STAT 350

Assignment 4

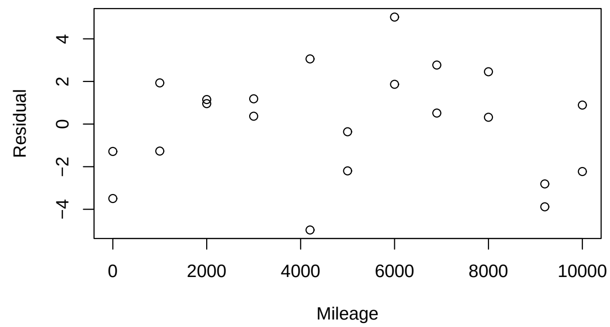
1. For the Mileage data in assignment 3 conduct a residual analysis and report your findings.

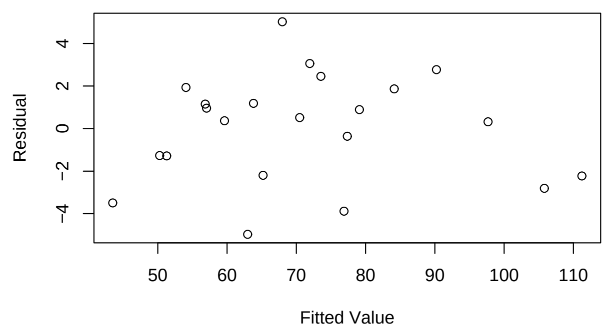
   I used the full model for this since my answers to assignment 3 suggested we needed the full two lines model. I plotted residuals versus mileage and residuals versus fitted value. Then I plotted squared residuals against mileage. I am looking to see no relation between the y and x values in these plots; I don't see any such so I don't see any problems here. I also made a Q-Q plot which seems quite adequately straight; I conclude the residuals are reasonably normally distributed.

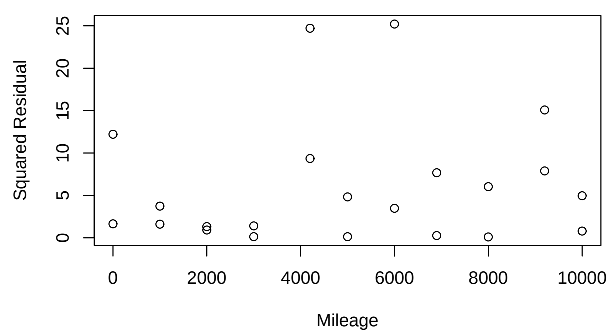
   The R code is

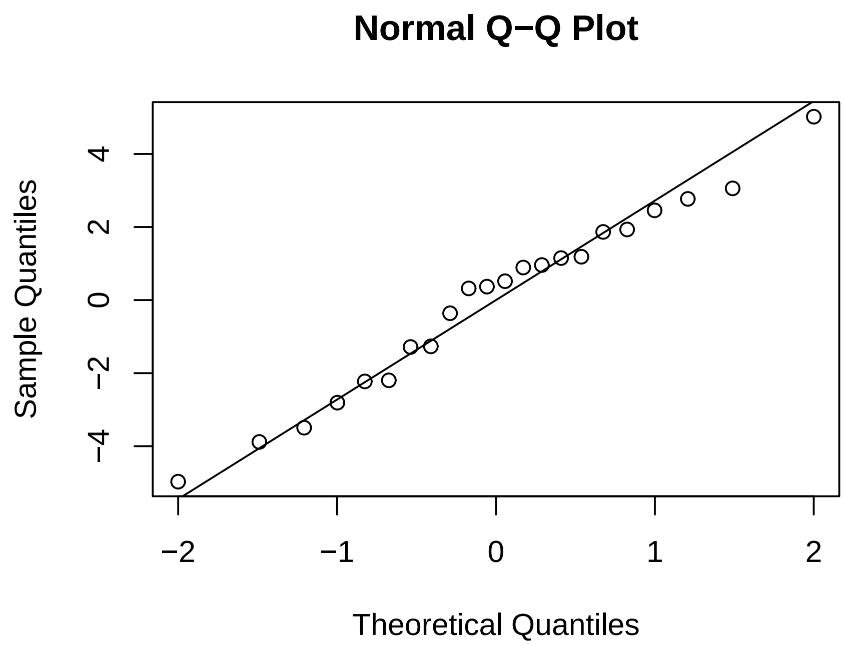
```
d = matrix(scan("mileage.dat"),ncol=4,byrow=T)
y = c(d[,2],d[,4])
x1 =c(d[,1],rep(0,11))
int1 = c(rep(1,11),rep(0,11))
int2 = c(rep(0,11),rep(1,11))
x2 =c(rep(0,11),d[,1])
fit = lm(y~int1+x1+int2+x2-1)
r = residuals(fit)
fv = fitted(fit)
postscript("r_vs_mileage.ps",width=6,height=4)
plot(x1+x2,r,xlab="Mileage",ylab="Residual")
dev.off()
postscript("r_vs_fitted.ps",width=6,height=4)
plot(fv,r,xlab="Fitted Value",ylab="Residual")
dev.off()
postscript("rsq_vs_mileage.ps",width=6,height=4)
plot(x1+x2,r^2,xlab="Mileage",ylab="Squared Residual")
dev.off()
postscript("QQMileage.ps",horizontal=F,width=5,height=4)
> qqnorm(residuals(fit))
> abline(a=0,b=summary(fit)$sigma)
> dev.off()
```

   The plots are


Residual
−4
−2
0
2
4
0
2000
4000
6000
8000
10000
Mileage


Residual
−4
−2
0
2
4
50
60
70
80
90
100
110
Fitted Value


Squared Residual
0
5
10
15
20
25
0
2000
4000
6000
8000
10000
Mileage



2. For the Wallabies data in assignment 3 conduct a residual analysis and report your findings. If the residual analysis suggests any refitting do that and report the results.

