DEPARTMENT OF STATISTICS

STATS 760 A Survey of Modern Applied Statistics

Assignment 1 2017 Model answers

Question 1. I will illustrate the process of building up the X-matrix using an example with factors A, B having 2 and 3 levels respectively and continuous variables X and Z. I will use the data in the data frame **`data`** constructed as follows:

```
> factors = expand.grid(A=factor(1:2), B=factor(1:3))
> data = data.frame(y=rnorm(6), factors, X=rnorm(6), Z=rnorm(6))
> data
            y A B          X            Z
1  0.61232562 1 1 -0.3232458 -0.084920449
2  1.68522307 2 1  1.4453469 -0.194883903
3 -0.31776920 1 2  0.9118442 -0.217020282
4 -0.07382695 2 2 -1.3044628  0.784233533
5  0.68356993 1 3 -0.3077209 -0.005710716
6  1.43643789 2 3  1.1927881 -0.520543495
```

The model has 6 planes, one for each factor level combination. There is not enough data to fit the model (we need at least 3 observations per plane) but this is big enough to illustrate the process of building up the X-matrix.

Now we start building up the model-matrix using the formula `y~A + B*X + B*Z` which expands to `y ~ A + B + X + Z + B:X + B:Z`. We start with the treatment contrasts.

```
> # first a column of ones
> myModelMatrix = as.matrix(rep(1,6))
> colnames(myModelMatrix)="(Int)"
> # add columns for factor A
> Xa = cbind(data$A=="1", data$A=="2")*1
> Ca = contr.treatment(2)
> XaCa = Xa%*%Ca
> colnames(XaCa) = "A2"
> myModelMatrix = cbind(myModelMatrix, XaCa)
> myModelMatrix
     (Int) A2
[1,]     1  0
[2,]     1  1
[3,]     1  0
[4,]     1  1
[5,]     1  0
[6,]     1  1
```

```
> # add columns for factor B
> Xb = cbind(data$B=="1", data$B=="2",data$B=="3")*1
> Cb = contr.treatment(3)
> XbCb = Xb%*%Cb
> colnames(XbCb) = c("B2","B3")
>
> myModelMatrix = cbind(myModelMatrix,XbCb)
> myModelMatrix
     (Int) A2 B2 B3
[1,]     1  0  0  0
[2,]     1  1  0  0
[3,]     1  0  1  0
[4,]     1  1  1  0
[5,]     1  0  0  1
[6,]     1  1  0  1
>
> # add column for X
> myModelMatrix = cbind(myModelMatrix,data$X)
> colnames(myModelMatrix)[5] = "X"
> myModelMatrix
     (Int) A2 B2 B3          X
[1,]     1  0  0  0 -0.3232458
[2,]     1  1  0  0  1.4453469
[3,]     1  0  1  0  0.9118442
[4,]     1  1  1  0 -1.3044628
[5,]     1  0  0  1 -0.3077209
[6,]     1  1  0  1  1.1927881
>
> # repeat for Z
>
> myModelMatrix = cbind(myModelMatrix,data$Z)
> colnames(myModelMatrix)[6] = "Z"
> myModelMatrix
     (Int) A2 B2 B3          X            Z
[1,]     1  0  0  0 -0.3232458 -0.084920449
[2,]     1  1  0  0  1.4453469 -0.194883903
[3,]     1  0  1  0  0.9118442 -0.217020282
[4,]     1  1  1  0 -1.3044628  0.784233533
[5,]     1  0  0  1 -0.3077209 -0.005710716
[6,]     1  1  0  1  1.1927881 -0.520543495
>
>
> # Add columns for B:X interactions
> myModelMatrix = cbind(myModelMatrix,XbCb* data$X)
> colnames(myModelMatrix)[7:8] = c("B2:X", "B3:X")
> myModelMatrix
     (Int) A2 B2 B3          X            Z       B2:X       B3:X
[1,]     1  0  0  0 -0.3232458 -0.084920449  0.0000000  0.0000000
[2,]     1  1  0  0  1.4453469 -0.194883903  0.0000000  0.0000000
```

```
[3,]     1  0  1  0  0.9118442 -0.217020282  0.9118442  0.0000000
[4,]     1  1  1  0 -1.3044628  0.784233533 -1.3044628  0.0000000
[5,]     1  0  0  1 -0.3077209 -0.005710716  0.0000000 -0.3077209
[6,]     1  1  0  1  1.1927881 -0.520543495  0.0000000  1.1927881
>
>
> # Add columns for B:Z interactions
> myModelMatrix = cbind(myModelMatrix,XbCb*data$Z)
> colnames(myModelMatrix)[9:10] = c("B2:Z", "B3:Z")
> myModelMatrix
     (Int) A2 B2 B3         X           Z       B2:X       B3:X       B2:Z         B3:Z
[1,]     1  0  0  0 -0.3232458 -0.084920449  0.0000000  0.0000000  0.0000000  0.000000000
[2,]     1  1  0  0  1.4453469 -0.194883903  0.0000000  0.0000000  0.0000000  0.000000000
[3,]     1  0  1  0  0.9118442 -0.217020282  0.9118442  0.0000000 -0.2170203  0.000000000
[4,]     1  1  1  0 -1.3044628  0.784233533 -1.3044628  0.0000000  0.7842335  0.000000000
[5,]     1  0  0  1 -0.3077209 -0.005710716  0.0000000 -0.3077209  0.0000000 -0.005710716
[6,]     1  1  0  1  1.1927881 -0.520543495  0.0000000  1.1927881  0.0000000 -0.520543495
>
```

We see that the beta coefficient (Int) is the y-intercept of the line corresponding to A=1 and B=1 (the 1,1 line.)

