

MATH 204 - ASSIGNMENT 2

SOLUTIONS

Note: In the analysis of these data sets, there is rarely a definitively 'correct' answer, so these solutions are model solutions that illustrate the use of the required approaches. My model solutions typically give more extensive detail that is required as I have tried to cover all the possible solution methods that you might give.

Q1. In a study of 111 patients who have undergone a kidney transplant and subsequently experienced organ rejection, the time (in months) until signs of rejection were diagnosed was recorded. The influence of genetic factors on time to rejection was studied. Two predictors were measured at time of transplantation:

- **Genotype:** The genotype of the patient at a specific genomic locus. With alleles G and g, the three possible genotypes forming the three levels of this **factor predictor** are
 - gg (labelled 0)
 - Gg (labelled 1)
 - GG (labelled 2)
- **Expression:** The level of expression of a second gene believed to be involved in immune defence. This variable is a **continuous covariate**.

The response (or dependent) variable is **Time**, the time in months until rejection is diagnosed. The data are available at

<https://www.math.mcgill.ca/dstephens/204/Assignments/A2/Rejection.csv>

Using the techniques studied in lectures for subgroup analyses, find the most suitable model for predicting time until rejection based on the genetic information. Assess the fit of your selected model.

- | | |
|---------------------|---------|
| (a) Model fitting | 8 Marks |
| (b) Model selection | 8 Marks |
| (c) Model checking | 4 Marks |

```
> #R Code for data entry and elementary processing
> fname<-"https://www.math.mcgill.ca/dstephens/204/Assignments/A2/Rejection.csv"
> Rejection<-read.csv(fname)
> Rejection$Genotype<-as.factor(Rejection$Genotype)
> str(Rejection)

+ 'data.frame': 111 obs. of 3 variables:
+ $ Genotype : Factor w/ 3 levels "gg","Gg","GG": 1 1 1 1 1 1 1 1 1 1 ...
+ $ Expression: num 8.04 7.79 8.08 6.68 8.23 ...
+ $ Time : num 17 18.5 19.1 19 20.4 ...

> nrow(Rejection)

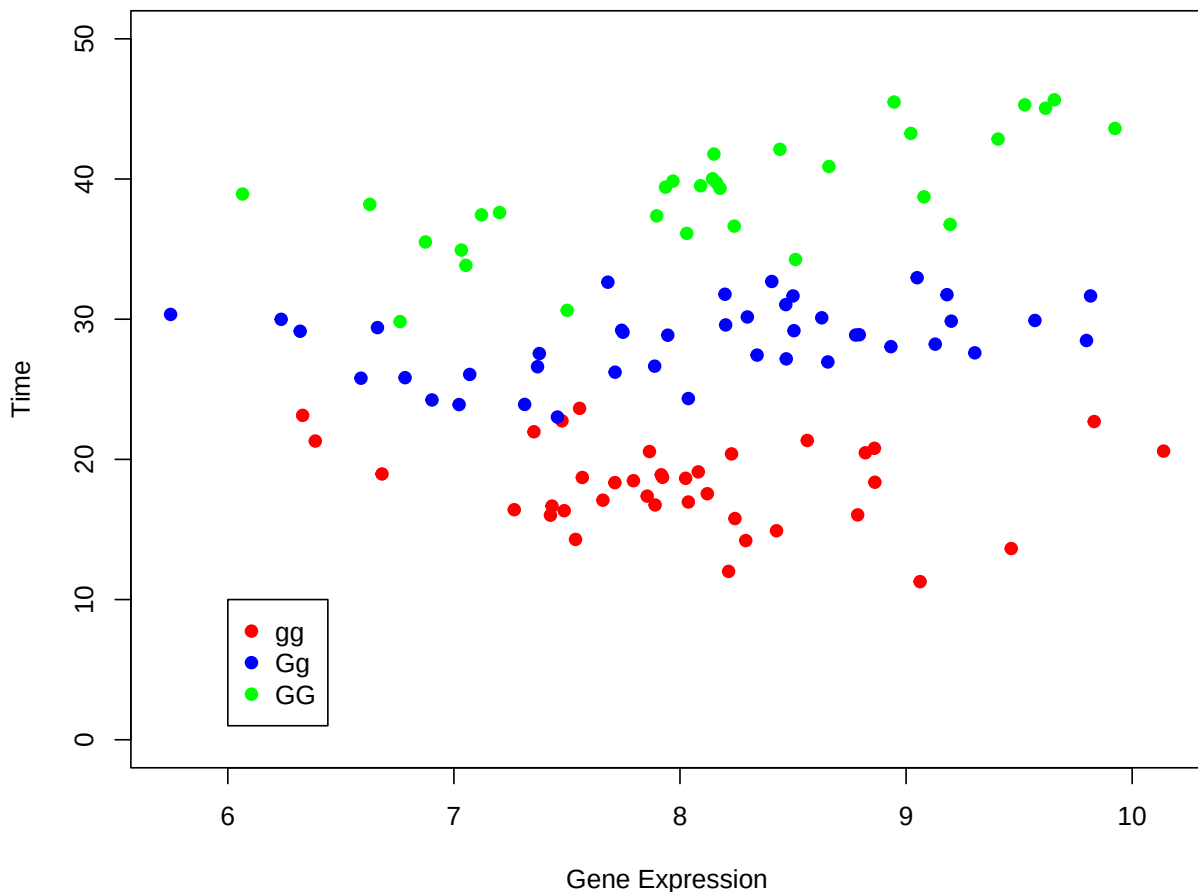
+ [1] 111

> table(Rejection$Genotype)

+
+ gg Gg GG
+ 38 42 31
```

Solution: We are here in the subgroup setting so we examine the data as a plot first.

```
> par(mar=c(4,4,2,2))
> plot(Rejection$Expression, Rejection$Time, xlab='Gene Expression', ylab='Time',
+      type='n',ylim=range(0,50))
> points(Rejection$Expression[Rejection$Genotype=='gg'],
+        Rejection$Time[Rejection$Genotype=='gg'],pch=19,col='red')
> points(Rejection$Expression[Rejection$Genotype=='Gg'],
+        Rejection$Time[Rejection$Genotype=='Gg'],pch=19,col='blue')
> points(Rejection$Expression[Rejection$Genotype=='GG'],
+        Rejection$Time[Rejection$Genotype=='GG'],pch=19,col='green')
> legend(6,10,c('gg','Gg','GG'),pch=19, col=c('red','blue','green'))
```



(a) We first fit the five standard models

```
> fit0<-lm(Time~1,data=Rejection)
> fit1<-lm(Time~Genotype,data=Rejection)
> fit2<-lm(Time~Expression,data=Rejection)
> fit3<-lm(Time~Expression+Genotype,data=Rejection)
> fit4<-lm(Time~Genotype+Expression+Genotype:Expression,data=Rejection)
> SSE0<-sum(residuals(fit0)^2)
> SSE1<-sum(residuals(fit1)^2)
> SSE2<-sum(residuals(fit2)^2)
> SSE3<-sum(residuals(fit3)^2)
> SSE4<-sum(residuals(fit4)^2)
```

This yields the following SSE table, where p is the total number of parameters.

Model	<i>SSE</i>	<i>p</i>
1	8560.592	1
Genotype	1106.476	3
Expression	8320.196	2
Genotype+Expression	987.661	4
Genotype+Expression+Genotype:Expression	824.176	6

This set of models is sufficient for the moment

8 Marks

- (b) For model selection we carry out the standard approaches

Backward selection:

```
> drop1(fit4, test='F')
+ Single term deletions
+
+ Model:
+ Time ~ Genotype + Expression + Genotype:Expression
+
+               Df Sum of Sq  RSS   AIC F value    Pr(>F)
+ <none>                    824.2 234.5
+ Genotype:Expression  2      163.5 987.7 250.6   10.41 7.48e-05 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This result indicates that the Model 4 cannot be simplified.

Forward selection: We begin with Model 0 and attempt to add Genotype as this yields a much smaller *SSE*

```
> anova(fit0,fit1,test='F')
+ Analysis of Variance Table
+
+ Model 1: Time ~ 1
+ Model 2: Time ~ Genotype
+   Res.Df  RSS Df Sum of Sq    F Pr(>F)
+ 1      110 8561
+ 2      108 1106   2      7454 363.8 <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is highly significant, so we attempt to add Expression:

```
> anova(fit1,fit3,test='F')
+ Analysis of Variance Table
+
+ Model 1: Time ~ Genotype
+ Model 2: Time ~ Expression + Genotype
+   Res.Df  RSS Df Sum of Sq    F    Pr(>F)
+ 1      108 1106.5
+ 2      107  987.7   1      118.8 12.87 0.000505 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is also highly significant, so we attempt to add the interaction:

```
> anova(fit3,fit4,test='F')
+ Analysis of Variance Table
+
+ Model 1: Time ~ Expression + Genotype
+ Model 2: Time ~ Genotype + Expression + Genotype:Expression
+   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
+ 1      107 987.7
+ 2      105 824.2  2      163.5 10.41 7.48e-05 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is also significant, so we conclude that both Backward and Forward selection lead us to the model

Genotype+Expression+Genotype:Expression

Here is the fit summary

```
> summary(fit4)
+
+ Call:
+ lm(formula = Time ~ Genotype + Expression + Genotype:Expression,
+   data = Rejection)
+
+ Residuals:
+   Min       1Q   Median       3Q      Max
+ -6.569 -1.711  0.244  2.169  5.760
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)      23.278      4.535   5.13 1.3e-06 ***
+ GenotypeGg       -2.243      5.748  -0.39  0.70
+ GenotypeGG       -7.127      6.181  -1.15  0.25
+ Expression       -0.634      0.562  -1.13  0.26
+ GenotypeGg:Expression  1.561      0.711   2.20  0.03 *
+ GenotypeGG:Expression  3.440      0.759   4.53 1.6e-05 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 2.8 on 105 degrees of freedom
+ Multiple R-squared:  0.904, Adjusted R-squared:  0.899
+ F-statistic: 197 on 5 and 105 DF, p-value: <2e-16
```

This yields a high R-squared value, so the fit has good explanatory power. The individual t-tests indicate that the intercepts are not very different, but the slopes are. 8 Marks

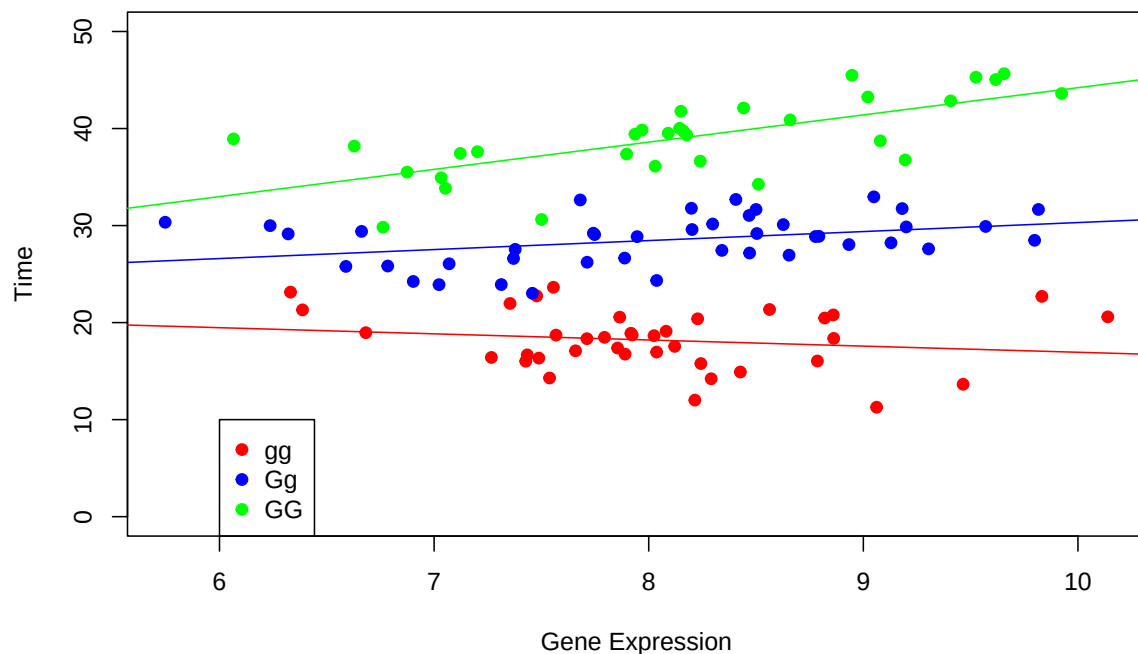
(c) Model checking: we begin with the data plots with the lines of best fit

```
> par(mar=c(4,4,1,2))
> plot(Rejection$Expression, Rejection$Time,
+   xlab='Gene Expression', ylab='Time', type='n',ylim=range(0,50))
> points(Rejection$Expression[Rejection$Genotype=='gg'],
+   Rejection$Time[Rejection$Genotype=='gg'],pch=19,col='red')
> points(Rejection$Expression[Rejection$Genotype=='Gg'],
+   Rejection$Time[Rejection$Genotype=='Gg'],pch=19,col='blue')
```

```

> points(Rejection$Expression[Rejection$Genotype=='GG'],
+       Rejection$Time[Rejection$Genotype=='GG'],pch=19,col='green')
> legend(6,10,c('gg','Gg','GG'),pch=19,col=c('red','blue','green'))
> be.hat<-coef(fit4)
> abline(be.hat[1],be.hat[4],col='red')
> abline(be.hat[1]+be.hat[2],be.hat[4]+be.hat[5],col='blue')
> abline(be.hat[1]+be.hat[3],be.hat[4]+be.hat[6],col='green')