   For this problem I am going to USE SAS to compute residuals, internally studentized residuals, externally studentized residuals and PRESS residuals. Then I will plot then using R. My final fitted model in assignment 3 regressed `nitexc` *on only* `nitin` *so that is the model here.*

```
proc glm  data=nit;
   model nitexc = nitin ;
   output out=outdat r=resid
   student=isr press=press rstudent=rsr ;
run ;
proc print data=outdat;
run ;
```

   The `proc print` *produces*

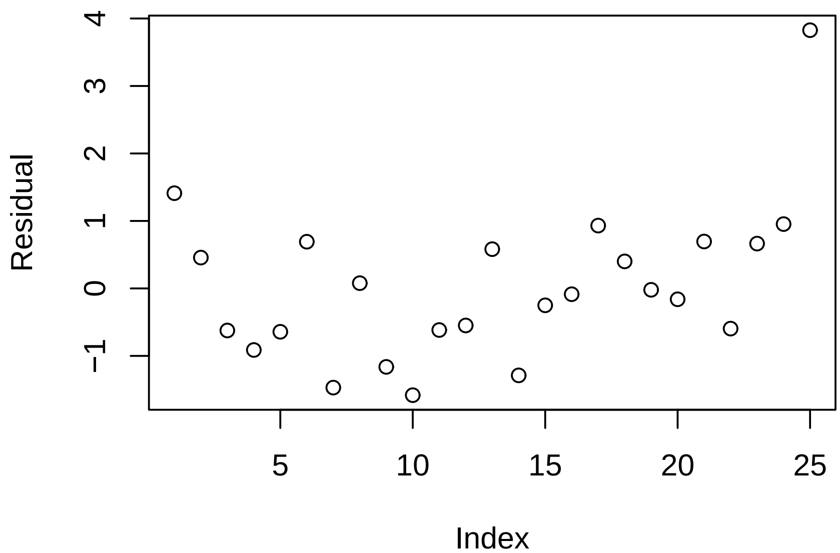
Obs	nitexc	weight	dryin	wetin	nitin	resid	isr	press	rsr
1	162	3386	166	417	54	25.9783	1.38176	28.6541	1.41123
2	174	3033	181	409	99	8.8658	0.46463	9.4936	0.45657
3	119	3477	134	250	46	-11.8461	-0.63212	-13.1511	-0.62367
4	205	3278	226	392	188	-17.7124	-0.91582	-18.4619	-0.91248

5	312	3368	265	474	345	-12.2829	-0.65116	-13.4588	-0.64280
6	157	2932	214	516	66	13.2150	0.69973	14.4453	0.69175
7	184	3128	303	716	171	-27.7143	-1.43444	-28.9464	-1.47021
8	155	3251	176	271	81	1.5108	0.07959	1.6349	0.07785
9	192	3396	213	377	175	-22.3021	-1.15396	-23.2793	-1.16276
10	331	3497	299	505	399	-28.2179	-1.53263	-32.4551	-1.58189
11	114	3182	128	284	38	-11.6705	-0.62489	-13.0454	-0.61641
12	159	3234	196	343	106	-10.6628	-0.55780	-11.3768	-0.54927
13	260	3139	362	776	228	11.4098	0.59019	11.9024	0.58163
14	265	3434	350	589	291	-24.3478	-1.27052	-25.8486	-1.28864
15	387	2970	329	553	449	-4.5652	-0.25577	-5.5871	-0.25051
16	146	3230	229	462	72	-1.6667	-0.08806	-1.8142	-0.08614
17	233	3470	329	674	176	18.0510	0.93393	18.8393	0.93123
18	261	3000	357	771	235	7.8812	0.40786	8.2293	0.40035
19	287	3224	344	749	288	-0.4070	-0.02122	-0.4315	-0.02076
20	412	3366	362	607	485	-2.8553	-0.16457	-3.6983	-0.16105
21	174	3264	299	654	92	13.3944	0.70332	14.3985	0.69538
22	171	3292	217	512	126	-11.6017	-0.60423	-12.2693	-0.59570
23	259	3525	350	668	224	12.9976	0.67217	13.5528	0.66395
24	298	3036	297	658	276	18.3564	0.95500	19.3711	0.95310
25	407	3356	292	481	386	56.1924	3.03176	63.7746	3.82679

