frac{RegSS}{TotalSS}.$$

Now $ESS = \sum(Y_i - \hat{\mu}_i)^2$ and $TotalSS = \sum(Y_i - \bar{Y})^2$. If we regress Y_i on $\hat{\mu}_i$ then the correlation coefficient is

$$r = \frac{\sum(Y_i - \bar{Y})(\hat{\mu}_i - \bar{\hat{\mu}}_i)}{\sqrt{\sum(Y_i - \bar{Y})^2 \sum(\hat{\mu}_i - \bar{\hat{\mu}}_i)^2}}$$

Squaring r we see that we should prove

$$(RegSS)\sum(\hat{\mu}_i - \bar{\hat{\mu}}_i)^2 = \left[\sum(Y_i - \bar{Y})(\hat{\mu}_i - \bar{\hat{\mu}})\right]^2$$

In lecture 4 or so I showed you that the Regression Sum of Squares is

$$RegSS = \sum(\hat{\mu}_i - \bar{\hat{\mu}}_i)^2$$

Now

$$\begin{aligned}
\sum(Y_i - \bar{Y})(\hat{\mu}_i - \bar{\hat{\mu}}) &= \sum(Y_i - \hat{\mu}_i + \hat{\mu}_i - \bar{Y})(\hat{\mu}_i - \bar{\hat{\mu}}) \\
&= \sum(Y_i - \hat{\mu}_i)\hat{\mu}_i + \sum(\hat{\mu}_i - \bar{Y})(\hat{\mu}_i - \bar{\hat{\mu}}) \\
&\quad -\bar{\hat{\mu}}\sum(Y_i - \hat{\mu}_i)
\end{aligned}$$

The first term is 0 because the vector of residuals is orthogonal to the fitted vector. The last term is 0 because the sum of the residuals is 0 in any model with an intercept. In the middle term $\bar{Y} = \bar{\hat{\mu}}$ because $\sum Y_i - \sum \hat{\mu}_i = \sum(Y_i - \hat{\mu}_i) = 0$. Hence

$$\sum(Y_i - \bar{Y})(\hat{\mu}_i - \bar{\hat{\mu}}) = \sum(\hat{\mu}_i - \bar{\hat{\mu}}_i)^2$$

and this finishes the problem.

7.46 The reduced models are

a)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$$

b)

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_4\sqrt{X_{i,1}} + \epsilon_i$$

c)

$$Y_i - 5X_{i1} - 5X_{i2} = \beta_0 + \beta_3 X_{i1}X_{i2} + \beta_4\sqrt{X_{i,1}} + \epsilon_i$$

d)

$$Y_i - 7\sqrt{X_{i,1}} = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i1}X_{i2} + \epsilon_i$$

9.25 If X is invertible then $(X^TX)^{-1} = X^{-1}[X^T]^{-1}$ so that

$$H = X(X^TX)^{-1}X^T = XX^{-1}[X^T]^{-1}X^T = I^2 = I.$$

The diagonal elements of the identity matrix are all 1 and $\hat{Y} = HY = Y$ so that $\hat{Y}_i = Y_i$.

Part B

1. In assignment 2 in question 2 you dealt with variables $Y_1, \ldots, Y_4$. In class I stated that the if the covariance between two components of a multivariate normal vector is 0 then the components are independent, but I indicated a proof only when the multivariate normal distribution in question has a density. In this case the variance matrix is singular so there is no density. However, in terms of the original Z it is possible to find two independent functions of Z such that Y_1, Y_2, Y_3 are a function of the first function while Y_4 is a function of the second.

   (a) Let $U_1 = (2Z_1 - Z_2 - Z_3)/3$, $U_2 = (2Z_2 - Z_1 - Z_3)/3$ and $U_3 = (Z_1 + Z_2 + Z_3)/3$. Show that $U = (U_1, U_2, U_3)^T$ has a multivariate normal distribution and identify the mean and variance of U.

   Solution: We have $U = AZ$ where the matrix A is

$$A = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

   Thus U has a multivariate normal distribution with mean $A0 = 0$ and variance AA^T which may be multiplied out to give

$$AA^T = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$$

   (b) Use the result in class, for multivariate normals which have a density to show that (U_1, U_2) is independent of U_3.