```

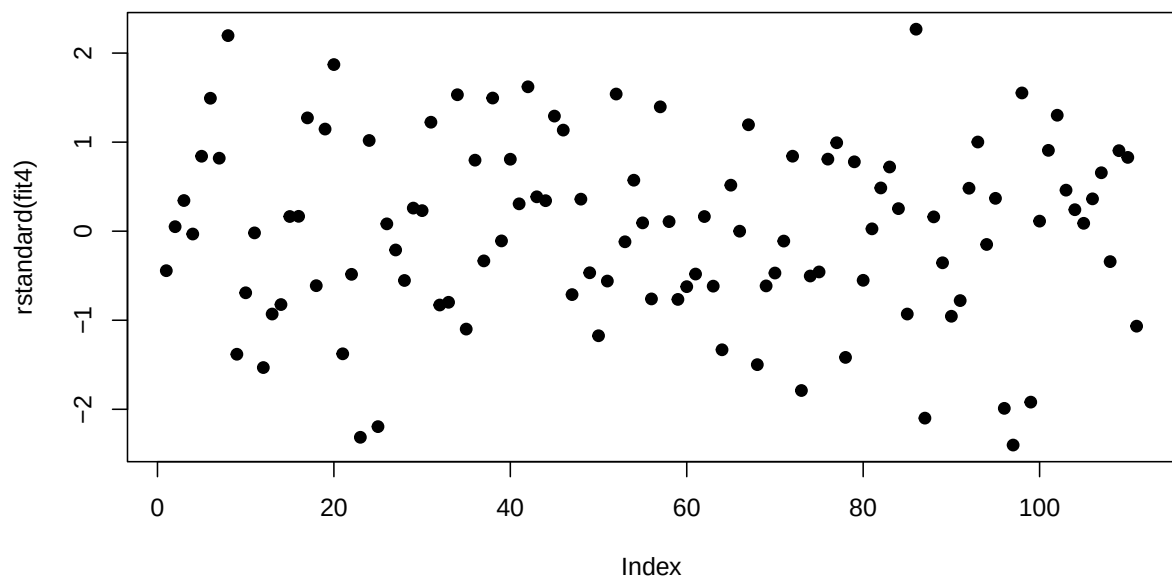


For the residual plots, we begin with a standard plot of the individual residuals:

```

> par(mar=c(4,4,1,1))
> plot(rstandard(fit4),pch=19)

```

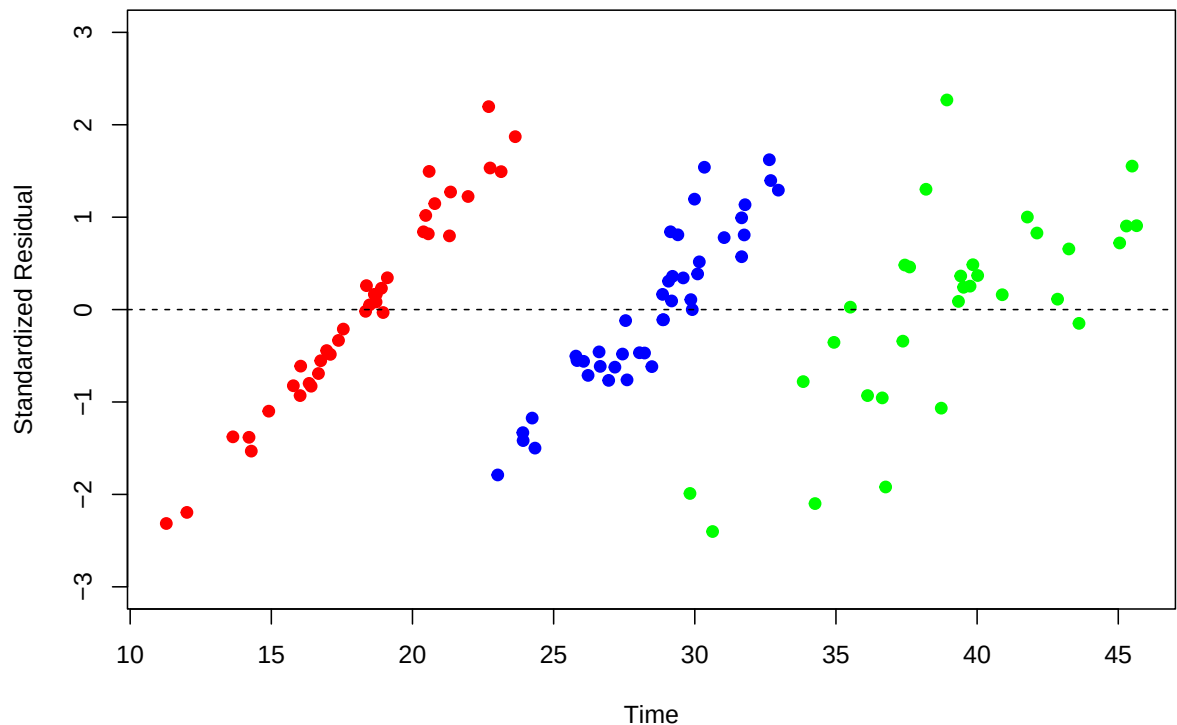


This patternless plot seems adequate. Now we plot against the Expression predictor

```

> par(mar=c(4,4,1,1))
> zvec<-rstandard(fit4)
> plot(Rejection$Time,zvec,ylim=range(-3,3),type='n',
+      pch=19,xlab="Time",ylab="Standardized Residual")
> points(Rejection$Time[Rejection$Genotype=='gg'],
+        zvec[Rejection$Genotype=='gg'],pch=19,col='red')
> points(Rejection$Time[Rejection$Genotype=='Gg'],
+        zvec[Rejection$Genotype=='Gg'],pch=19,col='blue')
> points(Rejection$Time[Rejection$Genotype=='GG'],
+        zvec[Rejection$Genotype=='GG'],pch=19,col='green')
> abline(h=0,lty=2)

```

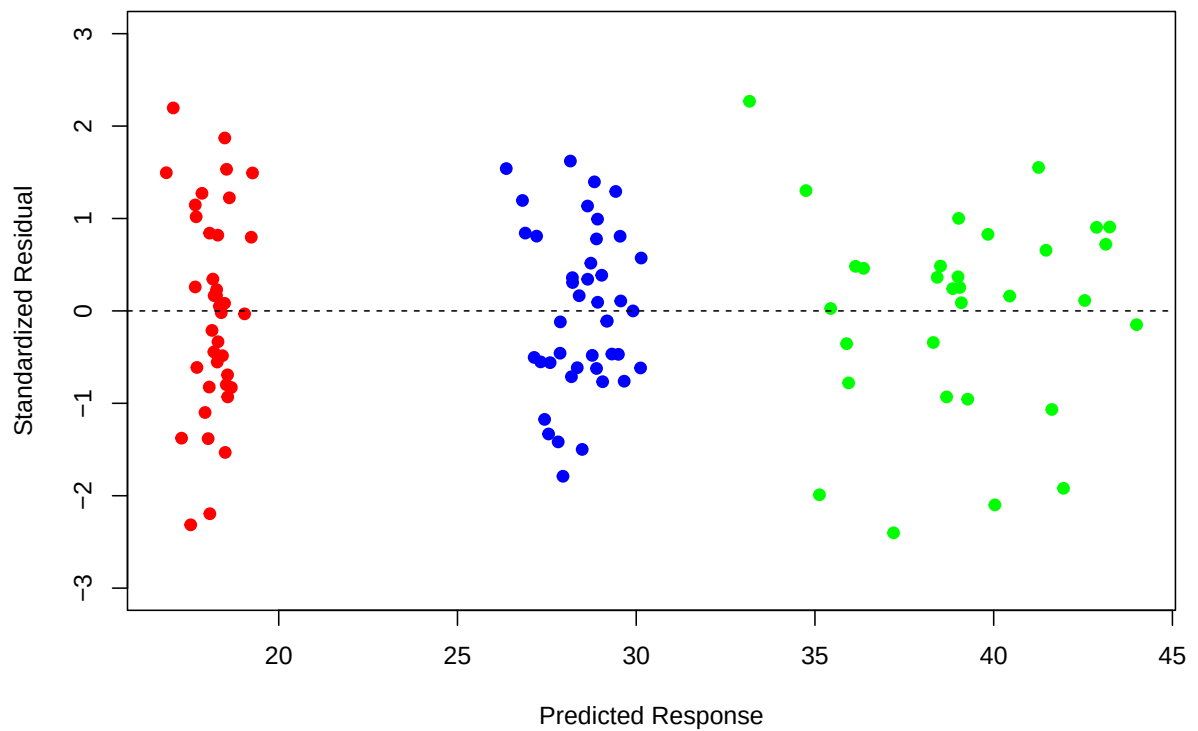


Due to the presence of the interaction, this plot is also adequate. Finally, we plot against the predicted response $\hat{y}$.

```

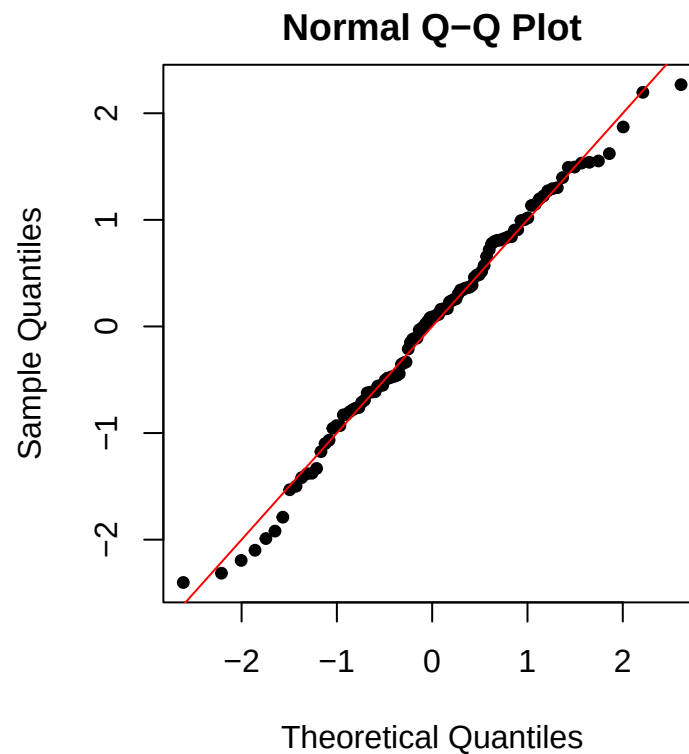
> par(mar=c(4,4,1,1))
> yvec<-predict(fit4)
> plot(yvec,zvec,ylim=range(-3,3),type='n',
+      pch=19,xlab="Predicted Response",ylab="Standardized Residual")
> points(yvec[Rejection$Genotype=='gg'],
+        zvec[Rejection$Genotype=='gg'],pch=19,col='red')
> points(yvec[Rejection$Genotype=='Gg'],
+        zvec[Rejection$Genotype=='Gg'],pch=19,col='blue')
> points(yvec[Rejection$Genotype=='GG'],
+        zvec[Rejection$Genotype=='GG'],pch=19,col='green')
> abline(h=0,lty=2)

```



Finally, a Q-Q plot can confirm that there are no serious deviations from normality.

```
> par(mar=c(4,4,2,1), pty='s')
> qqnorm(rstandard(fit4), pch=19, cex=0.75)
> abline(0,1,col='red')
```



This plot is certainly reasonable, and supports the normality assumption.

4 Marks

Q2. In an experimental study of the benefits of three dietary regimen on muscle development, 36 subjects aged 25-35 and 36 subjects aged 36-45 were recruited. After six weeks on study, the subjects underwent a strength test, and the result was recorded. At the end of study, the body mass of the subjects was also recorded. The variables recorded in the data set are

- **Age:** Age group
 - 25-35 (labelled 1)
 - 36-45 (labelled 2)
- **Regimen:** Dietary regimen followed
 - Diet 1
 - Diet 2
 - Standard
- **Mass:** body mass (in kg) at end of study

The response (or dependent) variable Y is the strength test result at end of study. The data are available at

<https://www.math.mcgill.ca/dstephens/204/Assignments/A2/DietStudy.csv>

Using the techniques studied in lectures, find the most suitable model for predicting test strength. Assess the fit of your selected model.