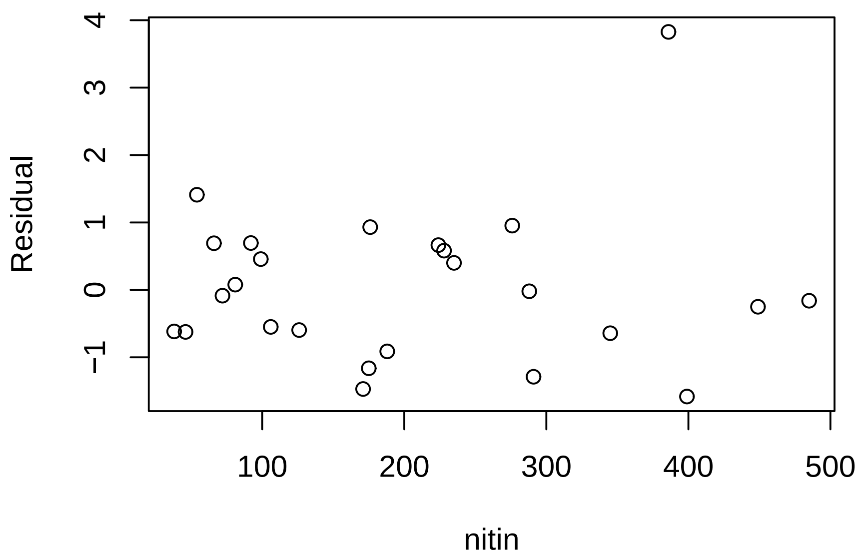
It will be seen that the externally studentized residual for observation 25 is very large. The model clearly fits badly for this data point. That outlier is visible in all the plots below. Except for that outlier I see no other problems. It would be best to rerun the fit without that data point – I asked you to do that but I won't do that here.

Now for some plots:

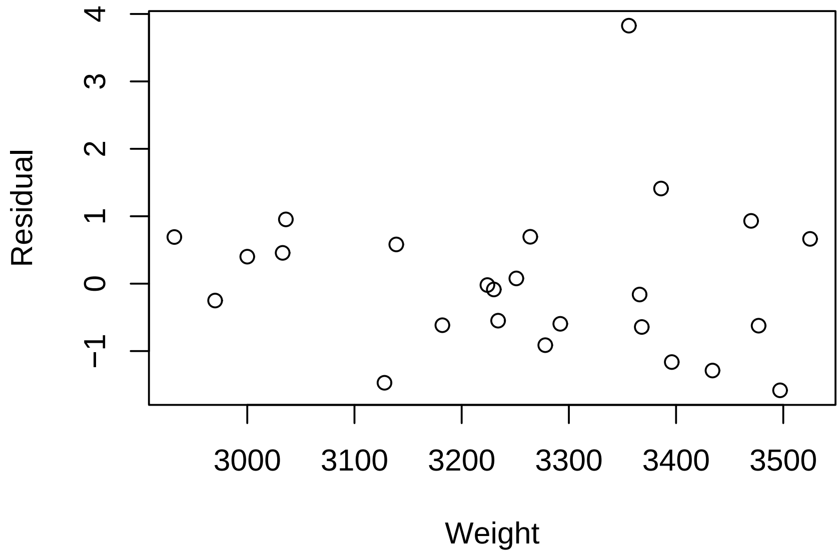
Index Plot of Externally Studentized Residuals



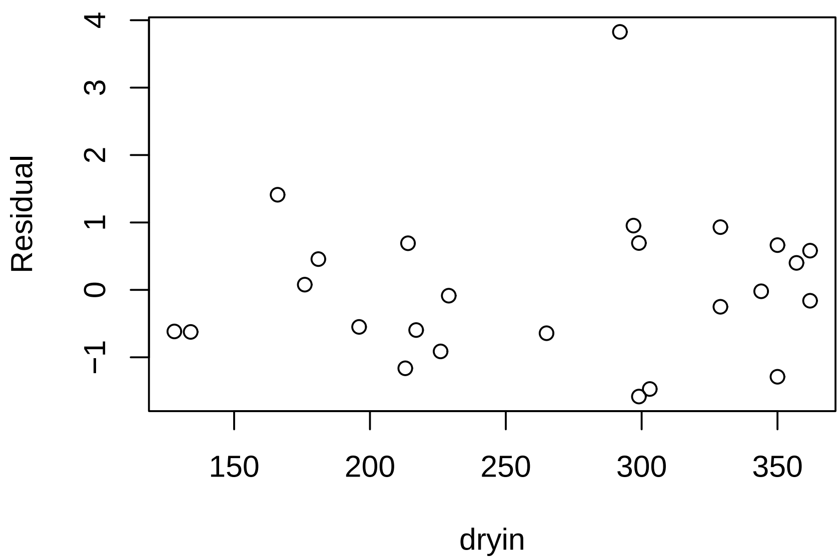
Plot of Externally Studentized Residuals against nitin