   Solution: The variance covariance matrix in the previous part is block diagonal with a 2 by 2 block and a 1 by 1 block, so the first two components are independent of the last component.

   (c) Express Y_3 as a function of U.

   Solution: $Y_3 = X_3 - \bar{X} = \sigma(Z_3 - \bar{Z})$. Now $(Z_1 - \bar{Z}) + (Z_2 - \bar{Z}) + (Z_3 - \bar{Z}) = 0$ so

$$Z_3 - \bar{Z} = -U_1 - U_2$$

   and

$$Y_3 = -\sigma(U_1 + U_2).$$

   (d) Use the fact that if X_1 and X_2 are independent then so are $G(X_1)$ and $H(X_2)$ for any functions G and H to show that Y_1, Y_2, Y_3 is independent of Y_4.

   Solution: The function G is

$$G(U_1, U_2) = [\sigma U_1, \sigma U_2, -\sigma(U_1 + U_2)] = [Y_1, Y_2, Y_3]$$

   while

$$H(U_3) = \sigma U_3 = \bar{X} = Y_4$$

   Since U_3 is independent of U_1, U_2 we see Y_4 is independent of Y_1, Y_2, Y_3.

(e) Express the sample variance of the $X_i, i = 1, 2, 3$ in terms of U and use this to show that $(n-1)s_X^2/\sigma^2$ has a χ^2 distribution on 2 degrees of freedom.

Solution: This question is wrong! You can write

$$[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2]/2$$

as

$$\sigma^2(U_1^2 + U_2^2 + (-(U_1 + U_2))^2)$$

but this is not a straightforward sum of two squares of standard normals as written. The trick to do the real question is to write the numerator in the sample variance for a sample of 3 as the sum of the numerator for the sample variance of the sample X_1, X_2 of size 2 plus another term. These two terms turn out to be independent χ^2 variables except for a factor of σ^2.

2. In class I discussed the general formula for a multivariate normal density. Suppose that Z_1 and Z_2 are independent standard normal variables. Assume that $X_1 = aZ_1 + bZ_2 + c$ and $X_2 = dZ_1 + eZ_2 + f$. Find the joint density of X_1 and X_2 by evaluating the formulas I gave in class. Express $P(X_1 \le t)$ as a double integral. I want to see the integrand and the limits of integration but you need not try to do the integral.

   Solution: The density in class was

$$\frac{1}{2\pi\sqrt{det(AA^T)}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

   where $\Sigma = AA^T$.

   The entries in μ are c and f and we have

$$A = \begin{bmatrix} a & b \\ d & e \end{bmatrix}$$

   Then

$$\Sigma = AA^T = \begin{bmatrix} a^2 + b^2 & ad + be \\ ad + be & d^2 + e^2 \end{bmatrix}$$

   and

$$(AA^T)^{-1} = \begin{bmatrix} \frac{d^2+e^2}{(ae-bd)^2} & -\frac{ad+be}{(ae-bd)^2} \\ -\frac{ad+be}{(ae-bd)^2} & \frac{a^2+b^2}{(ae-bd)^2} \end{bmatrix}$$

   Putting together all the algebra gives

$$f(x_1, x_2) = \frac{1}{2\pi|ae - bd|} \exp\left[-\frac{1}{2}\frac{q(x_1, x_2, a, b, c, d, e, f)}{(ae-bd)^2}\right].$$

   where the exponent is the quadratic function

$$\begin{aligned} q(x_1, x_2, a, b, c, d, e, f) &= [(x_1 - c)^2(d^2 + e^2) \\ &\quad -2(ad + be)(x_1 - c)(x_2 - f) \\ &\quad +(x_2 - f)^2(a^2 + b^2)]. \end{aligned}$$

To compute $P(X_1 \le t)$ we take the joint density of X_1 and X_2 and integrate it over the set of (x_1, x_2) such that $x_1 \le t$ to get

$$P(X_1 \le t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t} f(x_1, x_2) dx_1 dx_2 \,.$$