- | | |
|---------------------|---------|
| (a) Model fitting | 8 Marks |
| (b) Model selection | 8 Marks |
| (c) Model checking | 4 Marks |

```
> #R Code for data entry and elementary processing
> fname<-"https://www.math.mcgill.ca/dstephens/204/Assignments/A2/DietStudy.csv"
> DietStudy<-read.csv(fname)
> DietStudy$Age<-as.factor(DietStudy$Age)
> DietStudy$Regimen<-factor(DietStudy$Regimen, levels=c('Standard','Diet 1','Diet 2'))
> str(DietStudy)

+ 'data.frame': 72 obs. of 4 variables:
+ $ Age      : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
+ $ Regimen: Factor w/ 3 levels "Standard","Diet 1",...: 2 2 2 2 2 2 2 2 2 2 ...
+ $ Mass     : num  67.2 64.8 68 66.9 69.5 69.1 70.3 64.4 65.8 64.1 ...
+ $ Y        : num  198 201 198 196 201 ...

> nrow(DietStudy)

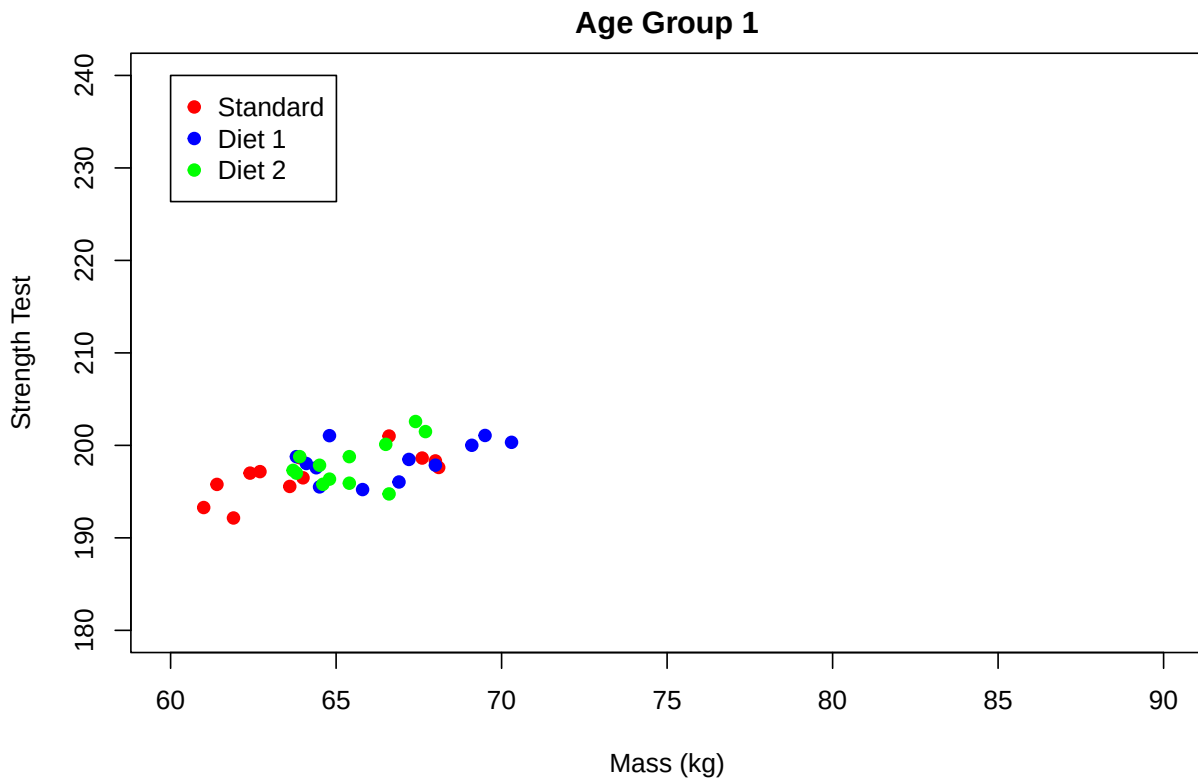
+ [1] 72

> table(DietStudy$Age, DietStudy$Regimen)

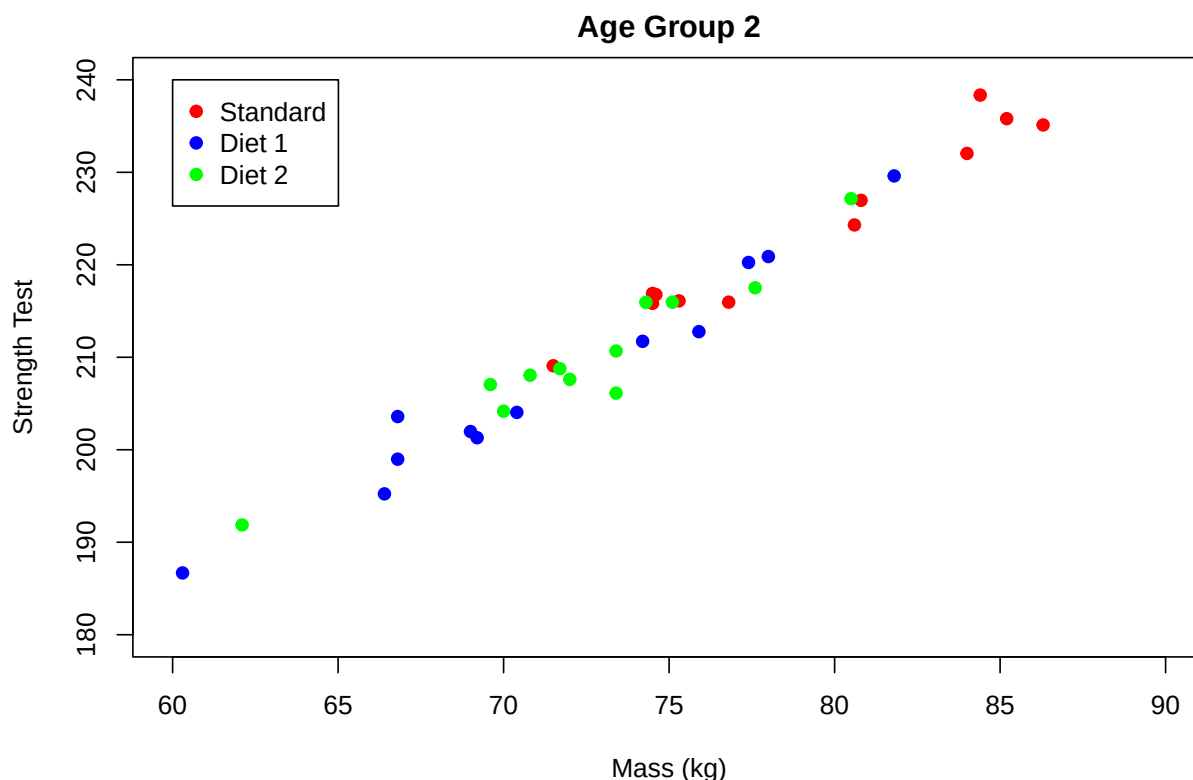
+
+      Standard Diet 1 Diet 2
+ 1          12      12      12
+ 2          12      12      12
```

Solution: We are here in the subgroup setting so we examine the data as a plot first. We plot the two Age groups separately.

```
> Diet1<-subset(DietStudy, Age == '1')
> Diet2<-subset(DietStudy, Age == '2')
> par(mar=c(4,4,2,2))
> plot(Diet1$Mass, Diet1$Y, xlab='Mass (kg)', ylab='Strength Test',
+      type='n',xlim=range(60,90),ylim=range(180,240))
> points(Diet1$Mass[Diet1$Regimen=='Standard'],
+        Diet1$Y[Diet1$Regimen=='Standard'],pch=19,col='red')
> points(Diet1$Mass[Diet1$Regimen=='Diet 1'],
+        Diet1$Y[Diet1$Regimen=='Diet 1'],pch=19,col='blue')
> points(Diet1$Mass[Diet1$Regimen=='Diet 2'],
+        Diet1$Y[Diet1$Regimen=='Diet 2'],pch=19,col='green')
> legend(60,240,c('Standard','Diet 1','Diet 2'),pch=19, col=c('red','blue','green'))
> title('Age Group 1')
```



```
> plot(Diet2$Mass, Diet2$Y, xlab='Mass (kg)', ylab='Strength Test',
+      type='n',xlim=range(60,90),ylim=range(180,240))
> points(Diet2$Mass[Diet2$Regimen=='Standard'],
+        Diet2$Y[Diet2$Regimen=='Standard'],pch=19,col='red')
> points(Diet2$Mass[Diet2$Regimen=='Diet 1'],
+        Diet2$Y[Diet2$Regimen=='Diet 1'],pch=19,col='blue')
> points(Diet2$Mass[Diet2$Regimen=='Diet 2'],
+        Diet2$Y[Diet2$Regimen=='Diet 2'],pch=19,col='green')
> legend(60,240,c('Standard','Diet 1','Diet 2'),pch=19, col=c('red','blue','green'))
> title('Age Group 2')
```



We get a sense from these plots that

- Mass certainly has an effect, but it is different in the two Age groups
- Regimen does not seem to have an effect once that Mass and Age are accounted for.

This suggests that a model of the form $\text{Mass} + \text{Age} + \text{Mass}:\text{Age}$ might be favoured, although Regimen might be included as a further main effect.

We proceed with an analysis, beginning with the full factorial model:

```
> fit1<-lm(Y~Age*Regimen*Mass,data=DietStudy)
> print(summary(aov(fit1)), digits=6)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
+ Age	1	4557.74	4557.74	967.96399	< 2.22e-16 ***
+ Regimen	2	753.92	376.96	80.05854	< 2.22e-16 ***
+ Mass	1	4252.39	4252.39	903.11375	< 2.22e-16 ***
+ Age:Regimen	2	8.88	4.44	0.94247	0.39536
+ Age:Mass	1	224.35	224.35	47.64794	3.6642e-09 ***
+ Regimen:Mass	2	0.42	0.21	0.04464	0.95638
+ Age:Regimen:Mass	2	6.73	3.36	0.71448	0.49356
+ Residuals	60	282.51	4.71		
+ ---					

+ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The study here is **balanced**, so we can rely more on the ANOVA F-tests for the factors, however it is still the case that because we have interactions between the factors and the continuous predictor, the F-test results are not directly reliable. For example, if we rearrange the terms in the model, some of the F-test results change.

```
> fit1a<-lm(Y~Regimen*Age*Mass,data=DietStudy)
> print(summary(aov(fit1a)), digits=6)

+               Df Sum Sq Mean Sq  F value    Pr(>F)
+ Regimen         2   753.92   376.96  80.05854 < 2.22e-16 ***
+ Age             1  4557.74  4557.74 967.96399 < 2.22e-16 ***
+ Mass            1  4252.39  4252.39 903.11375 < 2.22e-16 ***
+ Regimen:Age      2     8.88    4.44   0.94247   0.39536
+ Regimen:Mass     2     4.16    2.08   0.44179   0.64496
+ Age:Mass         1   220.61  220.61  46.85362  4.596e-09 ***
+ Regimen:Age:Mass 2     6.73    3.36   0.71448   0.49356
+ Residuals       60   282.51    4.71
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the *SSR* quantity for Age:Mass interaction is different in the second analysis, even though the model fitted is the same.

(a) Given the exploratory analysis, it seems that a model simplification to

$$\text{Age} + \text{Regimen} + \text{Mass} + \text{Age:Mass}$$

should be supported

```
> fit2<-lm(Y~Age+Regimen+Mass+Age:Mass, data=DietStudy)
> anova(fit1,fit2,test='F')

+ Analysis of Variance Table
+
+ Model 1: Y ~ Age * Regimen * Mass
+ Model 2: Y ~ Age + Regimen + Mass + Age:Mass
+   Res.Df  RSS Df Sum of Sq    F Pr(>F)
+ 1       60 282.5
+ 2       66 301.0 -6    -18.48 0.654  0.687
```

This supports the model simplification. We assess whether a further simplification is supported using drop1:

```
> drop1(fit2, test='F')

+ Single term deletions
+
+ Model:
+ Y ~ Age + Regimen + Mass + Age:Mass
+               Df Sum of Sq  RSS   AIC F value    Pr(>F)
+ <none>                 301.0 115.0
+ Regimen    2         5.27 306.3 112.2   0.578    0.564
+ Age:Mass   1        221.90 522.9 152.8  48.657 1.79e-09 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This suggests that Regimen can in fact be removed without significantly compromising the fit.