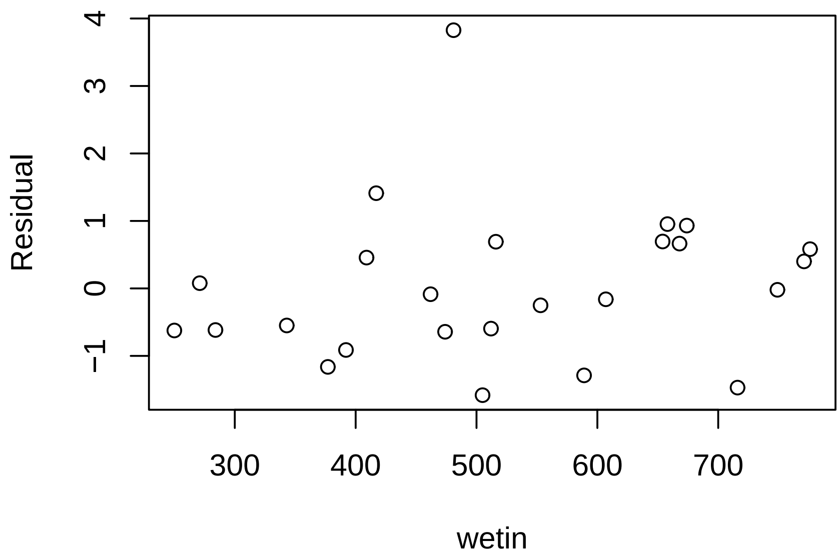
Plot of Externally Studentized Residuals against weight



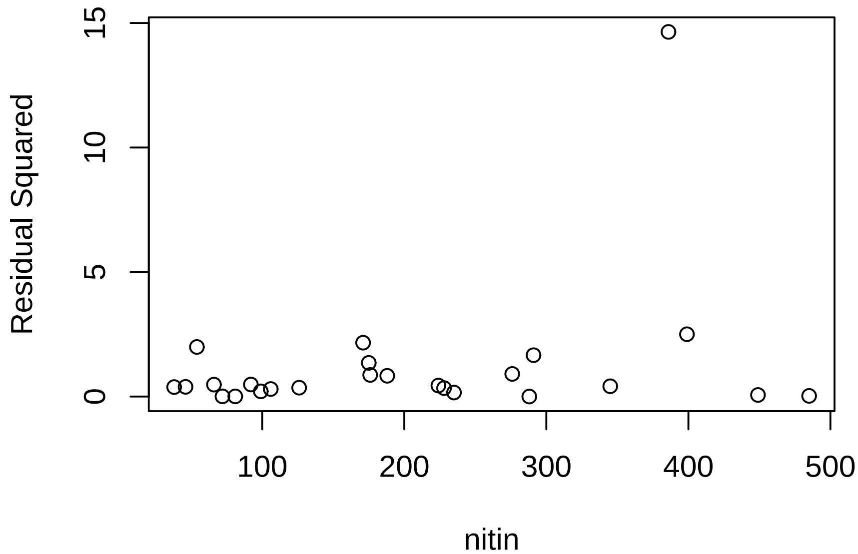
Plot of Externally Studentized Residuals against dryin



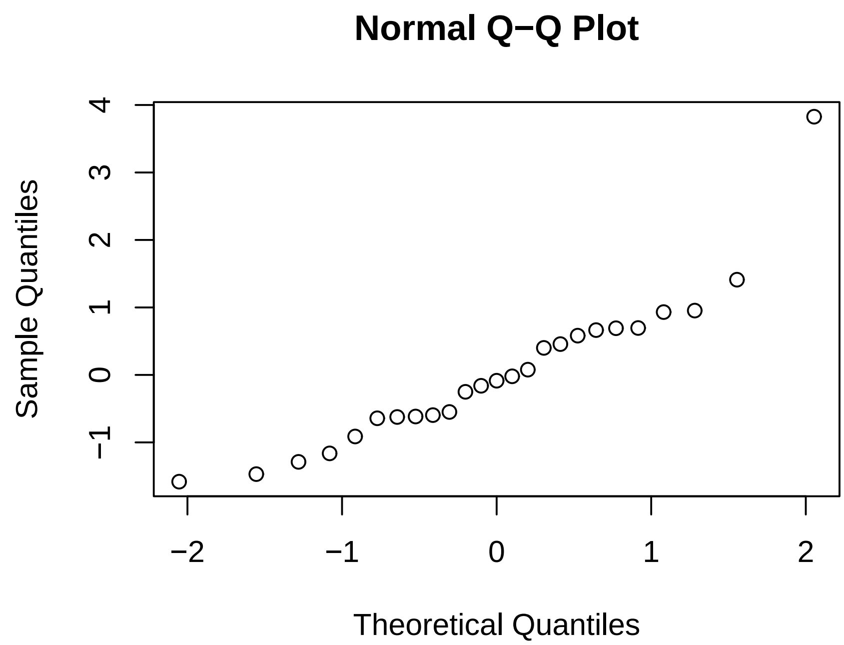
Plot of Externally Studentized Residuals against wetin



