# MATH 423/533 - ASSIGNMENT 4
# SOLUTIONS

## INTRODUCTION

This assignment concerns the use of factor predictors in linear regression modelling, and focusses on models with two factors $X_1$ and $X_2$ with $M_1$ and $M_2$ levels. Terminology that is commonly used is

- **one-way layout**: This means a data set/model with a **single** factor predictor; the two models that can be fitted are

| Model | Description | R |
|---|---|---|
| 1 | Intercept only | `lm(y~1)` |
| $1 + X_1$ | Main effect | `lm(y~x1)` |

- **two-way layout**: This means a data set/model with **two** factor predictors; the five models that can be fitted are

| Model | Description | R |
|---|---|---|
| 1 | Intercept only | `lm(y~1)` |
| $1 + X_1$ | Main effect of $X_1$ | `lm(y~x1)` |
| $1 + X_2$ | Main effect of $X_2$ | `lm(y~x2)` |
| $1 + X_1 + X_2$ | Main effects model | `lm(y~x1+x2)` |
| $1 + X_1 + X_2 + X_1.X_2$ | Main effects plus interaction | `lm(y~x1+x2+x1:x2)` |

The first four models are nested inside the main effects plus interaction model; the modelled mean for that model is

$$\beta_0 + \underbrace{\sum_{j=1}^{M_1-1} \beta^{\mathsf{C}}_{1j} \mathbb{1}_j(x_{i1})}_{\text{main effect of } X_1} + \underbrace{\sum_{l=1}^{M_2-1} \beta^{\mathsf{C}}_{2l} \mathbb{1}_l(x_{i2})}_{\text{main effect of } X_2} + \underbrace{\sum_{j=1}^{M_1-1} \sum_{l=1}^{M_2-1} \beta^{\mathsf{C}}_{12jl} \mathbb{1}_j(x_{i1}) \mathbb{1}_l(x_{i2})}_{\text{interaction}}.$$

For each data point, only one term in each summation is non-zero as

$$\mathbb{1}_j(x_{i1}) = 1 \Longleftrightarrow x_{i1} = j \qquad \mathbb{1}_j(x_{i1})\mathbb{1}_l(x_{i2}) = 1 \Longleftrightarrow x_{i1} = j \text{ and } x_{i2} = l.$$

As described in lectures, the default setting in R is to use this **contrast** parameterization; the estimates of the parameters

$$\beta_0, \beta^{\mathsf{C}}_{1j}, \beta^{\mathsf{C}}_{2l}, \beta^{\mathsf{C}}_{12jl}$$

are reported in the output of R. The default baseline level is the one with the first label when levels are ordered alphabetically.

1. The data set `TestScores.csv` contains data on standardized math test scores of 45 students from three Faculties in a University.

   (a) Using the `lm` and `anova` functions, assess whether there is any evidence that there is a difference between the test scores of students from the three Faculties. Justify your conclusions with suitable `R` output. 3 Marks

   (b) Report the estimated mean scores, with associated standard errors, for students from each of the three faculties. 3 Marks

2. The data set `Filter.csv` contains data on the noise emission level of 36 cars. The cars are categorized using the `carsize` factor predictor that takes three levels, and two different noise filters are studied – the filter factor predictor `type` therefore takes two levels (`normal filter` and `Octel filter`)

   (a) For these data, form a table containing the number of model parameters $p$ and the sum of squared residuals $SS_{Res}$ for the five models listed on page 1 in this two-way layout. 5 Marks

   (b) Using a standard (partial) F-test, report the result of a comparison of the two models

   "Reduced" : $\mathbb{E}[Y_i|\mathbf{x}_i]$ : `1 + carsize`
   "Full" : $\mathbb{E}[Y_i|\mathbf{x}_i]$ : `1 + carsize + type + carsize:type`

   Report the p-value from the test using the `pf()` function in `R`. For example, if the degrees of freedom of the Fisher-F distribution are 2 and 11, and the F statistic is 11.30, we compute the critical value and p-value as follows:

```
> df1<-2;df2<-11
> (crit.value<-qf(0.95,df1,df2))
[1] 3.982298
>
> fstat<-11.30
> (pvalue<-1-pf(fstat,df1,df2))
[1] 0.00215176
```

   3 Marks

3. The data set `PatSat.csv` contains information on patient satisfaction for 25 patients having undergone treatment at a hospital for the same condition. There are four predictors: `Age` (age of patient in years), `Severity` (severity score for condition) and `Anxiety` (anxiety score for patient) are continuous predictors, whereas `Surgery` is a factor predictor with two levels (`No` and `Yes`) recording whether surgery was needed.

   Is there any evidence in these data that having surgery (as opposed to not having surgery) significantly affected patient satisfaction ? Justify your answer using linear modelling and statistical testing, making sure that you include in your modelling all predictors that influence the outcome measure. 6 Marks

   *(Hint: a simple comparison of responses for the two surgery groups may not be sufficient to answer the research question if age, severity or anxiety also influence the outcome.)*

## EXTRA QUESTION FOR STUDENTS IN MATH 533

Compute the matrix $\mathbf{X}^\top\mathbf{X}$ for

(a) the main effect model in Q1;

(b) the main effects only model in Q2;

(c) the main effects plus interaction model in Q2

and hence comment on the orthogonality of the predictors in each case.