```
> fit3<-update(fit2,~.-Regimen)
> summary(fit3)
```

```

+
+ Call:
+ lm(formula = Y ~ Age + Mass + Age:Mass, data = DietStudy)
+
+ Residuals:
+   Min       1Q   Median       3Q      Max
+ -5.739 -1.543  0.236  1.295  4.947
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  158.079     10.168   15.55 < 2e-16 ***
+ Age2         -89.975     11.047   -8.15 1.2e-11 ***
+ Mass          0.606      0.155    3.90 0.00022 ***
+ Age2:Mass     1.352      0.166    8.15 1.2e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 2.12 on 68 degrees of freedom
+ Multiple R-squared:  0.97, Adjusted R-squared:  0.968
+ F-statistic: 724 on 3 and 68 DF, p-value: <2e-16

```

A further simplification is not supported however:

```

> drop1(fit3, test='F')
+ Single term deletions
+
+ Model:
+ Y ~ Age + Mass + Age:Mass
+      Df Sum of Sq  RSS   AIC F value    Pr(>F)
+ <none>                 306.3 112.2
+ Age:Mass  1       299.4 605.6 159.3   66.46 1.16e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

This leaves us with the originally proposed model

Age + Mass + Age:Mass

8 Marks

- (b) We have essentially implemented a backward selection in (a), so we first attempt a forward selection using the add1 function

```

> ffit0<-lm(Y~1, data=DietStudy)
> add1(ffit0, scope=~Age*Regimen*Mass, test='F')
+ Single term additions
+
+ Model:
+ Y ~ 1
+      Df Sum of Sq  RSS   AIC F value    Pr(>F)
+ <none>                 10087 357.8
+ Age      1       4558  5529 316.6   57.701 1.01e-10 ***
+ Regimen  2        754  9333 356.3    2.787  0.0686 .
+ Mass     1       9481   606 157.3 1095.736 < 2e-16 ***
+ ---

```

```

+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> ffit1<-update(ffit0,~.+Mass)
> add1(ffit1, scope=~Age*Regimen*Mass, test='F')
+ Single term additions
+
+ Model:
+ Y ~ Mass
+
+ Df Sum of Sq  RSS   AIC F value  Pr(>F)
+ <none>                605.7 157.3
+ Age      1         0.09 605.6 159.3   0.010 0.92095
+ Regimen  2        81.16 524.5 151.0   5.261 0.00751 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> ffit2<-update(ffit1,~.+Regimen)
> add1(ffit2, scope=~Age*Regimen*Mass, test='F')
+ Single term additions
+
+ Model:
+ Y ~ Mass + Regimen
+
+ Df Sum of Sq  RSS   AIC F value  Pr(>F)
+ <none>                524.5 151.0
+ Age      1         1.643 522.9 152.8   0.210 0.648
+ Regimen:Mass  2         1.971 522.6 154.7   0.124 0.883

```

This suggests that the forward model selection stops at the

Mass + Regimen

(FORWARD)

model.

```

> summary(ffit2)
+
+ Call:
+ lm(formula = Y ~ Mass + Regimen, data = DietStudy)
+
+ Residuals:
+    Min       1Q   Median       3Q      Max
+ -6.342 -1.750 -0.198  1.879  5.725
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)   84.1595     3.7728   22.31  <2e-16 ***
+ Mass           1.7591     0.0521   33.79  <2e-16 ***
+ RegimenDiet 1  -2.6406     0.8141   -3.24  0.0018 **
+ RegimenDiet 2  -1.3815     0.8140   -1.70  0.0942 .
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 2.78 on 68 degrees of freedom
+ Multiple R-squared:  0.948, Adjusted R-squared:  0.946
+ F-statistic: 413 on 3 and 68 DF, p-value: <2e-16

```

This gives a high R-squared value, so is quite a good model for explaining the observed variation, although perhaps not quite as good as the originally selected model.