Plot of ESR squared vs nitin



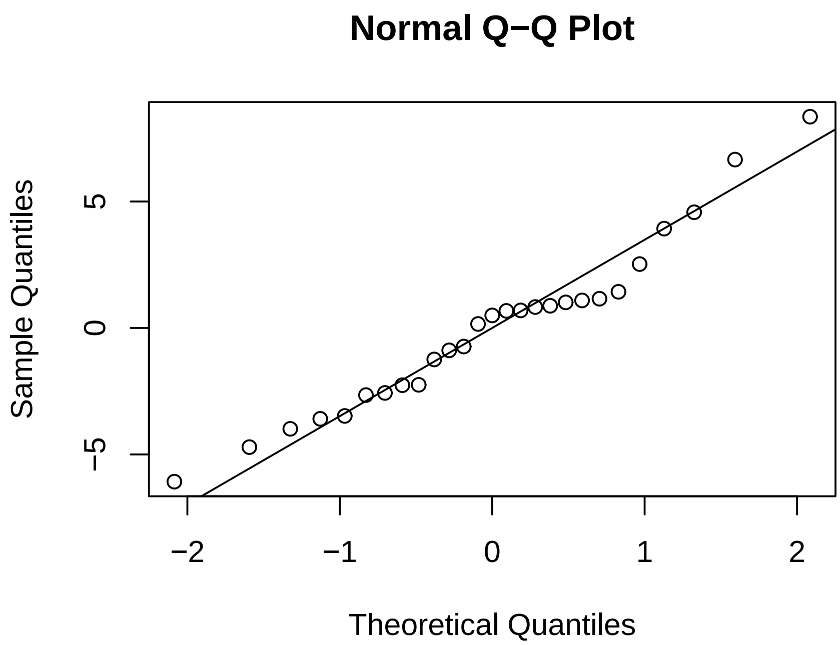
QQ plot of Externally Studentized Residuals

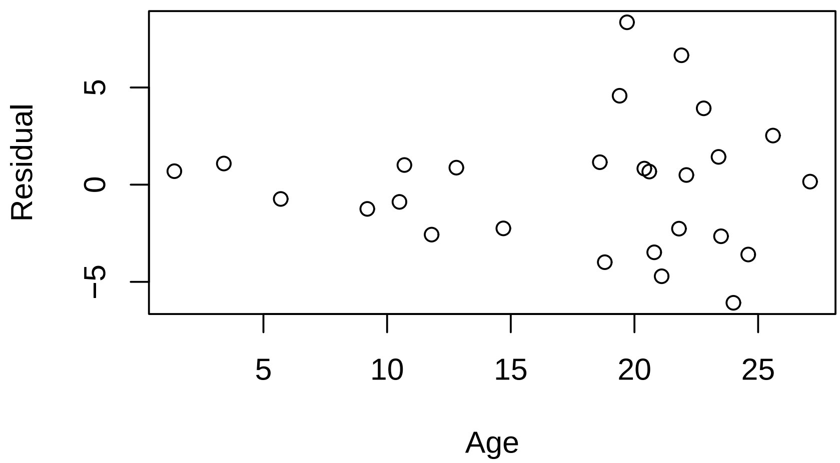


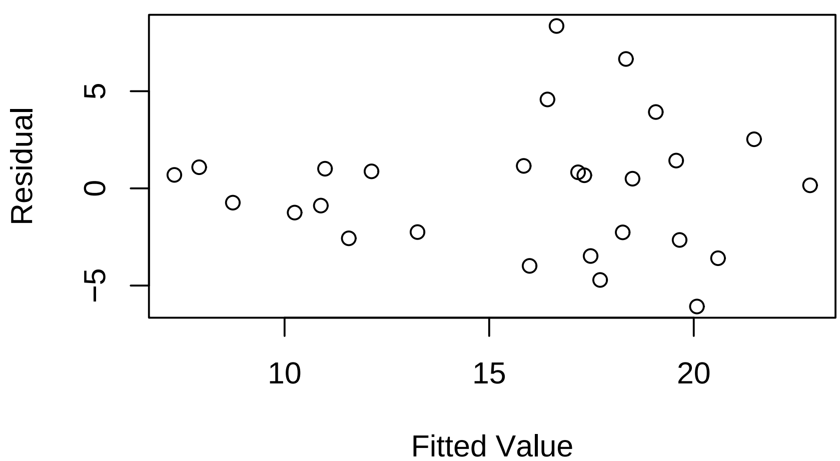
3. From the text page 336, 8.7 a.

The Q-Q plot is adequately straight. I have superimposed a straigh line with slope equal to 3.487, the estimate of σ and intercept 0 since the residuals and standard normal scores both have mean 0. So I see no great problems with the assumption of normality.