The beta coefficient A2 is the difference between the 2,1 line intercept and the 1,1 line intercept. (That is, with B at its baseline.) The interpretation of B2 and B3 is similar, except with A at its baseline.

The coefficient “X” is the x-slope of the baseline line (with A=1, B=1). The coefficient B2:X is the difference between the slope of lines with B=2 and the slope of the lines with B=1. Note the slopes do not depend on A, just B. The B3:X coefficient is the difference between the slope of lines with B=3 and the slope of the lines with B=1.

For the sum contrasts, we repeat the code above, changing contr.treatment to contr.sum. We will need to change the column labelling as well. The result is

```
     (Int) A1 B1 B2         X           Z       B1:X       B2:X         B1:Z         B2:Z
[1,]     1  1  1  0 -0.3232458 -0.084920449 -0.3232458  0.0000000 -0.084920449  0.000000000
[2,]     1 -1  1  0  1.4453469 -0.194883903  1.4453469  0.0000000 -0.194883903  0.000000000
[3,]     1  1  0  1  0.9118442 -0.217020282  0.0000000  0.9118442  0.000000000 -0.217020282
[4,]     1 -1  0  1 -1.3044628  0.784233533  0.0000000 -1.3044628  0.000000000  0.784233533
[5,]     1  1 -1 -1 -0.3077209 -0.005710716  0.3077209  0.3077209  0.005710716  0.005710716
[6,]     1 -1 -1 -1  1.1927881 -0.520543495 -1.1927881 -1.1927881  0.520543495  0.520543495
```

Thus, for example, the y-intercept of the 1,1 line (I11 say) is (Int)+ A1+B1. If we work out all six y-intercepts and average them, we see that (Int) is the overall average of the six y-intercepts.

The average of the 3 intercepts with A=1, minus the average intercept is

(I11+I12+ I13)/3 – (Int) = (I11-I21+I12-I22+I13-I23)/6

which reduces down to A1. B1 and B2 are interpreted similarly.

The average of all six X-slopes is

```
(s11+ s21 + s12+ s22 s13+ s23)/6
=[(X + B1:X) + (X+B1:X) +  (X+B2:X)  + (X+B2:X)  + (X-B1:X–B2:X) + (X-B1:X–B2:X)]/6
 = X
```

so the regression coefficient X is the average of the X-slopes.

The average of the 2 X-slopes, with B=1, (s11 and s21) minus the overall average slope is

(s11+ s21)/2 – (s11+ s21 + s12+ s22 s13+ s23)/6
= [(X + B1:X) + (X + B1:X)]/2 – X
=B1:X
The Z- coefficients are interpreted similarly.

Marks: 10 marks for the matrices, 5 for each contrast. If you interpreted the coefficients correctly you got an extra 5 marks for each type of contrast so 20 marks in all.

Comment on your answers: With a few happy exceptions, most of you are still struggling somewhat with interpretations, although the parts relating to formation of the model matrix were well done. Part of the problem is that some of you applied incorrect rounding to the matrix relating the betas to the y-intercepts.

Question 2.

Any three non-parallel vectors whose first two elements are the same, whose 3rd and 4th elements are the same, and whose last 2 elements are the same will do. E.G

```
1          1  0  0
2          1  0  0
3          0  1  0
4          0  1  0
5          0  0  1
6          0  0  1
```

Then the beta's are just the means.

Or you could use the matrices produced by the sum and treatment contrasts, which have the usual interpretation:

```
> A=factor(rep(1:3,c(2,2,2)))
> # sum contrasts
> options(contrasts=c("contr.sum","contr.poly"))
> XS = model.matrix(~A)
> XS
  (Intercept) A1 A2
1           1  1  0
2           1  1  0
3           1  0  1
4           1  0  1
5           1 -1 -1
6           1 -1 -1


> #  treatment contrasts
> options(contrasts=c("contr.treatment","contr.poly"))
> XT = model.matrix(~A)
> XT
  (Intercept) A2 A3
1           1  0  0
2           1  0  0
3           1  1  0
4           1  1  0
5           1  0  1
6           1  0  1
```

Marks: 5 marks for writing down the X-matrices, 5 for interpreting the coefficients.

Question 3.

We have $X_T \beta_T = X_S \beta_S$, so $\beta_T = (X_T'X_T)^{-1} X_T'X_S\beta_S$. Thus T = $(X_T'X_T)^{-1} X_T'X_S$.

```
> round(solve(t(XT)%*%XT)%*% t(XT)%*%XS)
            (Intercept) A1 A2
(Intercept)           1  1  0
A2                    0 -1  1
A3                    0 -2 -1
```

5 marks for the formula, 5 marks for the actual numerical matrix.

Total for assignment: 40 marks