We now attempt a full backward selection:

```
> bfit1<-lm(Y~Age*Regimen*Mass,data=DietStudy)
> drop1(bfit1,test='F')

+ Single term deletions
+
+ Model:
+ Y ~ Age * Regimen * Mass
+               Df Sum of Sq   RSS   AIC F value Pr(>F)
+ <none>                        282.5 122.4
+ Age:Regimen:Mass  2         6.728 289.2 120.1   0.714   0.494
```

```
> bfit2<-update(bfit1,~.-Age:Regimen:Mass)
> drop1(bfit2,test='F')

+ Single term deletions
+
+ Model:
+ Y ~ Age + Regimen + Mass + Age:Regimen + Age:Mass + Regimen:Mass
+               Df Sum of Sq   RSS   AIC F value   Pr(>F)
+ <none>                        289.2 120.1
+ Age:Regimen  2         4.34 293.6 117.2   0.465   0.630
+ Age:Mass     1       220.61 509.9 158.9 47.289 3.51e-09 ***
+ Regimen:Mass 2         0.42 289.7 116.2   0.045   0.956
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> bfit3<-update(bfit2,~.-Regimen:Mass)
> drop1(bfit3,test='F')

+ Single term deletions
+
+ Model:
+ Y ~ Age + Regimen + Mass + Age:Regimen + Age:Mass
+               Df Sum of Sq   RSS   AIC F value   Pr(>F)
+ <none>                        289.7 116.2
+ Age:Regimen  2        11.33 301.0 115.0   1.252   0.293
+ Age:Mass     1       224.35 514.0 155.5 49.570 1.58e-09 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> bfit4<-update(bfit3,~.-Age:Regimen)
> drop1(bfit4,test='F')

+ Single term deletions
+
+ Model:
+ Y ~ Age + Regimen + Mass + Age:Mass
+               Df Sum of Sq   RSS   AIC F value   Pr(>F)
+ <none>                        301.0 115.0
+ Regimen      2         5.27 306.3 112.2   0.578   0.564
```

```
+ Age:Mass  1    221.90 522.9 152.8 48.657 1.79e-09 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> bfit5<-update(bfit4,~.-Regimen)
> drop1(bfit5,test='F')

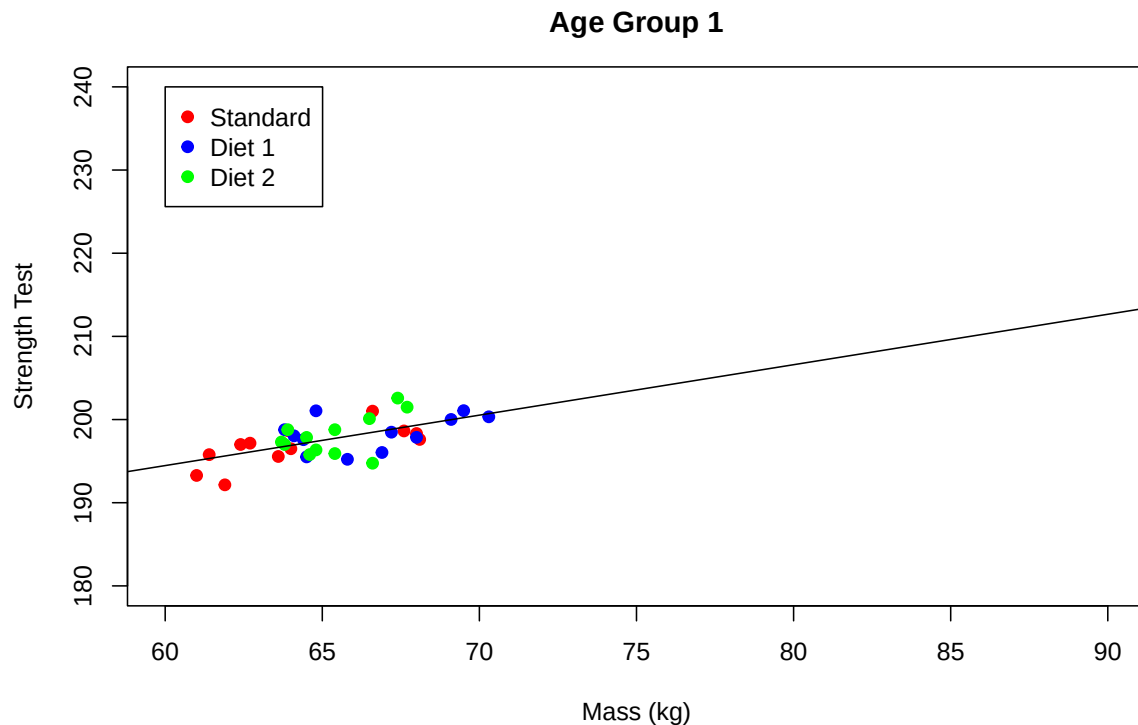
+ Single term deletions
+
+ Model:
+ Y ~ Age + Mass + Age:Mass
+
+   Df Sum of Sq  RSS   AIC F value    Pr(>F)
+ <none>                 306.3 112.2
+ Age:Mass  1      299.4 605.6 159.3   66.46 1.16e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

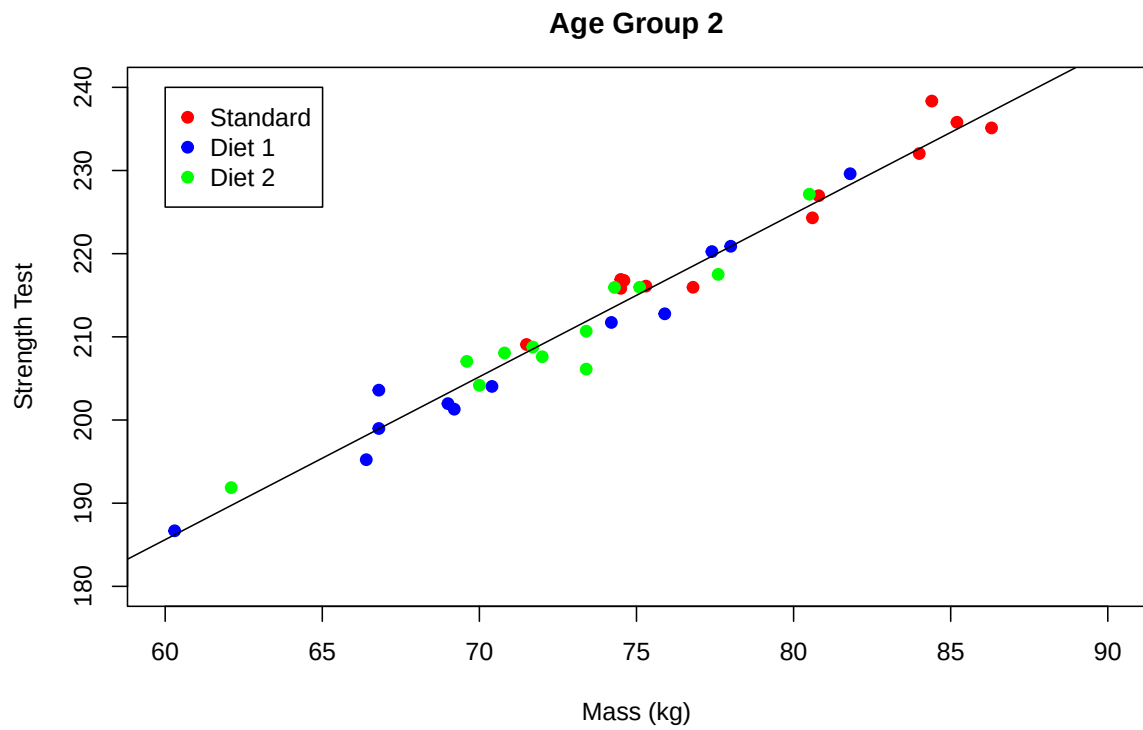
Backward selection stops at this point, with the model selected as

Age + Mass + Age:Mass. (BACKWARD)

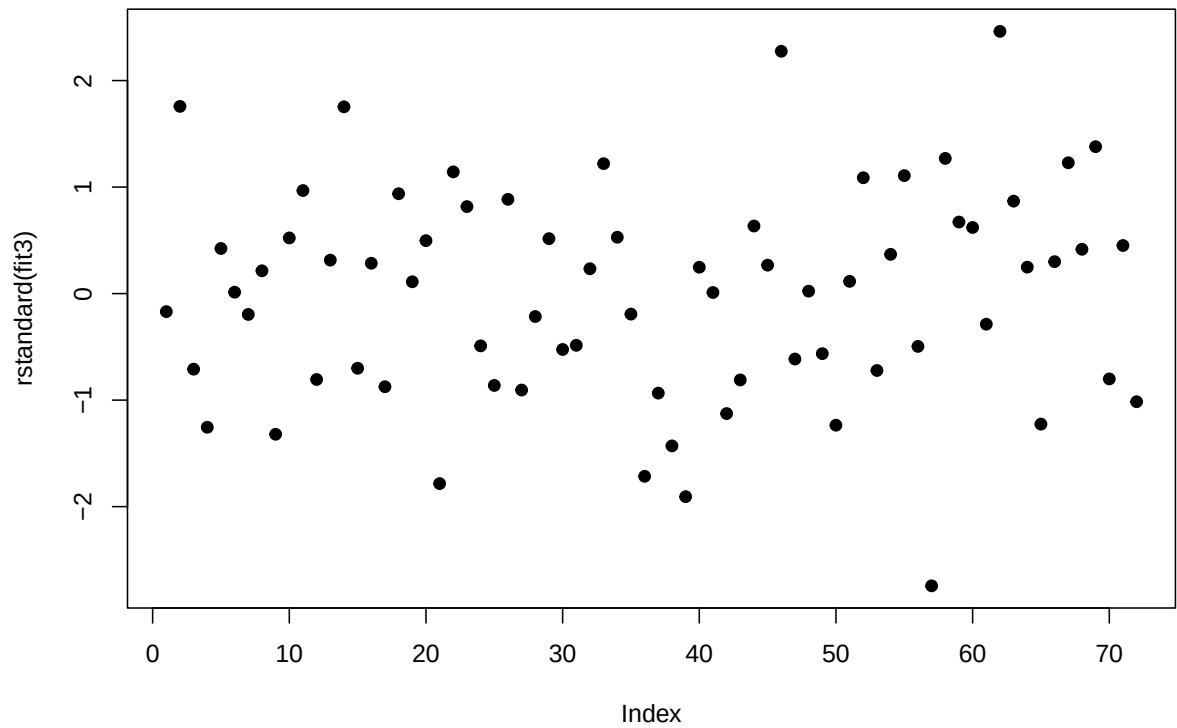
This model cannot be formally compared with the Forward selected model, as the two models are not nested. 8 Marks

- (c) We can first inspect the fits of the competing models (code here suppressed): first the Backward model

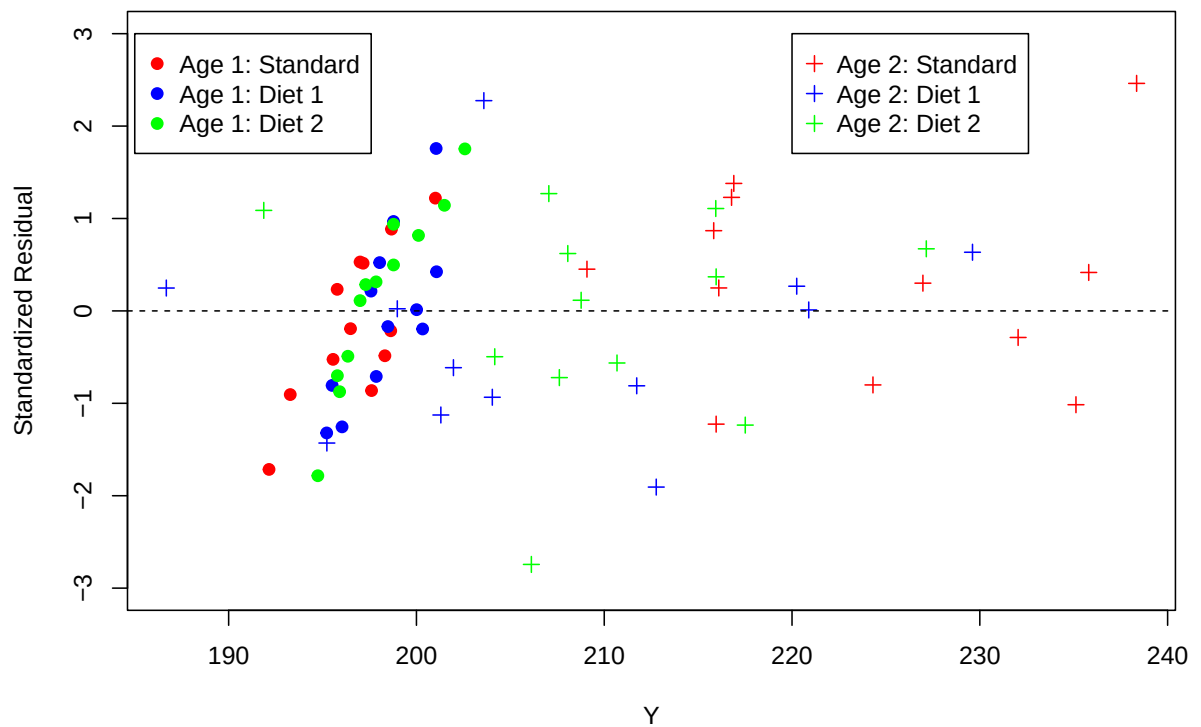




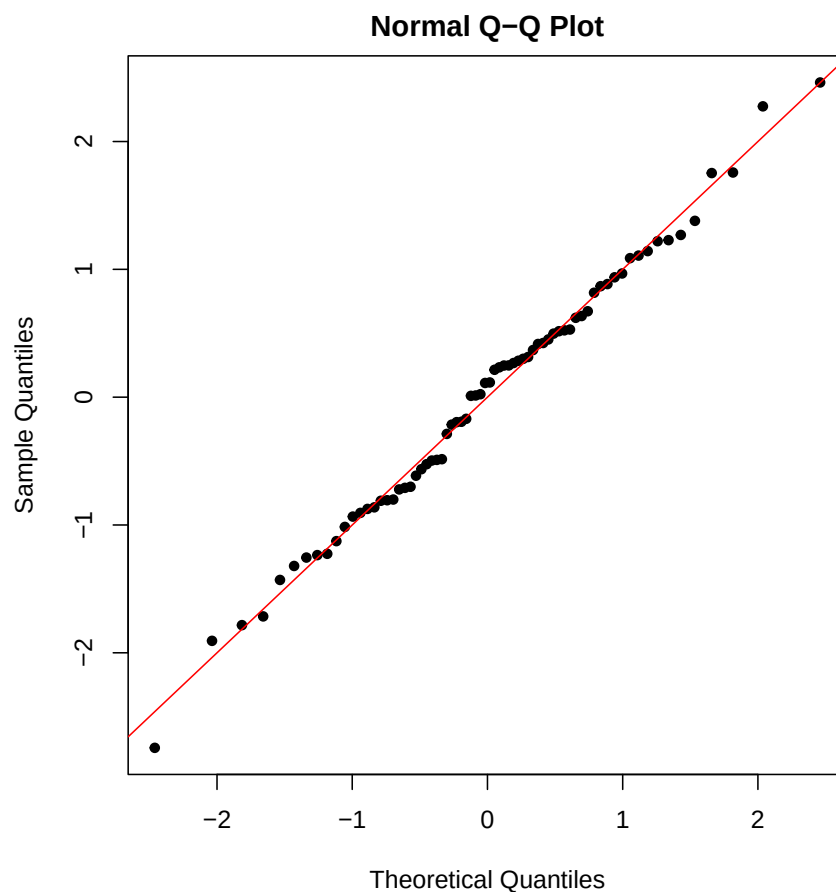
The fit is evidently quite good. For the residual plots, we begin with a standard plot of the individual residuals:



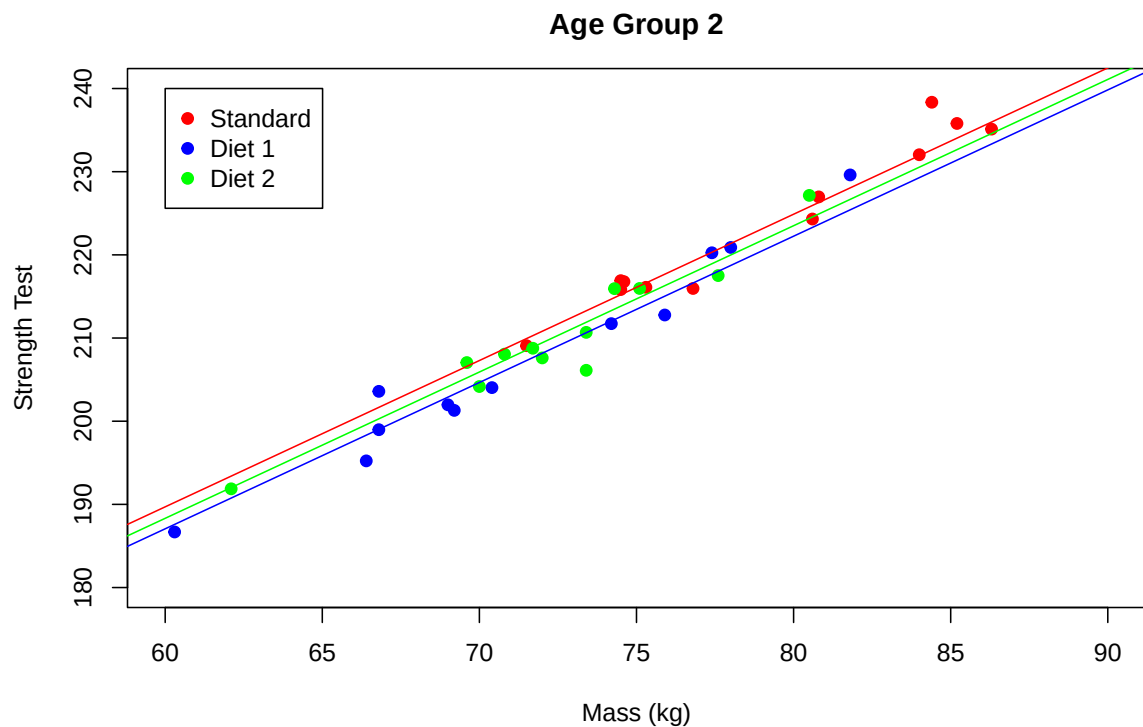
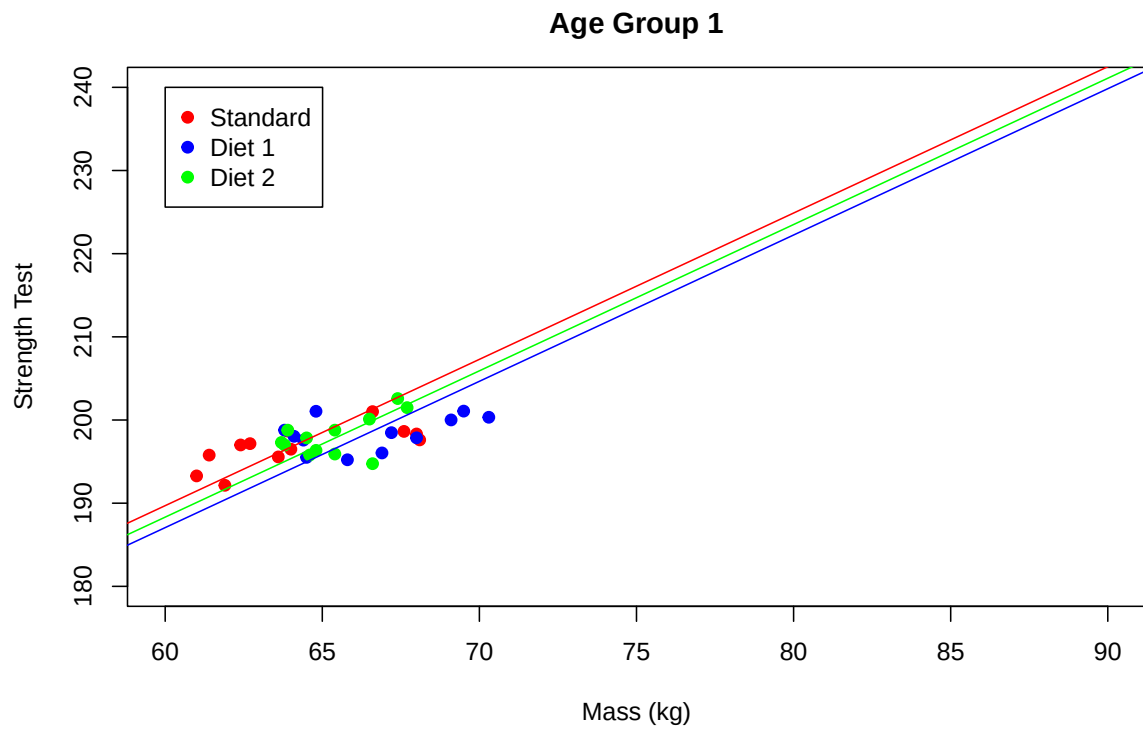
This patternless plot seems adequate. Now we plot against the Mass predictor



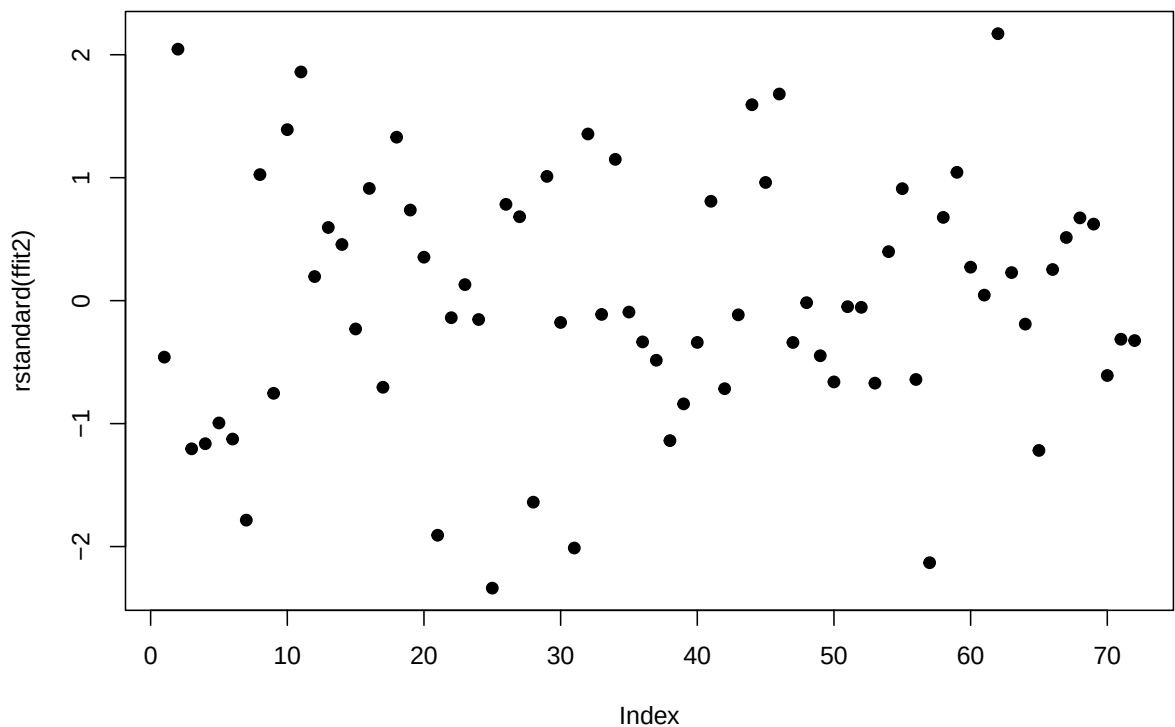
This is also adequate. A Q-Q plot can confirm that there are no serious deviations from normality.



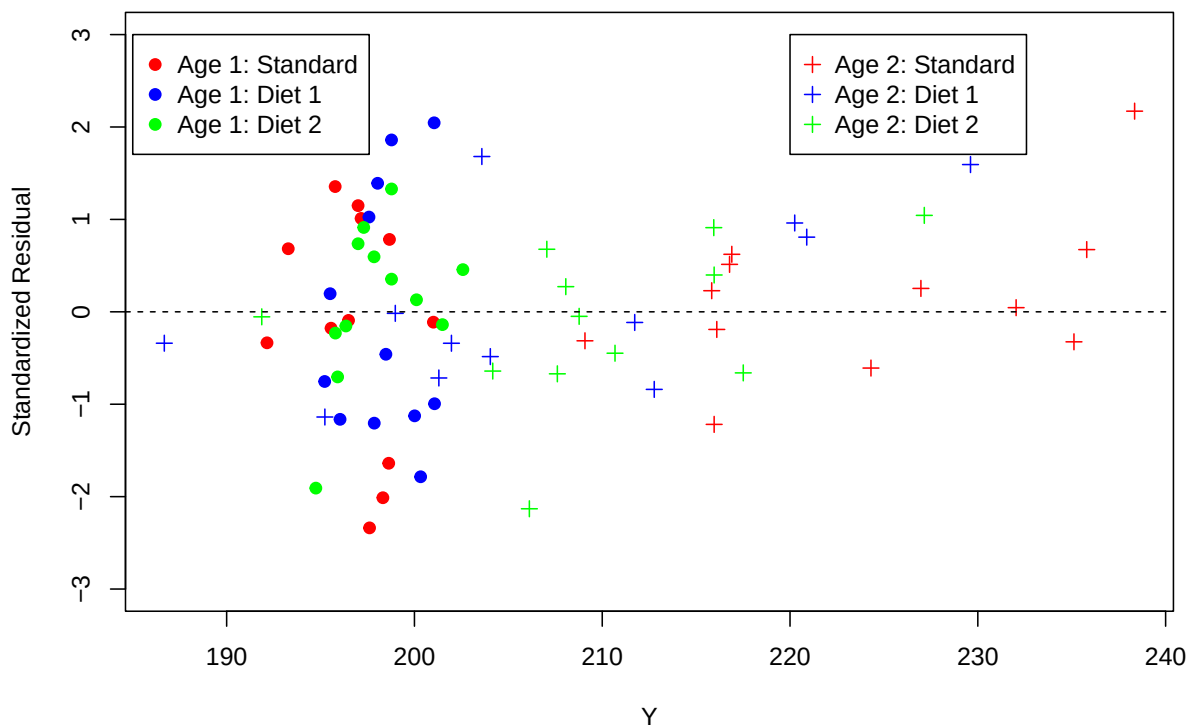
We now study the same plots for the Forward selected model.



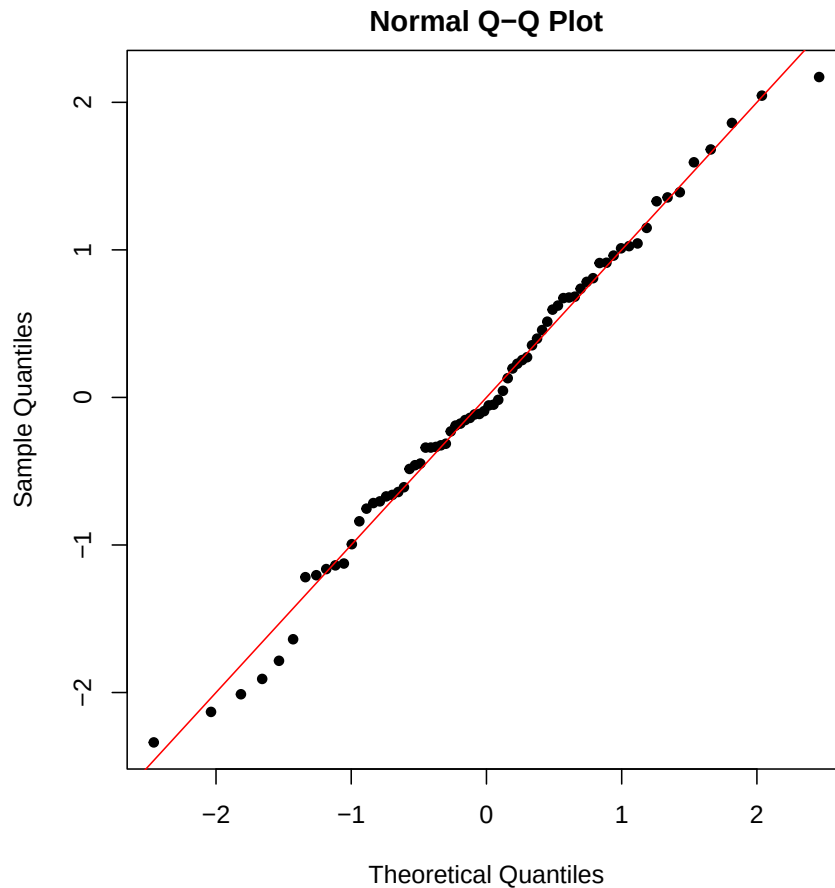
The fit is also quite good. For the residual plots, we begin with a standard plot of the individual residuals:



This patternless plot seems adequate. Now we plot against the Mass predictor



This is also adequate. A Q-Q plot can confirm that there are no serious deviations from normality. In summary, we really have two fairly competitive models, both of which do reasonably good jobs in explaining the observed variation in response. 4 Marks



Additional comments: The step function may be used to select the model automatically using an alternative approach based on the AIC metric: in this case, the Backward model is returned.

