MAST30025: Linear Statistical Models

Solutions to Week 9 Lab

1. Recall Question 5 from the Week 8 lab. In a manufacturing plant, filters are used to remove pollutants. We are interested in comparing the lifespan of 5 different types of filters. Six filters of each type are tested, and the time to failure in hours is given in the dataset `filters` (on the website, in `csv` format).

   (a) Is $\mu - \tau_1 + \tau_5$ estimable?
   Solution:

   ```
   > tt <- c(1,-1,0,0,0,1)
   > t(tt) %*% XtXc %*% t(X) %*% X
   ```

   ```
                 [,1] [,2] [,3] [,4] [,5] [,6]
   [1,] -5.551115e-17   -1    0    0    0    1
   ```

   No, it is not estimable.

   (b) Is $\tau_1 - \frac{1}{2}\tau_3 - \frac{1}{2}\tau_4$ estimable?
   Solution: Yes, as it is a treatment contrast.

   (c) In the week 8 lab you were asked to find two solutions to the normal equations. Verify that they produce the same estimate of $\tau_4 - \tau_5$.
   Solution:

   ```
   > b[5] - b[6]
   [1] -4
   > b2[5] - b2[6]
   [1] -4
   ```

   (d) Do your two solutions produce the same estimate of $2\mu + \tau_1$?
   Solution:

   ```
   > 2*b[1] + b[2]
   [1] 249.1667
   > 2*b2[1] + b2[2]
   [1] 469.3889
   ```

   No.

   (e) Write down the quantities corresponding to: (i) the lifespan of type 1 filters; (ii) the difference between the lifespans of type 2 and type 3 filters; (iii) the amount by which type 4 filters outlive the average filter; (iv) the expected total time to failure of a set of filters containing one of each type.
   Verify directly that all of these quantities are estimable, and estimate them.
   Solution: (i) $\mu + \tau_1$; (ii) $\tau_2 - \tau_3$; (iii) $\tau_4 - \frac{1}{5}\sum_{i=1}^{5}\tau_i$; (iv) $5\mu + \sum_{i=1}^{5}\tau_i$.

   ```
   > tt <- c(1,1,0,0,0,0)
   > t(tt) %*% XtXc %*% t(X) %*% X
   ```

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

   ```
   > tt%*%b
   ```

   ```
            [,1]
   [1,] 249.1667
   ```

   ```
   > tt <- c(0,0,1,-1,0,0)
   > t(tt) %*% XtXc %*% t(X) %*% X
   ```

```
              [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 5.551115e-17    0    1   -1    0    0
```

```
> tt%*%b
```

```
     [,1]
[1,] 21.5
```

```
> tt <- c(0,-1/5,-1/5,-1/5,4/5,-1/5)
> t(tt) %*% XtXc %*% t(X) %*% X
```

```
               [,1] [,2] [,3] [,4] [,5] [,6]
[1,] -1.387779e-17 -0.2 -0.2 -0.2  0.8 -0.2
```

```
> tt%*%b
```

```
         [,1]
[1,] 93.06667
```

```
> tt <- c(5,1,1,1,1,1)
> t(tt) %*% XtXc %*% t(X) %*% X
```

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

```
> tt%*%b
```

```
         [,1]
[1,] 1321.333
```

(f) Fit a `lm` model using `contr.treatment` contrasts (the default). This gives estimates of $\mu_1, \mu_2 - \mu_1, \ldots, \mu_5 - \mu_1$. Use these to estimate $\bar{\mu}, \mu_1 - \bar{\mu}, \ldots, \mu_5 - \bar{\mu}$. Check your answers by fitting a `contr.sum` model.

