Exercises, STK4040, week 42

October 11, 2007

Exercise 1

Consider the linear model $\boldsymbol{y} = \mathrm{X}\beta + \varepsilon$, where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathrm{I}_n)$. We assume that the $n \times (p+1)$ matrix X has full rank $p + 1 < n$, and that the first coloumn of X is $\mathbf{1}_n$.

The least squares estimate of β is $\hat{\beta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\boldsymbol{y}$. The predicted response values (A.K.A. fitted values) are $\hat{\boldsymbol{y}} = \mathrm{X}\hat{\beta}$. Let $\mathrm{H} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$. Then the predicted response values can be written as $\hat{\boldsymbol{y}} = \mathrm{H}\boldsymbol{y}$. (H is often called the 'hat matrix', because it transforms $\boldsymbol{y}$ into $\hat{\boldsymbol{y}}$.) The residuals are given by $\hat{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathrm{I} - \mathrm{H})\boldsymbol{y}$. Define $\mathrm{M} = \mathrm{I} - \mathrm{H}$. Then $\hat{\varepsilon} = \mathrm{M}\boldsymbol{y}$.

Prove the following:

1. H is idempotent (i.e., $\mathrm{HH} = \mathrm{H}$) and symmetric.
2. M is idempotent and symmetric.
3. $\mathrm{MX} = \mathbf{0}$, and $\mathrm{M}\mathbf{1}_n = \mathbf{0}$.
4. $\hat{\varepsilon} = \mathrm{M}\varepsilon$, $\mathrm{X}^T\hat{\varepsilon} = \mathbf{0}$, and $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.
5. $\hat{\boldsymbol{y}}^T\hat{\varepsilon} = 0$.
6. $\mathbf{1}_n^T\hat{\varepsilon} = 0$.

Exercise 2

Given the linear model $\boldsymbol{y} = \mathrm{X}_1\beta + \mathrm{X}_2\varphi + \varepsilon$, where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathrm{I}_n)$, and $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2)$ has full rank.

Show that $\hat{\beta}$ can be calculated by first regressing $\boldsymbol{y}$ and X_1 onto X_2, and then regress the residuals of $\boldsymbol{y}$ onto the residuals of X_1.

Exercise 3

The file `dataset_42.txt` contains sensory data from an experiment in which 105 persons tasted 6 different cheeses. Each person has given each cheese an integer score between 1 and 9, denoting how well they liked the cheese (9 is best and 1 is worst).

The file contains a variable `Code`, giving the sex of the person (`A` for female and `B` for male), and one variable for each cheese: `A_Cow_Full_fat`, `C_Cow_Full_fat`, `D_Cow_Low_fat`, `E_Cow_Low_fat`, `G_Buffalo_Full_fat`, and `I_Buffalo_Full_fat`. The cheeses A, C, D, and E are made from cow milk, while G and I are made from buffalo milk. The cheeses A, C, G, and I have regular fat content, while D and E are low fat cheeses.

Do a principal component analysis of the liking scores. How many components are important? Plot loadings and scores, and see if you can find any patterns. Tip: the data can be read in and set up in `R` like this:

```
tmp <- read.table("dataset_42.txt")
names(tmp)
Code <- tmp$Code
X <- as.matrix(tmp[,2:7])
```

Plot the scores with codes and/or colours to denote sex. Do you see a pattern now? Interpret the loadings. (Tip: You can plot the codes by using the argument `pch = as.character(Code))` in the plot function, and colours can be added with the argument `col = as.numeric(Code)`.)