**STATS 3001 / STATS 4104 / STATS 7054**
**Statistical Modelling III**
**Tutorial 2**
**2021**
**Solutions**

**QUESTIONS:**

1. Consider the multiple regression model,

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \text{ where } \boldsymbol{\mathcal{E}} \sim N_n(\mathbf{0}, \sigma^2 I)$$

and suppose an additional independent observation

$$Y_0 \sim N(\boldsymbol{x}_0^T\boldsymbol{\beta}, \sigma^2)$$

is to be made.

(a) Find the mean and variance of $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$.
(b) Find the mean and variance of $Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$.
(c) Hence find the distribution of

$$\frac{Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}}{\sigma\sqrt{1 + \boldsymbol{x_0}^T(X^TX)^{-1}\boldsymbol{x}_0}}.$$

(d) Hence what is the distribution of

$$\frac{Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}}{s_e\sqrt{1 + \boldsymbol{x_0}^T(X^TX)^{-1}\boldsymbol{x}_0}}$$

---

**SOLUTIONS:**

**a**

$$E(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T E(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0\boldsymbol{\beta}$$

and

$$\text{var}(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^T\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0 = \sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0.$$

**b**

$$E(Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}) = E(Y_0) - E(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0\boldsymbol{\beta} - \boldsymbol{x}_0\boldsymbol{\beta} = 0$$

and, since $Y_0$ and $\hat{\boldsymbol{\beta}}$ are independent,

$$\begin{aligned}\text{var}(Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}) &= \text{var}(Y_0) + \text{var}(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}) \\ &= \sigma^2 + \sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0 \\ &= \sigma^2(1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0).\end{aligned}$$

**c**

$$\frac{Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}}{\sigma\sqrt{1 + \boldsymbol{x_0}^T(X^TX)^{-1}\boldsymbol{x}_0}}. \sim N(0,1).$$

**d**

Recall that the $t_k$ distribution arises as the distribution of

$$T = \frac{Z}{\sqrt{V/k}}$$

where

$$Z \sim N(0,1)$$

and

$$V \sim \chi^2_k$$

independently.

Hence the distribution is

$$t_{n-p}$$

---

2. In this question, we are going to look at how to move from one parameterization to another.

   Consider an experiment with a categorical predictor with four levels.

   The models we will consider are

$$
\begin{aligned}
&M_1 : \eta_{ij} = \mu_i, \quad i = 1, 2, 3, 4; \quad j = 1, \ldots, n_i. \\
&M_2 : \eta_{ij} = \mu_{ref} + \alpha_i, \quad \alpha_1 = 0; \quad i = 1, 2, 3, 4; \quad j = 1, \ldots, n_i. \\
&M_3 : \eta_{ij} = \mu_{sum} + \beta_i, \quad \sum_{i=1}^{4} \beta_i = 0; \quad i = 1, 2, 3, 4; j = 1, \ldots, n_i.
\end{aligned}
$$

   Let

$$
\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix}; \boldsymbol{\alpha} = \begin{pmatrix} \mu_{ref} \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix}; \boldsymbol{\beta} = \begin{pmatrix} \mu_{sum} \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}
$$

   Note that $\beta_4$ is obtained by subtraction and therefore not included in the matrix formulation of the model.

   (a) Find the $4 \times 4$ matrices $C_1$ and $C_2$ such that

$$
\boldsymbol{\mu} = C_1 \boldsymbol{\alpha}
$$

   and

$$
\boldsymbol{\mu} = C_2 \boldsymbol{\beta}.
$$

   (b) By expressing the parameters

$$
\mu_{sum}, \beta_1, \beta_2, \beta_3
$$

   in terms of the group means:

$$
\mu_1, \mu_2, \mu_3, \mu_4,
$$

   obtain the inverse matrix $C_2^{-1}$.