Using a linear transformation, construct an orthogonal parameterization/predictor set for (a), such that in the new parameterization $\mathbf{X}^\top\mathbf{X}$ is a diagonal matrix.

*Hint: look up* ***polynomial contrasts*** *and how to implement them in* R.

5 Marks

# SOLUTIONS

1. The data set `TestScores.csv` contains data on standardized math test scores of 45 students from three Faculties in a University.

   (a) With the `TestScores.csv` file stored in the local directory, we implement as follows:

```
Q1data<-read.csv('TestScores.csv',header=TRUE)
Q1data$Faculty<-as.factor(Q1data$Faculty)
fitQ1<-lm(Score~Faculty,data=Q1data)
summary(fitQ1)

:
: Call:
: lm(formula = Score ~ Faculty, data = Q1data)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -15.800  -2.200   1.133   3.800   9.133
:
: Coefficients:
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept) 35.80000    1.59589  22.433  < 2e-16 ***
: Faculty2     0.06667    2.25694   0.030    0.977
: Faculty3    12.40000    2.25694   5.494 2.11e-06 ***
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 6.181 on 42 degrees of freedom
: Multiple R-squared:  0.488,Adjusted R-squared:  0.4636
: F-statistic: 20.02 on 2 and 42 DF,  p-value: 7.843e-07

anova(fitQ1)

: Analysis of Variance Table
:
: Response: Score
:           Df Sum Sq Mean Sq F value    Pr(>F)
: Faculty    2 1529.4  764.69  20.016 7.843e-07 ***
: Residuals 42 1604.5   38.20
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   We conclude, on the basis of the result of the global F test, with $p$-value equal to $7.8433837 \times 10^{-7}$, that there **is** a significant difference between the scores for faculties. To make this statement conclusively, we should also check the residual plots to assess the validity of the model assumptions, in particular the constant variance assumption (there is one mean per group, so the residuals will be zero mean in all groups) – see plot below, which indicates perhaps a smaller variance in Faculty 3 observations, although in the small sample this cannot be conclusively assessed.

   3 Marks

   (b) This can be done in two ways; either using the model formula removing the intercept

```
fitQ2<-lm(Score~-1+Faculty,data=Q1data)
summary(fitQ2)$coef[,1:2]

:          Estimate Std. Error
: Faculty1 35.80000   1.595894
: Faculty2 35.86667   1.595894
: Faculty3 48.20000   1.595894
```

or by using the linear transformation

$$\beta_1^G = \beta_0 \qquad \beta_2^G = \beta_0 + \beta_1^C \qquad \beta_3^G = \beta_0 + \beta_2^C$$

that is, using matrix

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

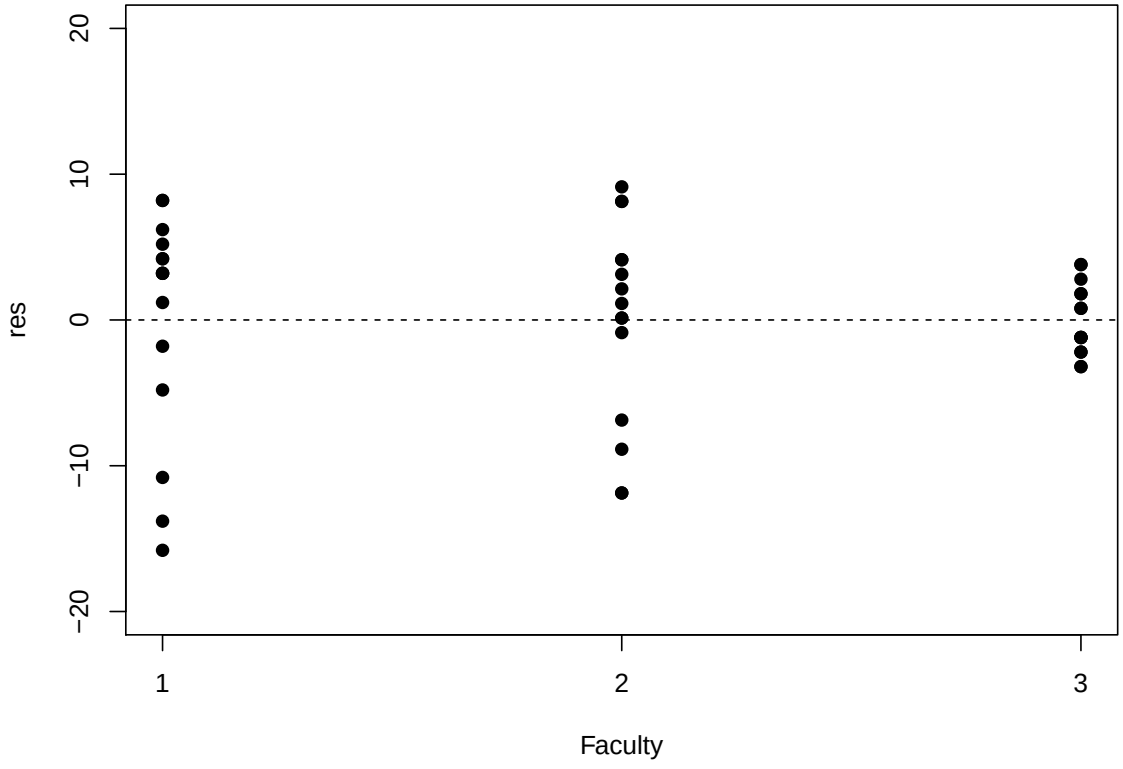
so that $\beta_G = \mathbf{L}\beta$. We can therefore compute from scratch:

```
L<-matrix(c(1,1,1,0,1,0,0,0,1),3,3,byrow=F)
X<-cbind(1,Q1data$Faculty=='2',Q1data$Faculty=='3')
beta.Sigma<-summary(fitQ1)$sigma^2 * solve(t(X)%*%X)
(betaG.ests<-L %*% coef(fitQ1))                          #Estimates

:          [,1]
: [1,] 35.80000
: [2,] 35.86667
: [3,] 48.20000

betaG.Sigma<-L%*%beta.Sigma%*%t(L)
sqrt(diag(betaG.Sigma))                                  #Standard errors

: [1] 1.595894 1.595894 1.595894
```

3 Marks

The residuals plot indicates that the residuals for Faculty 3 seem to perhaps have lower variance.

```
res.data<-data.frame(res=residuals(fitQ1),Faculty = Q1data$Faculty)
par(mar=c(4,4,1,2))
stripchart(res~Faculty,res.data,pch=19,vertical=T,ylim=range(-20,20),xlab='Faculty')
abline(h=0,lty=2)
```



2. (a) Computing these quantities is straightforward in R:

```
Q2data<-read.csv('Filter.csv',header=TRUE)
mod1<-lm(noise~1,data=Q2data);SS1<-round(anova(mod1)[1,2]/1000,3)
mod2<-lm(noise~carsize,data=Q2data);SS2<-round(anova(mod2)[2,2]/1000,3)
mod3<-lm(noise~type,data=Q2data);SS3<-round(anova(mod3)[2,2]/1000,3)
mod4<-lm(noise~carsize+type,data=Q2data);SS4<-round(anova(mod4)[3,2]/1000,3)
mod5<-lm(noise~carsize*type,data=Q2data);SS5<-round(anova(mod5)[4,2]/1000,3)
```

We may then put them in a table as follows

| Model | $SS_{Res}(\times 10^{-3})$ | $p$ |
|---|---|---|
| $1$ | 29.874 | 1 |
| $1+X_1$ | 3.823 | 3 |
| $1+X_2$ | 28.818 | 2 |
| $1+X_1+X_2$ | 2.767 | 4 |
| $1+X_1+X_2+X_1:X_2$ | 1.963 | 6 |

5 Marks

(b) We can do this easily from first principles

```
n<-nrow(Q2data)
fit1<-lm(noise~carsize,data=Q2data)
fit2<-lm(noise~carsize*type,data=Q2data)
SSRes1<-sum(residuals(fit1)^2)
p1<-length(coef(fit1))
SSRes2<-sum(residuals(fit2)^2)
p2<-length(coef(fit2))
(fstat<-((SSRes1-SSRes2)/(p2-p1))/(SSRes2/(n-p2)))

: [1] 9.47983

(pvalue<-1-pf(fstat,p2-p1,n-p2))