```
> fit.step<-step(fit1, direction='both', trace=0)
> summary(fit.step)

+
+ Call:
+ lm(formula = Y ~ Age + Mass + Age:Mass, data = DietStudy)
+
+ Residuals:
+   Min       1Q   Median       3Q      Max
+ -5.739 -1.543  0.236  1.295  4.947
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  158.079     10.168   15.55 < 2e-16 ***
+ Age2         -89.975     11.047   -8.15 1.2e-11 ***
+ Mass           0.606      0.155    3.90 0.00022 ***
+ Age2:Mass      1.352      0.166    8.15 1.2e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 2.12 on 68 degrees of freedom
+ Multiple R-squared:  0.97, Adjusted R-squared:  0.968
+ F-statistic: 724 on 3 and 68 DF, p-value: <2e-16
```

```

> anova(fit.step, fit3, test='F')
+ Analysis of Variance Table
+
+ Model 1: Y ~ Age + Mass + Age:Mass
+ Model 2: Y ~ Age + Mass + Age:Mass
+   Res.Df    RSS Df Sum of Sq F Pr(>F)
+ 1      68 306.3
+ 2      68 306.3  0          0

```

The `olsrr` library can be used to implement automatic F-test based selection

```

> # Stepwise selection using p-values (F-test based)
> fit.step2<-ols_step_both_p(fit1, pent = 0.05, prem = 0.05, details = FALSE)

+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
> summary(fit.step2$model)

+
+ Call:
+ lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
+     data = l)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -6.342  -1.750  -0.198   1.879   5.725
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)   84.1595     3.7728  22.31  <2e-16 ***
+ Mass          1.7591     0.0521  33.79  <2e-16 ***
+ RegimenDiet 1 -2.6406     0.8141  -3.24  0.0018 **
+ RegimenDiet 2 -1.3815     0.8140  -1.70  0.0942 .
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 2.78 on 68 degrees of freedom
+ Multiple R-squared:  0.948, Adjusted R-squared:  0.946
+ F-statistic: 413 on 3 and 68 DF, p-value: <2e-16

```

This is the Forward model.

Q3. The anticoagulant heparin is used for preventing blood clots of patients after surgery, and for the treatment and prevention of deep vein thrombosis. The most commonly used form of the drug requires careful monitoring, as misuse may lead to possible adverse outcomes. In a real clinical study concerning the efficacy of heparin, levels of the metabolite heparin sulphate in the blood of eligible recruited patients were monitored.

The file

<https://www.math.mcgill.ca/dstephens/204/Assignments/A2/HeparinDataSet.csv>

contains information on 90 patients (classified in the study via biological sex as 37 female, 53 male).

The data file contains five columns:

- **Sex:** Sex of patient (Female or Male) ;
- **Age:** Age of patient (years);
- **Weight:** Weight of patient (kg);
- **Dose:** Initial dose of heparin received (Low or High);
- **Heparin:** level of heparin sulphate in blood sample, measured 3 hours after dosing (relative scale).

Using these data, after selecting a suitable model, address the following questions (justifying your answers in each case):

- (a) Model selection step. 8 Marks
- (b) Does the Age of the patient influence the level of heparin sulphate in the blood ? 4 Marks
- (c) On average, do Male patients have a lower or higher level of heparin sulphate in their blood-stream compared to Female patients of the same weight and age, when both are dosed initially with a High dose ? 4 Marks
- (d) Predict the level of heparin sulphate that would be detected in the blood of a 40 year old male, weighing 100 kg, who is dosed on the Low dose. 4 Marks

```
> #R Code for data entry and elementary processing
> fname<-"https://www.math.mcgill.ca/dstephens/204/Assignments/A2/HeparinDataSet.csv"
> Heparin<-read.csv(fname)
> Heparin$Sex<-factor(Heparin$Sex,levels=c('Female','Male'))
> Heparin$Dose<-factor(Heparin$Dose,levels=c('Low','High'))
> str(Heparin)

+ 'data.frame': 90 obs. of 5 variables:
+ $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
+ $ Age      : int   51 48 65 44 43 54 44 36 43 55 ...
+ $ Weight   : num   146.9 115.1 83.5 94.9 122.5 ...
+ $ Dose     : Factor w/ 2 levels "Low","High": 2 2 1 2 2 1 1 2 2 1 ...
+ $ Heparin  : num   -1.65 1.49 1.76 1.32 -2.86 ...

> nrow(Heparin)

+ [1] 90

> table(Heparin$Sex, Heparin$Dose)

+
+           Low High
+  Female    22   31
+   Male     15   22
```

Solution:

- (a) We consider a sequence of nested models and compare them using the F-test procedure. In each case, we test the hypotheses:

H_0 : Reduced Model is Adequate Simplification of Complete Model

H_a : Reduced Model is Not Adequate Simplification of Complete Model

where the role of the “complete” and “reduced” depends on the comparison being carried out. Using backwards selection from the full model

Age*Sex*Dose*Weight (FULL)

including all interactions is possible. The ANOVA table for the full model is

```
> model.full<-lm(Heparin~Sex*Age*Weight*Dose,data=Heparin)
> summary(aov(model.full))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
+ Sex	1	244.2	244.2	267.07	< 2e-16 ***
+ Age	1	0.3	0.3	0.30	0.5825
+ Weight	1	258.4	258.4	282.60	< 2e-16 ***
+ Dose	1	32.4	32.4	35.39	8.3e-08 ***
+ Sex:Age	1	0.9	0.9	0.94	0.3365
+ Sex:Weight	1	6.8	6.8	7.46	0.0079 **
+ Age:Weight	1	1.8	1.8	1.93	0.1692
+ Sex:Dose	1	0.0	0.0	0.01	0.9186
+ Age:Dose	1	1.2	1.2	1.37	0.2462
+ Weight:Dose	1	0.5	0.5	0.55	0.4622
+ Sex:Age:Weight	1	3.5	3.5	3.82	0.0545 .
+ Sex:Age:Dose	1	0.0	0.0	0.01	0.9320
+ Sex:Weight:Dose	1	0.2	0.2	0.21	0.6472
+ Age:Weight:Dose	1	0.3	0.3	0.29	0.5936
+ Sex:Age:Weight:Dose	1	0.0	0.0	0.01	0.9368
+ Residuals	74	67.7	0.9		

+ Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

This listing can be used for exploratory purposes, but in this **unbalanced** study

```
> table(Heparin$Sex,Heparin$Dose)
```

	Low	High
+ Female	22	31
+ Male	15	22

and in the presence of a continuous covariate, the p-values in the ANOVA table should not be used to decide on the importance of a particular term. The sequential F-test procedure must be used for formal comparison.

The ANOVA table does, however, indicate that the three-way interactions are not significant in the model. Therefore, rather than doing full backward selection, it is justifiable to compare the model with all two-way interactions only, that is

Age + Sex + Dose + Weight + Age:Sex + Age:Dose + Age:Weight
+ Sex:Dose + Sex:Weight + Dose:Weight (MODEL 1)

This model can be written in R as follows:

```

> model.1<-lm(Heparin~(Sex+Age+Weight+Dose)^2,data=Heparin)
> anova(model.1,model.full,test="F")

+ Analysis of Variance Table
+
+ Model 1: Heparin ~ (Sex + Age + Weight + Dose)^2
+ Model 2: Heparin ~ Sex * Age * Weight * Dose
+   Res.Df  RSS Df Sum of Sq    F Pr(>F)
+ 1      79 71.63
+ 2      74 67.67   5      3.958 0.866  0.508

```

Thus we have

$$SSE_C = 67.674 \quad SSE_R = 71.632 \quad k - g = 5 \quad n - k - 1 = 74$$

so therefore $F = 0.8657$, which is not significant at the $\alpha = 0.05$ level, as the p-value is 0.5084. Thus the reduced model (MODEL 1) is an adequate simplification of the complete model (FULL).