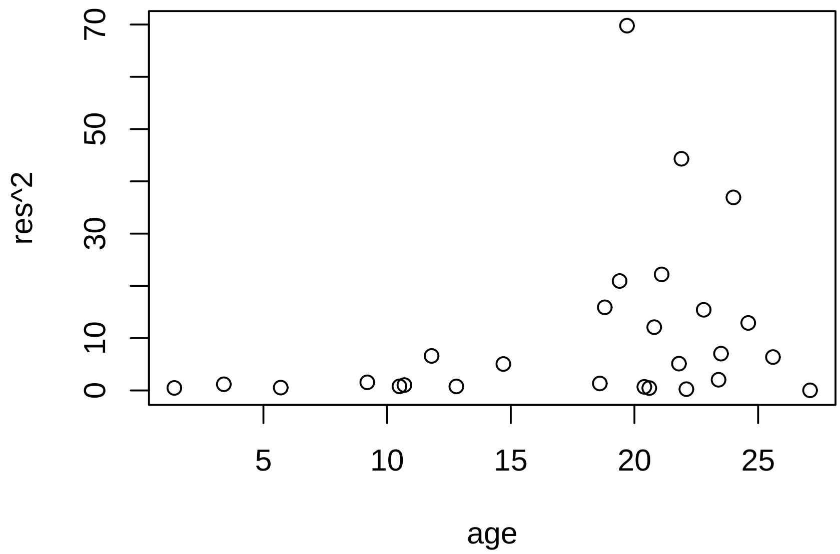
BUT: the plots of residual against both age and fitted value show a problem. The variance looks to be definitely bigger at Ages over 15 or so. The plot against fitted value is not better than that against age.


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
5
0
−5
−2
−1
0
1
2





We aren't asked but I tried plotting the squared residuals against age:



This graph makes it obvious that there is a problem! We can apply the Breusch Pagan test. The R code below does all this and gets a Breusch Pagan statistic equal to 622. For a chi-squared test on 1 degree of freedom this corresponds to a ridiculously small P-value which R calculates as 0.

The R code

```
d = matrix(scan("CH08PR06.txt"),ncol=2,byrow=T)

steroid =d[,2]
age = d[,1]
age2=age^2

fit = lm(steroid~age+age2)

res = residuals(fit)
fitted = fitted(fit)

postscript("Residuals.ps",horizontal=F,width=5,height=7)
par(mfcol=c(2,1))
plot(age,res,xlab="Age",ylab="Residual")
plot(fitted,res,xlab="Fitted Value",ylab="Residual")
dev.off()
postscript("QQ.ps",horizontal=F,width=5,height=4)
qqnorm(res)
```

```
abline(0,3.487)
dev.off()

 postscript("rsq.ps",horizontal=F,width=5,height=4)
 plot(age,res^2)
 dev.off()

r2=res^2

fit2 = lm(r2~age)

mse2= (summary(fit2)$sigma)^2

mse = (summary(fit)$sigma)^2

print("Breusch Pagan test statistic")
BP = (mse2/2)/(mse/length(steroid))^2
print(BP)

print("Breusch Pagan P value")
print(1-pchisq(BP,1))
```

The last bit produces the output:

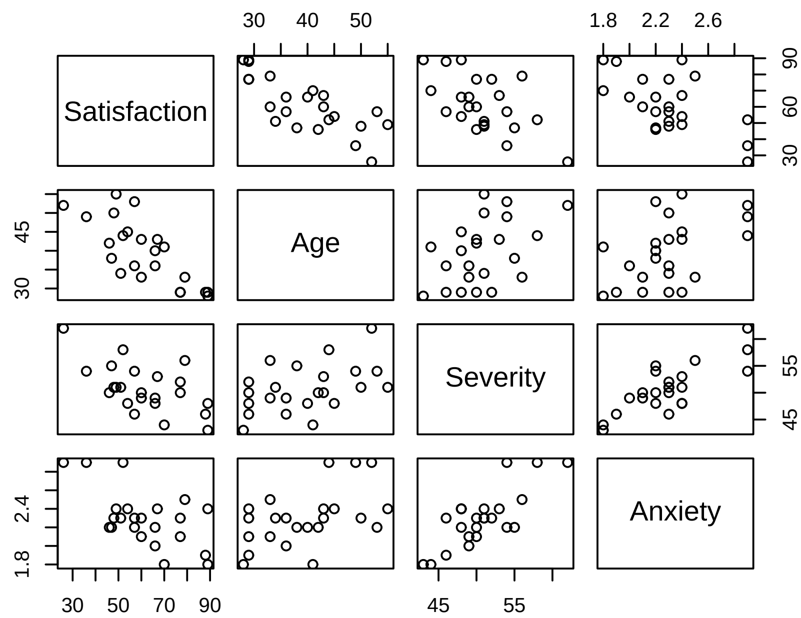
```
Read 54 items
[1] "Breusch Pagan test statistic"
[1] 622.6361
[1] "Breusch Pagan P value"
[1] 0
```

4. Analyze the patient satisfaction data from the text by doing:

   (a) 6.15 b through g (pp 250–251);

   6.15-17

   6.15 b *The pairwise scatterplot is*



The correlation matrix is (omitting redundant entries)

	Satisfaction	*Age*	*Severity*
Satisfaction	*1.000*		
Age	*-0.774*	*1.000*	
Severity	*-0.587*	*0.467*	*1.000*
Anxiety	*-0.602*	*0.498*	*0.795*

For the rest 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:

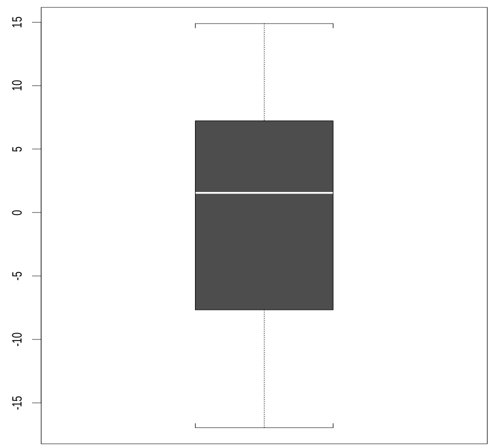
6.15 c *The fitted regression function is*