: [1] 0.0001460971
```

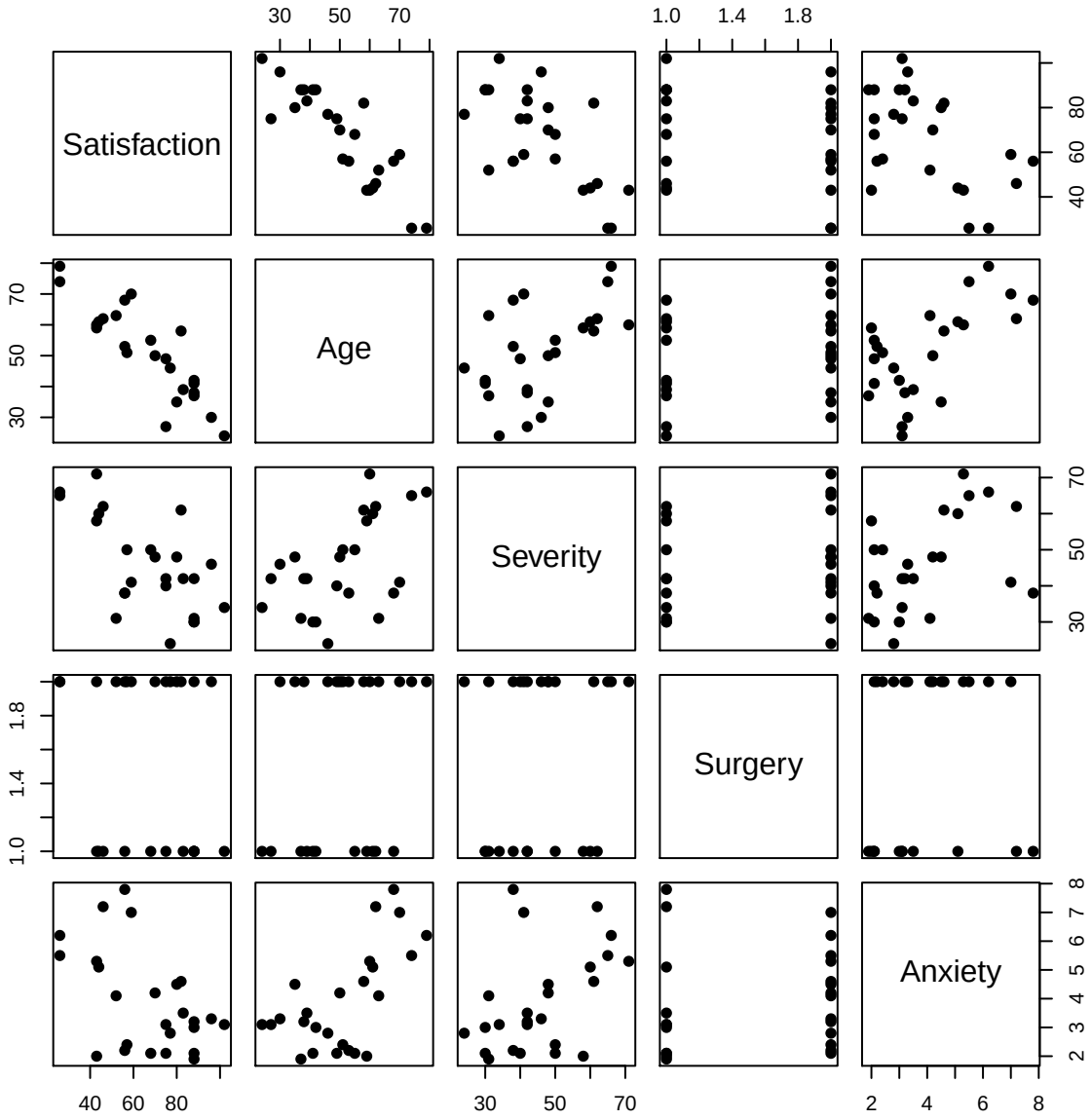
or using `anova` in R.

```
anova(fit1,fit2,test='F')

: Analysis of Variance Table
:
: Model 1: noise ~ carsize
: Model 2: noise ~ carsize * type
:   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
: 1     33 3822.9
: 2     30 1962.5  3    1860.4 9.4798 0.0001461 ***
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3 Marks

3. First we can explore the structure numerically (using correlation) and graphically (using a scatterplot):

```
Q3data<-read.csv('PatSat.csv',header=TRUE)
pairs(Q3data,pch=19)
```




```
Q3num<-Q3data;Q3num$Surgery<-as.numeric(Q3num$Surgery)-1
cor(Q3num)
```

```
:              Satisfaction        Age   Severity    Surgery    Anxiety
: Satisfaction    1.0000000 -0.8707049 -0.6531434 -0.1822682 -0.5127287
: Age            -0.8707049  1.0000000  0.5290246  0.2456932  0.6212453
: Severity       -0.6531434  0.5290246  1.0000000  0.1775101  0.4471567
: Surgery        -0.1822682  0.2456932  0.1775101  1.0000000  0.1096486
: Anxiety        -0.5127287  0.6212453  0.4471567  0.1096486  1.0000000
```

There appear to be some strong associations between the variables.

In a direct comparison between surgery groups, it seems that there is no effect of surgery on satisfaction level

```
fit0<-lm(Satisfaction~Surgery,data=Q3data)
print(summary(fit0), concise=TRUE)
```

```
:
: Call: lm(formula = Satisfaction ~ Surgery, data = Q3data)
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept)   71.000      6.433  11.036   <1e-04 ***
: SurgeryYes    -7.643      8.597  -0.889    0.383
:
: Residual standard error: 21.34 on 23 degrees of freedom
: Multiple R-squared:  0.03322,Adjusted R-squared:  -0.008812
: F-statistic: 0.7904 on 1 and 23 DF,  p-value: 0.3832
```

The task now is to assess whether any effect is being masked by the other variables. We start with the main effects model:

```
fit1<-lm(Satisfaction~Age+Severity+Surgery+Anxiety,data=Q3data)
print(summary(fit1), concise=TRUE)
```

```
:
: Call: lm(formula = Satisfaction ~ Age + Severity + Surgery + Anxiety,
:     data = Q3data)
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept) 140.1689     8.3191  16.849   <1e-04 ***
: Age          -1.1428     0.1904  -6.002   <1e-04 ***
: Severity     -0.4699     0.1866  -2.518   0.0204 *
: SurgeryYes    2.2259     4.1402   0.538   0.5968
: Anxiety       1.2673     1.4922   0.849   0.4058
:
: Residual standard error: 9.921 on 20 degrees of freedom
: Multiple R-squared:  0.8183,Adjusted R-squared:  0.7819
: F-statistic: 22.51 on 4 and 20 DF,  p-value: < 1e-04
```

```
drop1(fit1,test='F')
```

```
: Single term deletions
:
: Model:
: Satisfaction ~ Age + Severity + Surgery + Anxiety
:          Df Sum of Sq    RSS    AIC F value   Pr(>F)
: <none>                 1968.5 119.15
: Age       1    3545.1 5513.7 142.90 36.0182 7.22e-06 ***
: Severity  1     624.1 2592.6 124.04  6.3408  0.02043 *
: Surgery   1      28.4 1997.0 117.51  0.2890  0.59677
: Anxiety   1      71.0 2039.5 118.04  0.7212  0.40579
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems we can drop the predictors `Surgery` and `Anxiety`: we update the model as follows by omitting these variables, and computing the summaries and comparison statistics.