We then proceed with a sequence of model simplifications, dropping one term at a time: first, using the drop1 function, we have

```

> drop1(model.1,test="F")

+ Single term deletions
+
+ Model:
+ Heparin ~ (Sex + Age + Weight + Dose)^2
+
+   Df Sum of Sq  RSS   AIC F value    Pr(>F)
+ <none>                 71.63   1.456
+ Sex:Age      1      0.945 72.58   0.635    1.042 0.31052
+ Sex:Weight   1      7.979 79.61   8.961    8.799 0.00398 **
+ Sex:Dose     1      0.035 71.67  -0.500    0.038 0.84572
+ Age:Weight   1      0.535 72.17   0.126    0.590 0.44472
+ Age:Dose     1      1.305 72.94   1.082    1.440 0.23376
+ Weight:Dose  1      0.500 72.13   0.082    0.551 0.46014
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

It appears that dropping any one of the two-way interactions apart from the Sex.Weight interaction does not lead to a significant depreciation in fit. Thus, we may attempt to drop those terms:

```

> model.2<-update(model.1,~.-Sex:Age-Sex:Dose-Age:Weight-Age:Dose-Weight:Dose)
> anova(model.2,model.1,test="F")

+ Analysis of Variance Table
+
+ Model 1: Heparin ~ Sex + Age + Weight + Dose + Sex:Weight
+ Model 2: Heparin ~ (Sex + Age + Weight + Dose)^2
+   Res.Df  RSS Df Sum of Sq    F Pr(>F)
+ 1      84 76.00
+ 2      79 71.63   5      4.363 0.962  0.446

```

This confirms that the reduced model

$$\text{Sex} + \text{Age} + \text{Weight} + \text{Dose} + \text{Sex:Weight} \quad (\text{MODEL2})$$

is an adequate simplification of (MODEL 1), with a p-value of 0.4461. We attempt to simplify this model further:

```

> drop1(model.2,test="F")
+ Single term deletions
+
+ Model:
+ Heparin ~ Sex + Age + Weight + Dose + Sex:Weight
+      Df Sum of Sq    RSS    AIC F value    Pr(>F)
+ <none>                 76.00 -3.223
+ Age      1         1.29  77.29 -3.707    1.426  0.23570
+ Dose     1        34.71 110.71 28.636   38.367 2.08e-08 ***
+ Sex:Weight 1         6.83  82.83  2.528    7.555  0.00732 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Of these possible one-term deletions, we appear to be able to remove the Age main effect, but no other term. Thus a further model update is attempted.

```

> model.3<-update(model.2,~.-Age)
> drop1(model.3,test="F")
+ Single term deletions
+
+ Model:
+ Heparin ~ Sex + Weight + Dose + Sex:Weight
+      Df Sum of Sq    RSS    AIC F value    Pr(>F)
+ <none>                 77.29 -3.707
+ Dose     1        33.68 110.97 26.849   37.045 3.22e-08 ***
+ Sex:Weight 1         7.00  84.29  2.097    7.699  0.00679 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Thus the model

Sex + Weight + Dose + Sex:Weight (MODEL3)

cannot be simplified without leading to a significant depreciation of fit. Thus we settle on (MODEL3) as the final model. A model summary is as follows:

```

> summary(model.3)
+
+ Call:
+ lm(formula = Heparin ~ Sex + Weight + Dose + Sex:Weight, data = Heparin)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -2.073 -0.636 -0.116  0.646  2.460
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)   8.83886    0.77862   11.35 < 2e-16 ***
+ SexMale       0.97531    1.08828    0.90  0.3727
+ Weight      -0.06739    0.00794   -8.49 5.9e-13 ***
+ DoseHigh      1.26682    0.20814    6.09 3.2e-08 ***
+ SexMale:Weight -0.02952    0.01064   -2.77  0.0068 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+

```

```

+ Residual standard error: 0.954 on 85 degrees of freedom
+ Multiple R-squared:  0.875, Adjusted R-squared:  0.869
+ F-statistic: 149 on 4 and 85 DF,  p-value: <2e-16

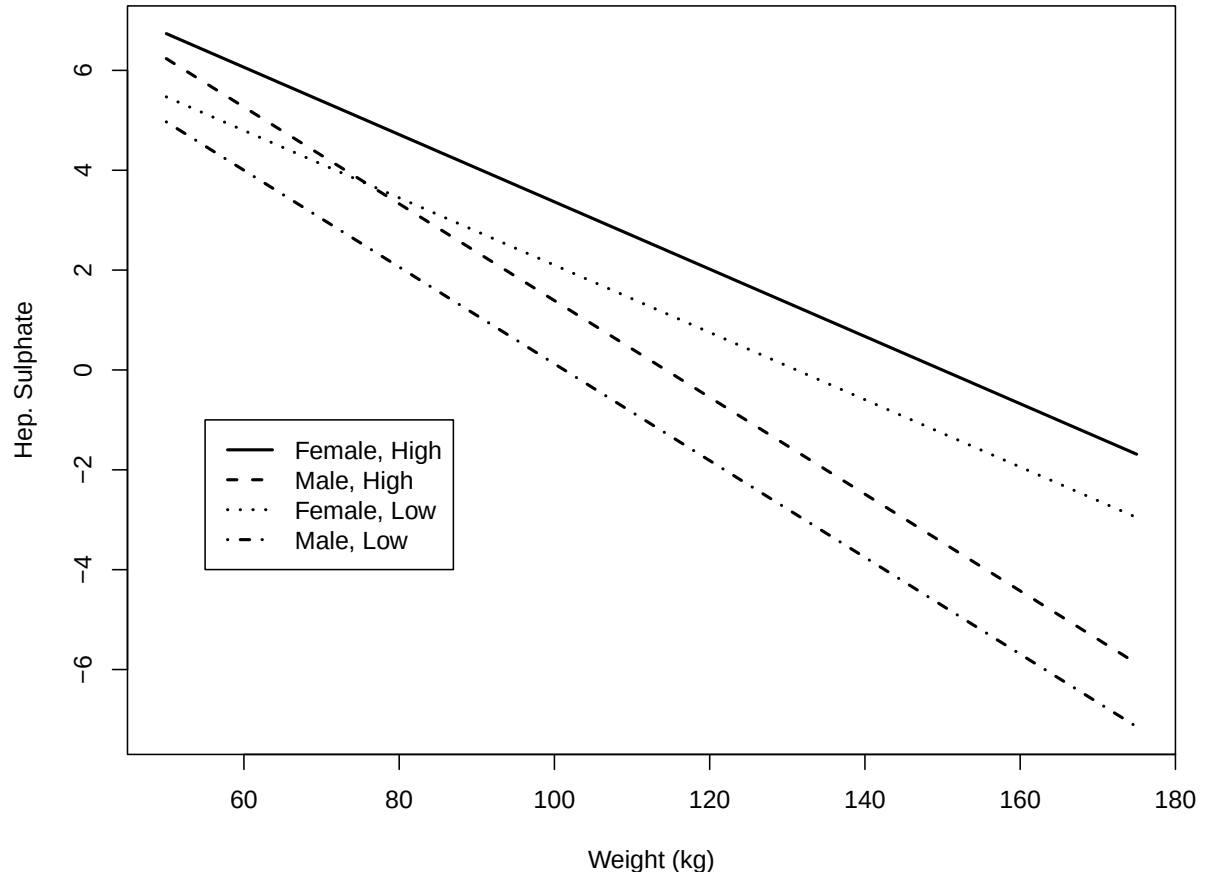
```

Here the **baseline** group is Female with High Dose. Due to the combinations of factors, we essentially have four subgroups, in each of which the effect of weight on heparin sulphate level is notable - Age is not influential. Thus we have a “four straight lines” model, as illustrated in the following plot: the lines for the different Dose levels are **parallel** for each of the two levels of Sex, as the Dose.Weight interaction is not included in the final model.

```

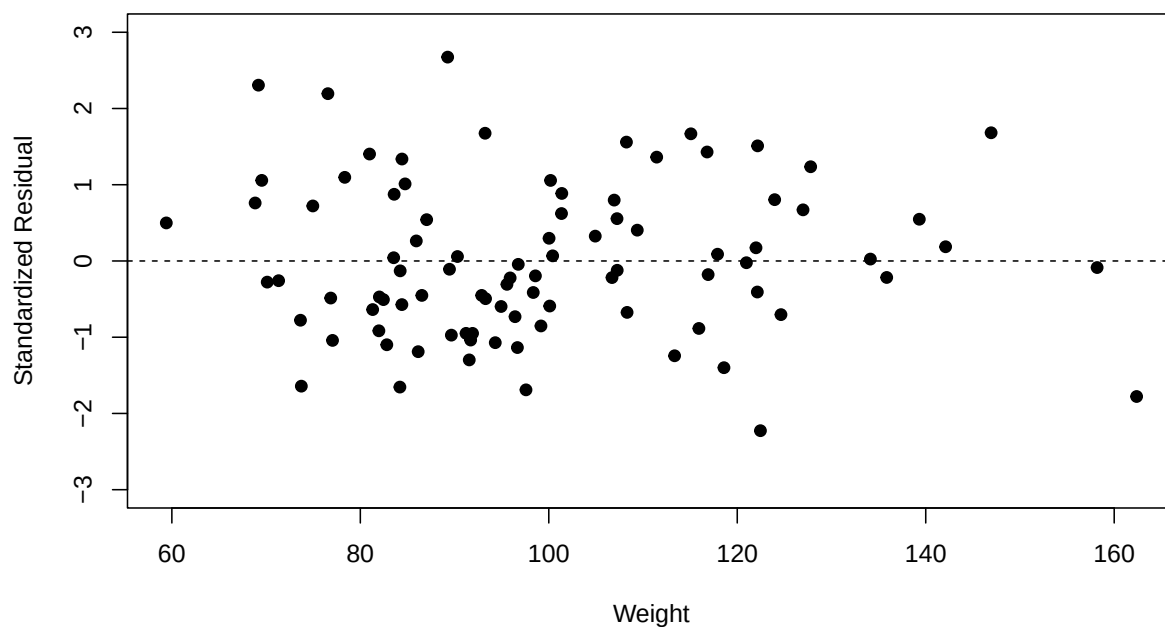
> W<-seq(50,175,by=1)
> D<-rep(rep(c("High","Low"),each=length(W)),2)
> S<-rep(c("Female","Male"),each=2*length(W))
> newd<-data.frame(Dose=as.factor(D),Sex=as.factor(S),Weight=rep(W,4))
> Ypred<-predict(model.3,newdata=newd)
>
> par(mar=c(4,4,1,1))
> plot(W,Ypred[D=="High" & S=="Female"],type="l",lwd=2,
+       ylim=range(Ypred),ylab="Hep. Sulphate",xlab="Weight (kg)")
> lines(W,Ypred[D=="High" & S=="Male"],lty=2,lwd=2)
> lines(W,Ypred[D=="Low" & S=="Female"],lty=3,lwd=2)
> lines(W,Ypred[D=="Low" & S=="Male"],lty=4,lwd=2)
> legend(55,-1,c("Female, High","Male, High","Female, Low","Male, Low"),
+       lty=1:4,lwd=2)

```

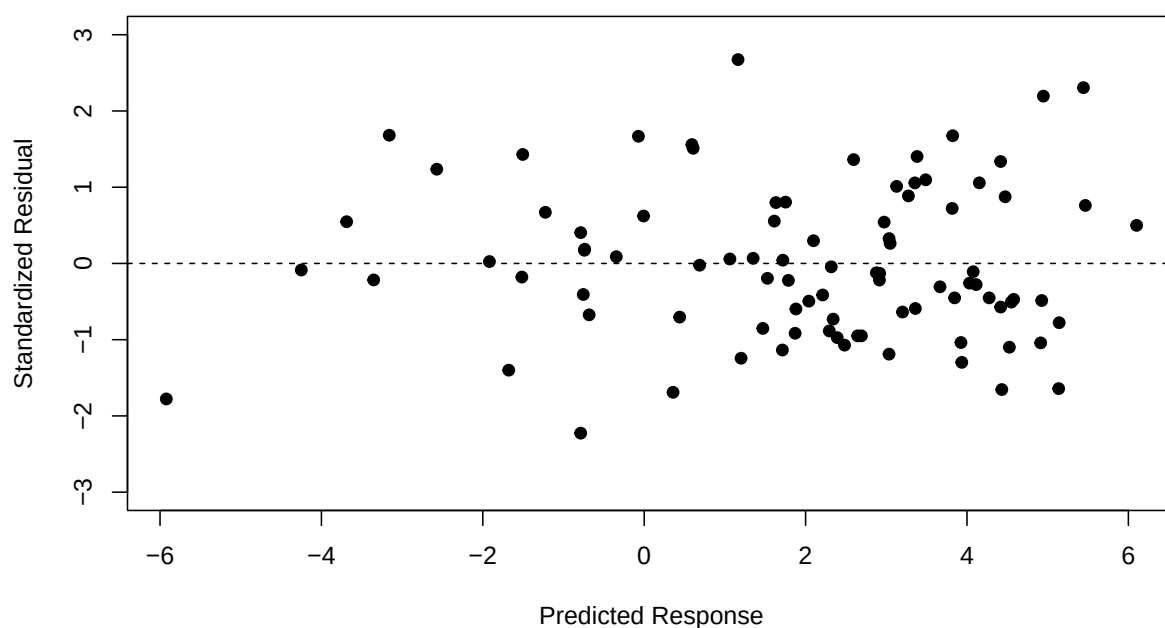


To check the fit we begin with the residuals:

```
> par(mar=c(4,4,2,1))
> plot(Heparin$Weight,rstandard(model.3),ylim=range(-3,3),
+      pch=19,xlab="Weight",ylab="Standardized Residual")
> abline(h=0,lty=2)
```

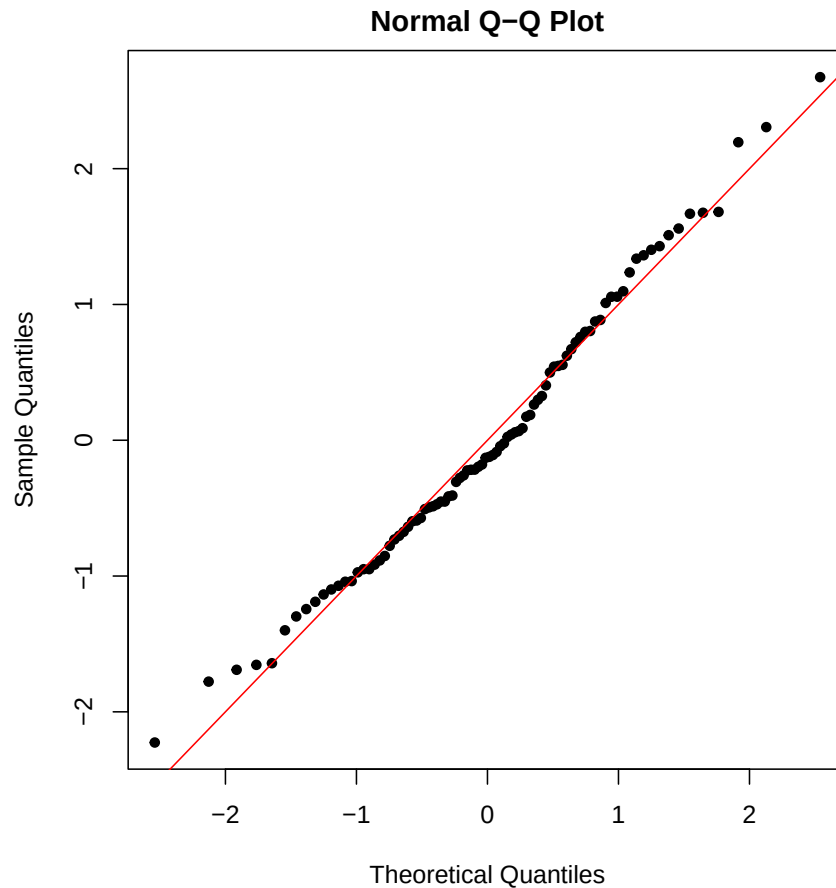


```
> plot(predict(model.3),rstandard(model.3),ylim=range(-3,3),
+      pch=19,xlab="Predicted Response",ylab="Standardized Residual")
> abline(h=0,lty=2)
```



Finally, a Q-Q plot can confirm that there are no serious deviations from normality.

```
> par(mar=c(4,4,2,1), pty='s')
> qqnorm(rstandard(model.3), pch=19, cex=0.8)
> abline(0,1,col='red')
```



Hence, overall, the model fits reasonably well, and the assumptions made about the residual errors seem to be justified.

(b) *Does the Age of the patient influence the level of heparin sulphate in the blood ?*

Age does not appear in my selected model, so Age does not influence the outcome. 4 Marks

(c) *Do Male patients have a lower or higher level of heparin sulphate in their bloodstream compared to Female patients of the same weight and age, when both are dosed initially with a High dose ?*

Here is the table of coefficients for the selected model:

```
> summary(model.3)$coef

+               Estimate Std. Error  t value    Pr(>|t|)
+ (Intercept)    8.8388578  0.77862396 11.351896 1.02812e-18
+ SexMale        0.9753098  1.08827888  0.896195 3.72680e-01
+ Weight        -0.0673871  0.00793993 -8.487114 5.89375e-13
+ DoseHigh       1.2668197  0.20813785  6.086446 3.21576e-08
+ SexMale:Weight -0.0295240  0.01064027 -2.774744 6.79142e-03

> be.hat<-coef(model.3)
```

Recalling that this is the contrast parameterization, we have the two modelled means as

$$\text{Females : } E[Y|X = x] = \beta_0 + \beta_{\text{HIGH}}^{(\text{DOSE})} + \beta_1 \text{ WEIGHT}$$

$$\text{Males : } E[Y|X = x] = \beta_0 + \beta_{\text{HIGH}}^{(\text{DOSE})} + \beta_{\text{MALE}}^{(\text{SEX})} + (\beta_1 + \beta_{1.\text{MALE}}^{(\text{SEX.WEIGHT})}) \text{ WEIGHT}$$

with the parameter estimates being

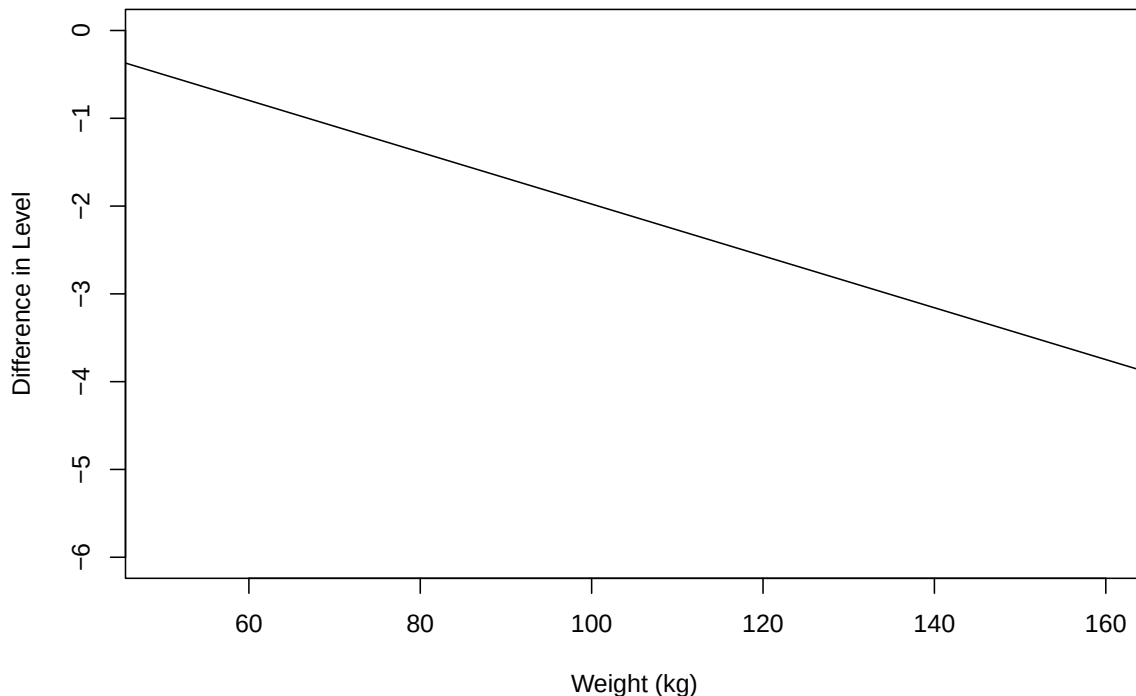
Intercept	$\hat{\beta}_0 = 8.8389$
SexMale	$\hat{\beta}_{\text{MALE}}^{(\text{SEX})} = 0.9753$
Weight	$\hat{\beta}_1 = -0.0674$
DoseHigh	$\hat{\beta}_{\text{HIGH}}^{(\text{DOSE})} = 1.2668$
SexMale:Weight	$\hat{\beta}_{1.\text{MALE}}^{(\text{SEX.WEIGHT})} = -0.0295$

Thus for fixed weight, the **difference** between Males and Females is then

$$\beta_{\text{MALE}}^{(\text{SEX})} + \beta_{1.\text{MALE}}^{(\text{SEX.WEIGHT})} \times \text{WEIGHT} = 0.9753 + (-0.0295) \times \text{WEIGHT}$$

We may display the difference in a plot

```
> w<-Heparin$Weight
> par(mar=c(4,4,2,2))
> plot(w,w*0,ylim=range(-6,0),xlim=c(50,160),type='n',
+      xlab='Weight (kg)', ylab='Difference in Level')
> abline(be.hat[2],be.hat[5])
```



For weights in the range 50-160kg, this difference is **negative**, and becomes more negative as weight increases. This Males do appear to have a **lower** heparin sulphate level. Whether this is a statistically significant difference cannot be answered without further calculation: at the lower end of the weight range we have

```

> W<-seq(60,80,by=10)
> D<-rep(c("High"),each=length(W))
> S<-rep(c("Female","Male"),length(W))
> newd2<-data.frame(Dose=as.factor(D),Sex=as.factor(S),Weight=rep(W,each=2))
> Ypred<-predict(model.3,newdata=newd2,interval="confidence")
> cbind.data.frame(newd2,Ypred)

+   Dose    Sex Weight      fit      lwr      upr
+ 1 High Female    60 6.06245 5.48604 6.63886
+ 2 High  Male    60 5.26632 4.48977 6.04287
+ 3 High Female    70 5.38858 4.93711 5.84005
+ 4 High  Male    70 4.29721 3.64450 4.94992
+ 5 High Female    80 4.71471 4.36080 5.06861
+ 6 High  Male    80 3.32810 2.78977 3.86643

```

These results imply that there is not a significant difference for weights below 70kg.

4 Marks

- (d) Predict the level of heparin sulphate that would be detected in the blood of a 40 year old male, weighing 100 kg, who is dosed on the Low dose.

We have using predict

```

> newd3<-data.frame(Dose=as.factor("Low"),Sex=as.factor("Male"),Weight=100)
> Ypred<-predict(model.3,newdata=newd3,interval="confidence")
> Ypred

+      fit      lwr      upr
+ 1 0.123054 -0.300448 0.546557

```

The prediction is given by plugging estimated parameter values into the model formula, recalling the baseline categories.

$$\beta_0 + \beta_{\text{MALE}}^{(\text{SEX})} + \beta_1 \text{WEIGHT} + \beta_{1.\text{MALE}}^{(\text{SEX}.\text{WEIGHT})} \text{WEIGHT}$$

that is

$$8.8389 + 0.9753 + (-0.0674) \times 100 + (-0.0295) \times 100 = 0.1231$$

4 Marks

Additional comments: The step function may be used to select the model automatically using an alternative approach based on the AIC metric:

```

> model.step<-step(model.full, direction='both', trace=0)
> summary(model.step)

+
+ Call:
+ lm(formula = Heparin ~ Sex + Age + Weight + Dose + Sex:Age +
+   Sex:Weight + Age:Weight + Sex:Age:Weight, data = Heparin)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -2.0988 -0.6120 -0.0392  0.6210  2.5225
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept)   -0.032534    3.487541   -0.01   0.9926

```

```

+ SexMale          12.175534    5.562555    2.19    0.0315 *
+ Age              0.195259    0.075103    2.60    0.0111 *
+ Weight           0.018293    0.034407    0.53    0.5964
+ DoseHigh         1.306065    0.203784    6.41    9.1e-09 ***
+ SexMale:Age      -0.252208    0.125357   -2.01    0.0476 *
+ SexMale:Weight   -0.151361    0.052972   -2.86    0.0054 **
+ Age:Weight       -0.001881    0.000737   -2.55    0.0125 *
+ SexMale:Age:Weight 0.002742    0.001190    2.30    0.0237 *
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.922 on 81 degrees of freedom
+ Multiple R-squared:  0.889, Adjusted R-squared:  0.878
+ F-statistic: 80.7 on 8 and 81 DF,  p-value: <2e-16

```

However, this produces a fairly complex model, and according to the F-test, it does not really improve on the selected model.

```

> anova(model.step, model.3, test='F')
+ Analysis of Variance Table
+
+ Model 1: Heparin ~ Sex + Age + Weight + Dose + Sex:Age + Sex:Weight +
+   Age:Weight + Sex:Age:Weight
+ Model 2: Heparin ~ Sex + Weight + Dose + Sex:Weight
+   Res.Df    RSS Df Sum of Sq    F Pr(>F)
+ 1      81 68.87
+ 2      85 77.29 -4    -8.411 2.473 0.0509 .
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The `olsrr` library can be used to implement automatic F-test based selection

```

> # Stepwise selection using p-values (F-test based)
> model.step2<-ols_step_both_p(model.full, pent = 0.05,
+   prem = 0.05, details = FALSE)
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
+ Note: model has aliased coefficients
+       sums of squares computed by model comparison
> summary(model.step2$model)
+
+ Call:
+ lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
+   data = l)
+
+ Residuals:
+   Min       1Q   Median       3Q      Max

```

```

+ -2.0845 -0.6317 -0.0816  0.6663  2.4985
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)   8.17689    0.95416   8.57  4.4e-13 ***
+ Weight       -0.06748    0.00792  -8.52  5.5e-13 ***
+ DoseHigh      1.29333    0.20880   6.19  2.1e-08 ***
+ Age           0.01493    0.01250   1.19   0.2357
+ SexMale       0.94981    1.08577   0.87   0.3842
+ Weight:SexMale -0.02918    0.01062  -2.75   0.0073 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.951 on 84 degrees of freedom
+ Multiple R-squared:  0.877, Adjusted R-squared:  0.87
+ F-statistic: 120 on 5 and 84 DF, p-value: <2e-16

```

We can see that the selected model includes **Age**, which is again not justified in a direct comparison with the earlier selected model.

```

> anova(model.step2$model,model.3, test='F')
+ Analysis of Variance Table
+
+ Model 1: Heparin ~ Weight + Sex:Weight + Dose + Age + Sex
+ Model 2: Heparin ~ Sex + Weight + Dose + Sex:Weight
+   Res.Df  RSS Df Sum of Sq    F Pr(>F)
+ 1      84 76.00
+ 2      85 77.29 -1      -1.29 1.426  0.236

```