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

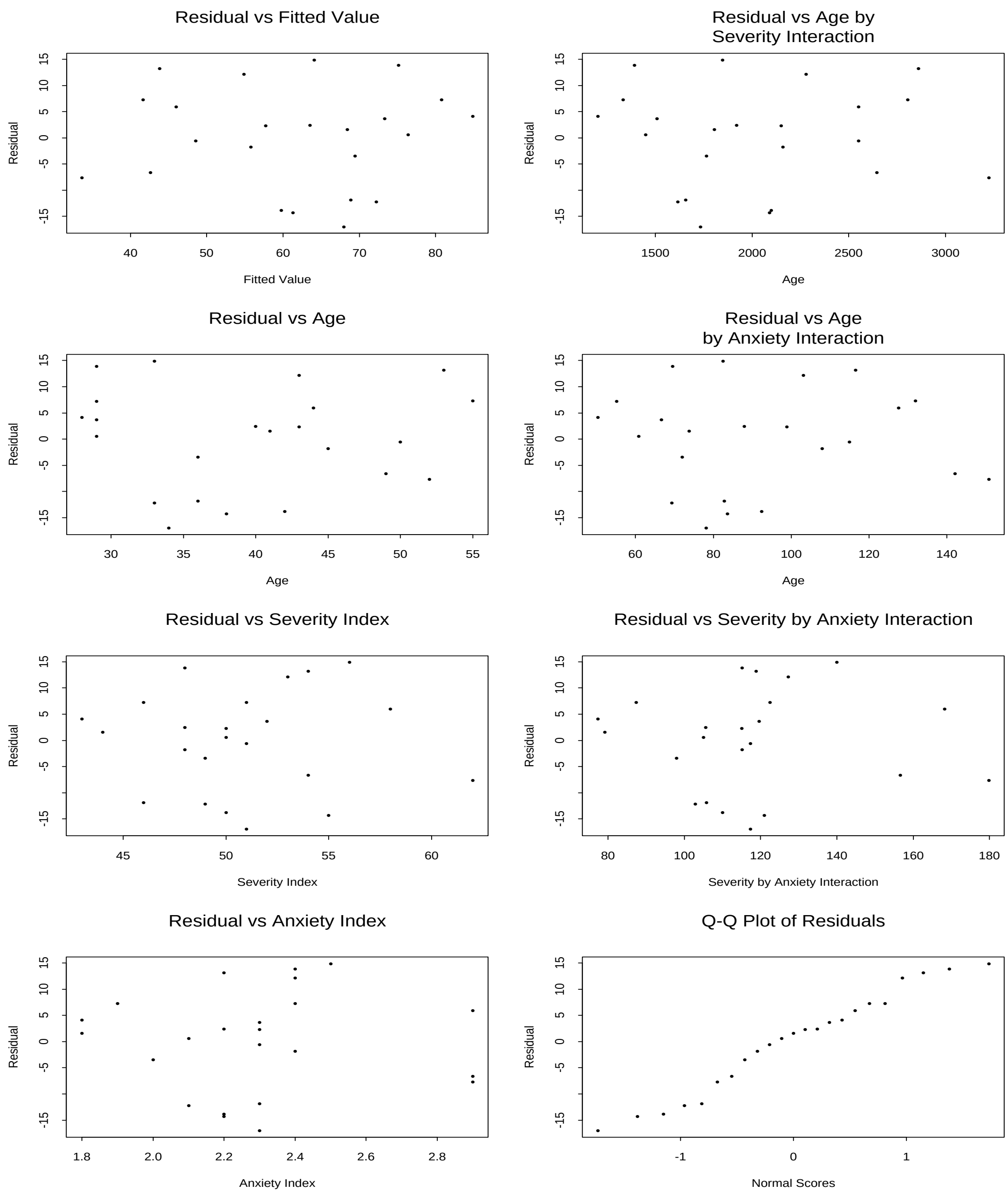
The question asks for an interpretation of b_2. Mathematically it means that holding Age and Anxiety constant an increase of 1 unit in severity of disease is associ-

ated 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.

(b) 6.16 (p251);

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.*

(c) 6.17 a (p 251);

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* $71.6 \pm 1.73(11)$.

(d) 7.9 (p 290);

This is an extra sum of squares test. The full model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

while the reduced model is

$$Y_i = \beta_0 - X_{1i} + \beta_3 X_{3i} + \epsilon_i$$

We can fit the reduced model by writing it as

$$Y_i + X + i1 = \beta_0 + \beta_3 X_{3i} + \epsilon_i$$

We then regress the variable on the left on just X_3. *SAS code is*

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

and edited output is

```
Dependent Variable: ynew
                            Sum of
 Source             DF     Squares    Mean Square  F Value Pr > F
 Model               1   753.437242  753.437242     7.35   0.0131
 Error              21  2151.519280  102.453299
 Corrected Total    22  2904.956522
```

The desired test statistic is

$$F = \frac{(2151.52 - 2011.58)/2}{2011.58/19} = 0.6609$$

From R using `pf(0.6608885,2,19,lower.tail=F)` *we find the P-value is 0.528 which is nowhere near significant. This hypothesis is accepted at the level* $\alpha = 0.025$*.*

(e) 7.26 (p 292).