```
fit2<-update(fit1,~.-Surgery-Anxiety)
print(summary(fit2), concise=TRUE)


:
: Call: lm(formula = Satisfaction ~ Age + Severity, data = Q3data)
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept) 139.9233     8.1002  17.274   <1e-04 ***
: Age          -1.0462     0.1573  -6.652   <1e-04 ***
: Severity     -0.4359     0.1788  -2.439   0.0233 *
:
: Residual standard error: 9.682 on 22 degrees of freedom
: Multiple R-squared:  0.8096,Adjusted R-squared:  0.7923
: F-statistic: 46.77 on 2 and 22 DF,  p-value: < 1e-04


anova(fit2,fit1,test='F')


: Analysis of Variance Table
:
: Model 1: Satisfaction ~ Age + Severity
: Model 2: Satisfaction ~ Age + Severity + Surgery + Anxiety
:   Res.Df    RSS Df Sum of Sq      F Pr(>F)
: 1     22 2062.3
: 2     20 1968.5  2    93.754 0.4763  0.628


AIC(fit1,fit2)


:      df      AIC
: fit1  6 192.1011
: fit2  4 189.2643
```

We now check whether the introduction of interactions has any affect:

```
fit3<-update(fit2,~.+Surgery*Age*Anxiety)
fit4<-update(fit3,~.-Surgery:Age:Anxiety)
anova(fit2,fit4,fit3,test='F')


: Analysis of Variance Table
:
: Model 1: Satisfaction ~ Age + Severity
: Model 2: Satisfaction ~ Age + Severity + Surgery + Anxiety + Age:Surgery +
:     Surgery:Anxiety + Age:Anxiety
: Model 3: Satisfaction ~ Age + Severity + Surgery + Anxiety + Age:Surgery +
:     Surgery:Anxiety + Age:Anxiety + Age:Surgery:Anxiety
:   Res.Df    RSS Df Sum of Sq      F Pr(>F)
: 1     22 2062.3
: 2     17 1739.6  5    322.72 0.5978 0.7023
: 3     16 1727.5  1     12.06 0.1117 0.7425


AIC(fit2,fit4,fit3)


:      df      AIC
: fit2  4 189.2643
: fit4  9 195.0097
: fit3 10 196.8358
```

These models do not improve the fit, it seems, so we attempt other models:

```
fit5<-update(fit2,~.+Age:Severity)
fit6<-update(fit5,~.+Surgery)
anova(fit2,fit5,fit6,test='F')
```

```
: Analysis of Variance Table
:
: Model 1: Satisfaction ~ Age + Severity
: Model 2: Satisfaction ~ Age + Severity + Age:Severity
: Model 3: Satisfaction ~ Age + Severity + Surgery + Age:Severity
:   Res.Df    RSS Df Sum of Sq      F Pr(>F)
: 1     22 2062.3
: 2     21 2032.7  1    29.549 0.2929 0.5944
: 3     20 2017.8  1    14.967 0.1483 0.7042
```

```
AIC(fit2,fit5,fit6)
```

```
:      df      AIC
: fit2  4 189.2643
: fit5  5 190.9035
: fit6  6 192.7187
```

```
BIC(fit2,fit5,fit6)
```

```
:      df      BIC
: fit2  4 194.1398
: fit5  5 196.9979
: fit6  6 200.0320
```

The model `Age + Severity` still seems to be preferred. We can also attempt some automatic methods for selection using the `step` function, and different starting models: we attempt comparisons using AIC for convenience, although other methods can be used.

```
step.fit1<-step(lm(Satisfaction~Age*Severity*Anxiety*Surgery,data=Q3data),k=2,trace=0)
print(summary(step.fit1),concise=T)
```

```
:
: Call: lm(formula = Satisfaction ~ Age + Severity + Anxiety + Surgery +
:     Age:Severity + Age:Anxiety + Severity:Anxiety + Age:Surgery +
:     Severity:Surgery + Anxiety:Surgery + Age:Severity:Anxiety +
:     Age:Anxiety:Surgery, data = Q3data)
:                          Estimate Std. Error t value Pr(>|t|)
: (Intercept)            239.72040  160.09378   1.497    0.160
: Age                     -2.01705    3.34228  -0.603    0.557
: Severity                -5.34778    4.21765  -1.268    0.229
: Anxiety                -25.67624   39.13060  -0.656    0.524
: SurgeryYes             132.08818  107.10595   1.233    0.241
: Age:Severity             0.06882    0.07897   0.872    0.401
: Age:Anxiety              0.41667    0.67947   0.613    0.551
: Severity:Anxiety         1.22388    0.95373   1.283    0.224
: Age:SurgeryYes          -3.34196    2.14892  -1.555    0.146
: Severity:SurgeryYes      1.03425    0.60872   1.699    0.115
: Anxiety:SurgeryYes     -40.96403   33.53151  -1.222    0.245
: Age:Severity:Anxiety    -0.02000    0.01639  -1.220    0.246
: Age:Anxiety:SurgeryYes   0.72484    0.54702   1.325    0.210
:
: Residual standard error: 10.36 on 12 degrees of freedom
: Multiple R-squared:  0.8811,Adjusted R-squared:  0.7623
: F-statistic: 7.412 on 12 and 12 DF,  p-value: 0.0007637
```

Starting with the model with up to the four-way interaction, the model gets simplified to the model above that includes some third order interactions. The $R^2$ statistic is up to 0.8811, although the adjusted $R^2$ statistic is lower, at 0.7623. However, more compellingly, the AIC value is 197.486, indicating an inferior model to the `fit2` model above. We also note that some of the estimated standard errors are large, indicating some multicollinearity and variance inflation.

We now attempt some simplification by dropping the three-way interactions