Solution:

```
> filters$type <- factor(filters$type)
> model <- lm(life ~ type, data = filters)
> mu <- model$coefficients + c(0, 1, 1, 1, 1)*model$coefficients[1]
> names(mu)[1] <- "type1"
> (mubar <- mean(mu))
```

```
[1] 264.2667
```

```
> mu - mubar
```

```
    type1     type2     type3     type4     type5
-15.10000 -76.76667 -98.26667  93.06667  97.06667
```

```
> contrasts(filters$type) <- contr.sum(5)
> model2 <- lm(life ~ type, data = filters)
> summary(model2)
```

```
Call:
lm(formula = life ~ type, data = filters)

Residuals:
     Min       1Q   Median       3Q      Max 
-226.333  -61.458   -2.833   39.625  250.667 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   264.27      22.59  11.700 1.23e-11 ***
type1         -15.10      45.17  -0.334   0.7410    
type2         -76.77      45.17  -1.699   0.1017    
type3         -98.27      45.17  -2.175   0.0393 *  
type4          93.07      45.17   2.060   0.0499 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 123.7 on 25 degrees of freedom
Multiple R-squared:  0.3468,	Adjusted R-squared:  0.2423
F-statistic: 3.319 on 4 and 25 DF,  p-value: 0.026
> -sum(model2$coefficients[-1]) # mu_5 - bar(mu)
[1] 97.06667
```

2. According to the Gauss-Markov theorem, the estimator for $\mathbf{t}^T\boldsymbol{\beta}$ with the lowest variance is $\mathbf{t}^T\mathbf{b}$. Assuming that $\mathbf{t}^T\boldsymbol{\beta}$ is estimable, show that this variance is $\sigma^2\mathbf{t}^T(X^TX)^c\mathbf{t}$.

   Solution:

$$\begin{aligned}
\text{var } \mathbf{t}^T\mathbf{b} &= \text{var } \mathbf{t}^T(X^TX)^cX^T\mathbf{y} \\
&= \mathbf{t}^T(X^TX)^cX^T\sigma^2 IX[(X^TX)^c]^T\mathbf{t} \\
&= \sigma^2\mathbf{t}^T(X^TX)^cX^TX(X^TX)^c\mathbf{t} \\
&= \sigma^2\mathbf{t}^T(X^TX)^c\mathbf{t}.
\end{aligned}$$

3. For the one-way classification model, with n_i observations in group i, show that

$$SS_{Reg} := \hat{\mathbf{y}}^T\hat{\mathbf{y}} = \mathbf{y}^TX(X^TX)^cX^T\mathbf{y} = \sum_{i=1}^{k}(\bar{y}_i)^2 n_i.$$

   Solution:

$$\begin{aligned}
SS_{Reg} &= \mathbf{y}^TX(X^TX)^cX^T\mathbf{y} \\
&= (X^T\mathbf{y})^T\mathbf{b} \\
&= \begin{bmatrix} \sum_{ij} y_{ij} & \sum_j y_{1j} & \sum_j y_{2j} & \dots & \sum_j y_{kj} \end{bmatrix} \begin{bmatrix} 0 \\ \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_k \end{bmatrix} \\
&= \sum_i \left(\sum_j y_{ij}\bar{y}_i \right) \\
&= \sum_i (\bar{y}_i)^2 n_i.
\end{aligned}$$

4. Consider the one-way classification model with 3 levels ($k = 3$). Find all estimable quantities of the form $\sum_{i=1}^{3} a_i\tau_i$.

   Solution: All estimable quantities are of the form $\mathbf{t}^T\boldsymbol{\beta}$, where

$$\begin{aligned}
\mathbf{t}^T &= \mathbf{t}^T(X^TX)^cX^TX \\
&= \mathbf{t}^T \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{n_1} & 0 & 0 \\ 0 & 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & 0 & \frac{1}{n_3} \end{bmatrix} \begin{bmatrix} n & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{bmatrix} \\
&= \mathbf{t}^T \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \\
\mathbf{t}^T \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} &= \mathbf{0}.
\end{aligned}$$

   Since we have $\mathbf{t}^T = \begin{bmatrix} 0 & a_1 & a_2 & a_3 \end{bmatrix}$, this gives $a_1 + a_2 + a_3 = 0$. In other words, only treatment contrasts are estimable.

5. Consider the two-way classification model

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}.$$

Suppose that you have at least one sample from each combination of factor levels.

Treatment contrasts for the first factor are defined here as $\sum_i a_i \tau_i$, where $\sum_i a_i = 0$. Show that these treatment contrasts are estimable.

Solution: We can write them as

$$\sum_i a_i \tau_i = \sum_i a_i(\mu + \tau_i + \beta_1),$$

and we know that $\mu + \tau_i + \beta_1$ is estimable as it is an element of $X\boldsymbol{\beta}$. Therefore these treatment contrasts are estimable. (And likewise, treatment contrasts in the second factor are also estimable.)