I used the following SAS code:

```
data patsat;
 infile '615.dat' firstobs=2;
 input Satisf Age Severity Anxiety ;
 proc glm  data=patsat;
  model Satisf = Age Severity ;
 run;
 proc glm  data=patsat;
  model Satisf = Anxiety Age;
 run;
 proc glm  data=patsat;
  model Satisf = Severity Age;
 run;
 proc glm  data=patsat;
  model Satisf = Severity;
 run;
```

The first `glm` *gives you* $SS(X_1)$, $SS(X_2|X_1)$ *and* $SS(X_1, X_2)$*. The second* `glm` *gives you* $SS(X_3)$, $SS(X_1|X_3)$ *and* $SS(X_1, X_3)$*. The fourth* `glm` *gives you* $SS(X_3)$, $SS(X_2|X_3)$ *and* $SS(X_2, X_3)$*. The fifth* `glm` *gives you* $SS(X_2)$*. All these may be found in the Type I sum of squares.*

We then have the following answers:

i. *The fitted regression function is*

$$\hat{Y}_i = 166.591 - 1.260 \times Age - 1.089 Severity$$

ii. *The coefficients are -1.260 an -1.089 in this reduced model where they were -1.210 and -0.666 in the full model. The coefficient of Age is very little changed by that of severity has become more negative.*

iii. *We have* $SS(X_1) = 1706.666$ *(see Type I SS for the first run of* `glm`*) and* $SS(X_1|X_3) = 1834.633$ *(Type I SS for the second run of* `glm`*) so they are somewhat different. We have* $SS(X_2) = 2120.659$ *and* $SS(X_2|X_3) = 402.784$ *which are very different.*

iv. *The point here is that when the correlation between two covariates is low then adjusting for one will make little difference to the Sum of Squares for the other. For Age and Anxiety the correlation is about 0.5 while for Severity and Anxiety it is about 0.8; the adjustment has a much bigger impact in the second case.*

SAS output:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4081.219492	2040.609746	19.77	<.0001
Error	20	2063.997899	103.199895		
Corrected Total	22	6145.217391			

R-Square	Coeff Var	Root MSE	Satisf Mean
0.664129	16.55924	10.15873	61.34783

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Age	1	3678.435847	3678.435847	35.64	<.0001
Severity	1	402.783645	402.783645	3.90	0.0622

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	1960.560918	1960.560918	19.00	0.0003
Severity	1	402.783645	402.783645	3.90	0.0622

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	166.5913303	24.90844062	6.69	<.0001
Age	-1.2604583	0.28918645	-4.36	0.0003
Severity	-1.0893177	0.55138923	-1.98	0.0622

Dependent Variable: Satisf

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4063.982298	2031.991149	19.53	<.0001
Error	20	2081.235094	104.061755		
Corrected Total	22	6145.217391			

R-Square	Coeff Var	Root MSE	Satisf Mean
0.661324	16.62824	10.20107	61.34783

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Anxiety	1	2229.349139	2229.349139	21.42	0.0002
Age	1	1834.633158	1834.633158	17.63	0.0004

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Anxiety	1	385.546451	385.546451	3.70	0.0686
Age	1	1834.633158	1834.633158	17.63	0.0004

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	147.0751185	16.73344897	8.79	<.0001
Anxiety	-15.8906357	8.25559710	-1.92	0.0686
Age	-1.2433613	0.29612038	-4.20	0.0004

Dependent Variable: Satisf

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4081.219492	2040.609746	19.77	<.0001
Error	20	2063.997899	103.199895		
Corrected Total	22	6145.217391			

R-Square	Coeff Var	Root MSE	Satisf Mean
0.664129	16.55924	10.15873	61.34783

Source	DF	Type I SS	Mean Square	F Value	Pr > F

Severity	1	2120.658574	2120.658574	20.55	0.0002
Age	1	1960.560918	1960.560918	19.00	0.0003

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Severity	1	402.783645	402.783645	3.90	0.0622
Age	1	1960.560918	1960.560918	19.00	0.0003

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	166.5913303	24.90844062	6.69	<.0001
Severity	-1.0893177	0.55138923	-1.98	0.0622
Age	-1.2604583	0.28918645	-4.36	0.0003

Dependent Variable: Satisf

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2120.658574	2120.658574	11.07	0.0032
Error	21	4024.558818	191.645658		
Corrected Total	22	6145.217391			

R-Square	Coeff Var	Root MSE	Satisf Mean
0.345091	22.56578	13.84361	61.34783

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Severity	1	2120.658574	2120.658574	11.07	0.0032

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Severity	1	2120.658574	2120.658574	11.07	0.0032

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	173.6140281	33.87239519	5.13	<.0001
Severity	-2.2107214	0.66458133	-3.33	0.0032