```
step.fit11<-update(step.fit1,~.-Age:Severity:Anxiety-Age:Anxiety:Surgery)
anova(step.fit11,step.fit1)


: Analysis of Variance Table
:
: Model 1: Satisfaction ~ Age + Severity + Anxiety + Surgery + Age:Severity +
:     Age:Anxiety + Severity:Anxiety + Age:Surgery + Severity:Surgery +
:     Anxiety:Surgery
: Model 2: Satisfaction ~ Age + Severity + Anxiety + Surgery + Age:Severity +
:     Age:Anxiety + Severity:Anxiety + Age:Surgery + Severity:Surgery +
:     Anxiety:Surgery + Age:Severity:Anxiety + Age:Anxiety:Surgery
:   Res.Df    RSS Df Sum of Sq      F Pr(>F)
: 1     14 1520.1
: 2     12 1287.5  2    232.56 1.0838 0.3693


AIC(fit2,step.fit11)


:            df      AIC
: fit2        4 189.2643
: step.fit11 12 197.6376
```

which seems a legitimate simplification, and then re-attempting an automatic fit using `step` but with a different starting model including the two-way interactions:

```
step.fit2<-step(lm(Satisfaction~(Age+Severity+Anxiety+Surgery)^2,data=Q3data),k=2,trace=0
print(summary(step.fit2),concise=T)


:
: Call: lm(formula = Satisfaction ~ Age + Severity + Surgery + Age:Surgery +
:     Severity:Surgery, data = Q3data)
:                     Estimate Std. Error t value Pr(>|t|)
: (Intercept)         145.1381    11.4691  12.655  < 1e-04 ***
: Age                  -0.7182     0.2575  -2.789  0.01169 *
: Severity             -0.9342     0.3122  -2.993  0.00748 **
: SurgeryYes           -5.2017    16.2534  -0.320  0.75243
: Age:SurgeryYes       -0.4996     0.3260  -1.532  0.14194
: Severity:SurgeryYes   0.7085     0.3771   1.879  0.07571 .
:
: Residual standard error: 9.443 on 19 degrees of freedom
: Multiple R-squared:  0.8436,Adjusted R-squared:  0.8024
: F-statistic: 20.49 on 5 and 19 DF,  p-value: < 1e-04


AIC(fit2,fit5,step.fit2)


