MATH 423/533 - ASSIGNMENT 4

To be handed in not later than 11:59pm, 9th December 2017.
Please submit your solutions as pdf via myCourses. Include relevant `R` code.

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

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

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

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

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.

QUESTIONS

All data sets can be found at

`http://www.math.mcgill.ca/yyang/regression/data/XXXXXX.csv`

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. 15 Marks

   (b) Report the estimated mean scores, with associated standard errors, for students from each of the three faculties. 15 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. 25 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
```

   15 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. 30 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.

25 Marks