

## ST503: Homework 02

Arnab Maity

### Problem 1 (40 pts)

Consider the linear model where the response vector is  $y = [y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}]^T$ , the parameter vector is  $\beta = [\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3]^T$ , and the model matrix  $X$  is as follows:

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    0    1    0    0
## [2,]    1    1    0    0    1    0
## [3,]    1    1    0    0    0    1
## [4,]    1    0    1    1    0    0
## [5,]    1    0    1    0    1    0
## [6,]    1    0    1    0    0    1
```

- (A) What is the rank of  $X$ ?
- (B) Write the normal equations. Explain why the normal equations have infinitely many solutions.
- (C) Show that  $\alpha_1 - \alpha_2$  is estimable. Don't use any software.
- (D) Show that  $\beta_1 - 2\beta_2 + \beta_3$  is estimable. Don't use any software.
- (E) Use R to check your answers in part (C) and (D).

### Problem 2 (40 pts)

The dataset `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors.

- (a) Present the output. What percentage of variation in the response is explained by these predictors?
- (b) Which observation has the largest (positive) residual? Give the case number.
- (c) Compute the mean and median of the residuals.
- (d) Compute the correlation of the residuals with the fitted values.
- (e) Compute the correlation of the residuals with the income.
- (f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

### Problem 3 (20 pts)

The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988. Fit a model with weekly wages as the response and years of education and experience as predictors. Report and give a simple interpretation to the regression coefficient for years of education. Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education. Which interpretation is more natural?