:           df      AIC
: fit2       4 189.2643
: fit5       5 190.9035
: step.fit2  7 190.3508
```

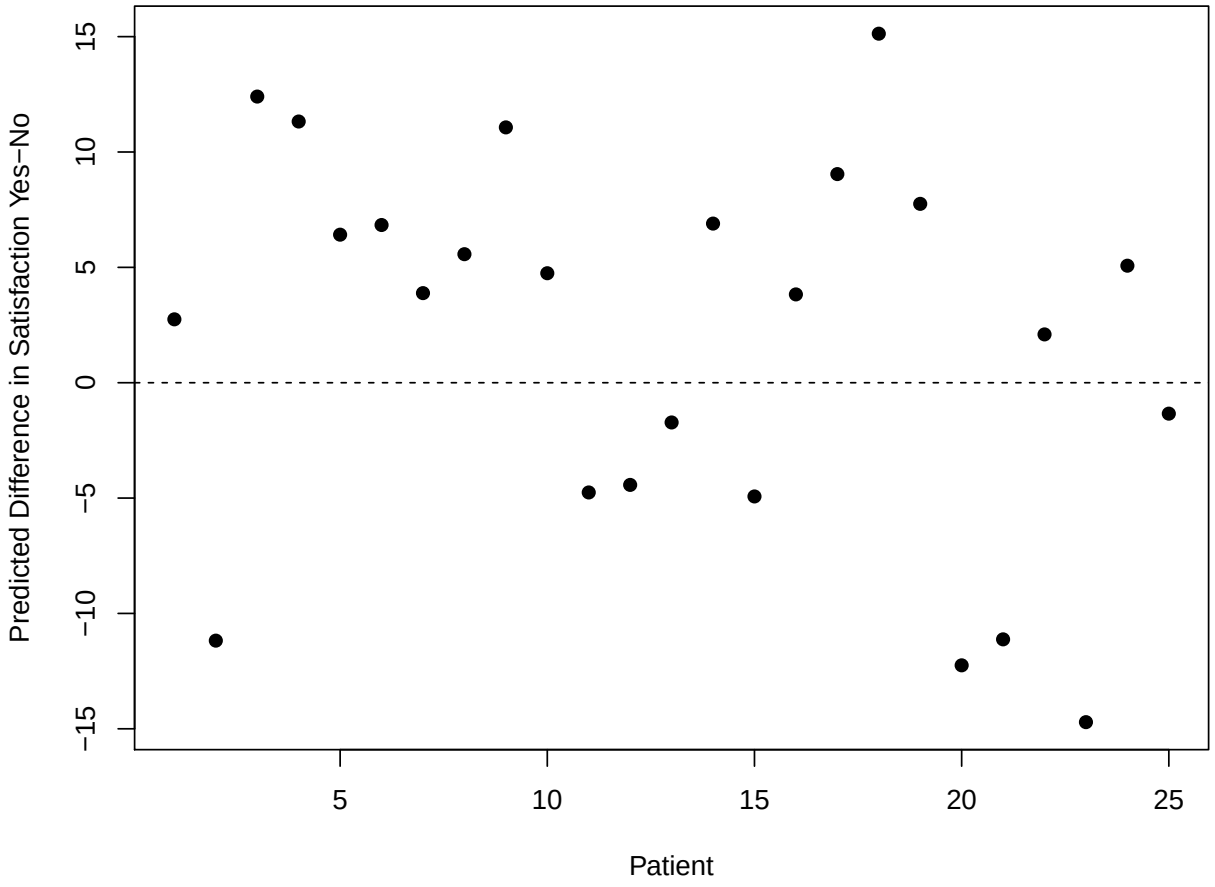
The AIC values inform is that the three models `fit2`, `fit5` and `step.fit2` have very similar qualities; recall that these three models are

- `fit2`: `Satisfaction = Age + Severity`
- `fit5`: `Satisfaction = Age + Severity + Age:Severity`
- `step.fit2`: `Satisfaction = Age + Severity + Surgery + Age:Surgery + Severity:Surgery`

| Model | $SS_{Res}$ | AIC | BIC | $R^2_{Adj}$ |
|---|---|---|---|---|
| `fit2` | 2062.286 | 189.264 | 194.14 | 0.792 |
| `fit5` | 2032.737 | 190.903 | 196.998 | 0.786 |
| `step.fit2` | 1694.31 | 190.351 | 198.883 | 0.802 |

These three models could be equally well supported by the data, and only the last depends on the `Surgery` factor. Therefore, the conclusion is not at all clear. To attempt to quantify the effect of surgery, we look at fitted values for the last models when `Surgery` is set to `No` for all patients, and then to `Yes` for all patients, and then look at the difference in Satisfaction Score.

```
No.data<-Q3data;No.data$Surgery<-as.factor('No')
Yes.data<-Q3data;Yes.data$Surgery<-as.factor('Yes')
No.fit<-predict(step.fit2,newdata=No.data)
Yes.fit<-predict(step.fit2,newdata=Yes.data)
par(mar=c(4,4,1,1))
plot(Yes.fit-No.fit,pch=19,xlab='Patient',ylab='Predicted Difference in Satisfaction Yes-
abline(h=0,lty=2)
```




We might therefore conclude that surgery has a small effect to improve patient satisfaction overall (most of these differences between `Surgery==Yes` and `Surgery==No` are positive), but it is not clear that it is statistically significant. 6 Marks

In each case we can extract the design matrix using `model.matrix()` in R:

(a) **the main effect model in Q1;**

```
Xa<-model.matrix(fitQ1)[,]
(XatXa<-data.matrix(t(Xa) %*% Xa))
```

```
:             (Intercept) Faculty2 Faculty3
: (Intercept)          45       15       15
: Faculty2             15       15        0
: Faculty3             15        0       15
```

(b) **the main effects only model in Q2;**

```
Xb<-model.matrix(mod4)[,]
colnames(Xb)<-c('Int','car=medium','car=small','type=Octel')
(XbtXb<-data.matrix(t(Xb) %*% Xb))
```

```
:             Int car=medium car=small type=Octel
: Int          36         12        12         18
: car=medium   12         12         0          6
: car=small    12          0        12          6
: type=Octel   18          6         6         18
```

(c) **the main effects plus interaction model in Q2**

```
Xc<-model.matrix(mod5)[,]
colnames(Xc)<-c('Int','car=medium','car=small','type=Octel','med:Oct.','small:Oct.')
(XctXc<-data.matrix(t(Xc) %*% Xc))
```

```
:             Int car=medium car=small type=Octel med:Oct. small:Oct.
: Int          36         12        12         18        6          6
: car=medium   12         12         0          6        6          0
: car=small    12          0        12          6        0          6
: type=Octel   18          6         6         18        6          6
: med:Oct.      6          6         0          6        6          0
: small:Oct.    6          0         6          6        0          6
```

These matrices are not diagonal, and therefore the dummy predictors given by the indicator functions on page 1 are not orthogonal. 3 Marks

The simplest way to obtain an orthogonal parameterization in (a) is to reparameterize to use the group means, as in Q1 (b), by removing the intercept: 2 Marks