   (c) Find the $4 \times 4$ matrix $A$ such that

$$
\boldsymbol{\beta} = A \boldsymbol{\alpha}
$$

(d) Studies conducted at the University of Melbourne indicate that there may be a difference between the pain thresholds of blonds and brunettes. Men and women of various ages were divided into four categories according to hair colour: light blond, dark blond, light brunette, and dark brunette. The purpose of the experiment was to determine whether hair colour is related to the amount of pain produced by common types of mishaps and assorted types of trauma. Each person in the experiment was given a pain threshold score based on his or her performance in a pain sensitivity test (the higher the score, the higher the person's pain tolerance). Shown below is the output from analysis in R.

```
pacman::p_load(tidyverse)
pain <- read_csv(here::here("data", "pain.csv"))
```

```
pain %>% count(HairColour)
```

```
## # A tibble: 4 x 2
##   HairColour        n
##   <chr>         <int>
## 1 DarkBlond         5
## 2 DarkBrunette      5
## 3 LightBlond        5
## 4 LightBrunette     4
```

```
lm1 <- lm(Pain ~ HairColour, data = pain)
broom::tidy(lm1)
```

```
## # A tibble: 4 x 5
##   term                    estimate std.error statistic  p.value
##   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                51.2       3.65     14.0  5.07e-10
## 2 HairColourDarkBrunette    -13.8       5.17     -2.67 1.75e- 2
## 3 HairColourLightBlond        8.00      5.17      1.55 1.43e- 1
## 4 HairColourLightBrunette    -8.7       5.48     -1.59 1.33e- 1
```

```
lm2 <- lm(Pain ~ HairColour, data = pain,
          contrasts = list(HairColour = "contr.sum"))
broom::tidy(lm2)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    47.6       1.88     25.3  1.05e-13
## 2 HairColour1     3.63      3.20      1.13 2.75e- 1
```

```
## 3 HairColour2   -10.2       3.20     -3.18 6.19e- 3
## 4 HairColour3    11.6       3.20      3.64 2.44e- 3
```

```
pain %>%
  group_by(HairColour) %>%
  summarise(mu = mean(Pain))
```

```
## # A tibble: 4 x 2
##   HairColour       mu
##   <chr>         <dbl>
## 1 DarkBlond      51.2
## 2 DarkBrunette   37.4
## 3 LightBlond     59.2
## 4 LightBrunette  42.5
```

Using the matrices from above show that the parameter estimates are equivalent.

---

**SOLUTIONS:**

**a**

$$C_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

**b**

Recall for the zero sum constraints that $\mu_{sum}$ is the average of the group means,

$$\mu_{sum} = \frac{1}{4}(\mu_1 + \mu_2 + \mu_3 + \mu_4)$$

and the individual effects are then

$$\beta_i = \mu_i - \mu_{sum}$$

for $i = 1, 2, 3$.

Note that $\beta_4$ is not required because it can obtained by subtraction.

Expressing this in matrix notation gives

$$C_2^{-1} = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \end{pmatrix}.$$

**c**

$$\begin{aligned} C_1\boldsymbol{\alpha} &= \boldsymbol{\mu} = C_2\boldsymbol{\beta} \\ \implies \boldsymbol{\beta} &= C_2^{-1}C_1\boldsymbol{\alpha} \end{aligned}$$

Hence

$$\begin{aligned} A &= C_2^{-1}C_1 \\ &= \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \\ &= \frac{1}{4}\begin{pmatrix} 4 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 \\ 0 & 3 & -1 & -1 \\ 0 & -1 & 3 & -1 \end{pmatrix}. \end{aligned}$$

**d**

Observe that

$$\frac{1}{4}\begin{pmatrix} 4 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 \\ 0 & 3 & -1 & -1 \\ 0 & -1 & 3 & -1 \end{pmatrix}\begin{pmatrix} 51.2 \\ -13.8 \\ 8.0 \\ -8.7 \end{pmatrix} = \begin{pmatrix} 47.575 \\ 3.625 \\ -10.175 \\ 11.625 \end{pmatrix}$$

as required.

$$\begin{aligned} \boldsymbol{\mu} &= C_1\boldsymbol{\alpha} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 51.2 \\ -13.8 \\ 8.0 \\ -8.7 \end{pmatrix} \\ &= \begin{pmatrix} 51.2 \\ 37.4 \\ 59.2 \\ 42.5 \end{pmatrix} \end{aligned}$$