```
fitQ2<-lm(Score~-1+Faculty,data=Q1data)
Xd<-model.matrix(fitQ2)[,]
(XdtXd<-data.matrix(t(Xd) %*% Xd))
```

```
:          Faculty1 Faculty2 Faculty3
: Faculty1       15        0        0
: Faculty2        0       15        0
: Faculty3        0        0       15
```

However, note that does NOT work for multi-factor models. For example,

```
mod4.2<-lm(noise~-1+carsize+type,data=Q2data)
Xe<-model.matrix(mod4.2)[,]
colnames(Xe)<-c('car=large','car=medium','car=small','type=Octel')
(XetXe<-data.matrix(t(Xe) %*% Xe))

:            car=large car=medium car=small type=Octel
: car=large         12          0         0          6
: car=medium         0         12         0          6
: car=small          0          0        12          6
: type=Octel         6          6         6         18
```

In order to obtain a fully orthogonal general representation, *polynomial* contrasts can be used. The `contr.poly` function defines a new parameterization where

```
Q2data.poly<-Q2data
contrasts(Q2data.poly$carsize)<-contr.poly(3)
contrasts(Q2data.poly$type)<-contr.poly(2)
mod4.3<-lm(noise~carsize+type,data=Q2data.poly)
Xf<-model.matrix(mod4.3)[,]
(XftXf<-round(data.matrix(t(Xf) %*% Xf),6))

:             (Intercept) carsize.L carsize.Q type.L
: (Intercept)          36         0         0      0
: carsize.L             0        12         0      0
: carsize.Q             0         0        12      0
: type.L                0         0         0     18

mod5.1<-lm(noise~carsize*type,data=Q2data.poly)
Xg<-model.matrix(mod5.1)[,]
colnames(Xg)<-c('Int','c.L','c.Q','t.L','c.L:t.L','c.Q:t.L')
(XgtXg<-round(data.matrix(t(Xg) %*% Xg),6))

:         Int c.L c.Q t.L c.L:t.L c.Q:t.L
: Int      36   0   0   0       0       0
: c.L       0  12   0   0       0       0
: c.Q       0   0  12   0       0       0
: t.L       0   0   0  18       0       0
: c.L:t.L   0   0   0   0       6       0
: c.Q:t.L   0   0   0   0       0       6
```

The `contr.poly` parameterization uses a representation of the predictor columns based on orthogonal polynomials, and it is possible to use the resulting contrasts especially if it is believed that the factor levels are equally spaced on an underlying continuum; the idea is that the factor levels on an underlying ordinal scale can be represented by real values, and then for a factor with $M$ levels, one may represent the effects of the levels by using polynomials of increasing order $1, 2, \ldots, M-1$, and hence look for polynomial patterns in the data. See for example

`http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm`

Alternatively, a direct reparameterization into an orthogonal matrix can be obtained by decomposing any $\mathbf{X}$ matrix: for example, using standard Gram-Schmidt or QR decomposition.

Note, however, that such a reparameterization typically changes the interpretation of the estimates, and will change the result of the inference, but not the ANOVA tests; for example

```
round(summary(mod5)$coef,3)

:                                Estimate Std. Error t value
: (Intercept)                     775.000      3.302 234.711
: carsizemedium car                70.833      4.670  15.169
: carsizesmall car                 50.833      4.670  10.886
```

```
: typeOctel filter                       -5.000      4.670  -1.071
: carsizemedium car:typeOctel filter  -19.167      6.604  -2.902
: carsizesmall car:typeOctel filter     1.667      6.604   0.252
:                                    Pr(>|t|)
: (Intercept)                           0.000
: carsizemedium car                     0.000
: carsizesmall car                      0.000
: typeOctel filter                      0.293
: carsizemedium car:typeOctel filter    0.007
: carsizesmall car:typeOctel filter     0.802
```

```
round(summary(mod5.1)$coef,3)
```

```
:                  Estimate Std. Error t value Pr(>|t|)
: (Intercept)       810.139      1.348 600.989    0.000
: carsize.L          36.534      2.335  15.647    0.000
: carsize.Q         -28.918      2.335 -12.385    0.000
: type.L             -7.660      1.906  -4.018    0.000
: carsize.L:type.L    0.833      3.302   0.252    0.802
: carsize.Q:type.L   11.547      3.302   3.497    0.001
```

```
anova(mod5,test='F')
```

```
: Analysis of Variance Table
:
: Response: noise
:              Df  Sum Sq Mean Sq  F value    Pr(>F)
: carsize       2 26051.4 13025.7 199.1189 < 2.2e-16 ***
: type          1  1056.2  1056.2  16.1465 0.0003631 ***
: carsize:type  2   804.2   402.1   6.1465 0.0057915 **
: Residuals    30  1962.5    65.4
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod5.1,test='F')
```

```
: Analysis of Variance Table
:
: Response: noise
:              Df  Sum Sq Mean Sq  F value    Pr(>F)
: carsize       2 26051.4 13025.7 199.1189 < 2.2e-16 ***
: type          1  1056.2  1056.2  16.1465 0.0003631 ***
: carsize:type  2   804.2   402.1   6.1465 0.0057915 **
: Residuals    30  1962.5    65.4
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```