HW7

Sam Olson

Q1

An experiment was conducted to compare the effectiveness of two sports drinks (denoted 1 and 2). The subjects included 60 males between the ages of 18 and 31. Each subject rode a stationary bicycle until his muscles were depleted of energy, rested for two hours, and biked again until exhaustion. During the rest period, each subject drank one of the two sports drinks as assigned by the researchers. Each subject's performance on the second round of biking following the rest period was assigned a score between 0 and 100 based on the energy expended prior to exhaustion. Higher scores were indicative of better performance.

20 of the 60 subjects repeated the bike-rest-bike trial on a second occasion separated from the first by approximately three weeks. These subjects drank one sports drink during the first trial and the other during the second trial. The drink order was randomized for each subject by the researchers, even though previous research suggested no performance difference in repeated trials when three weeks passed between trials. The other 40 subjects performed the trial only a single time, drinking a randomly assigned sports drink during the rest period. 20 of these subjects received sports drink 1, and the other 20 received sports drink 2. A portion of the entire data set is provided in the following table.

```
Subject  Drink 1  Drink 2
1        45       52
2        69       73
...      ...      ...
20       29       46
21       35       -
22       81       -
...      ...      ...
40       55       -
41       -        17
42       -        54
...      ...      ...
60       -        61
```

Subjects 1 through 20 in the table above represent the 20 subjects who performed the trial separately for each of the sports drinks. Note that the data set contains no information about which drink was received in the first trial and which drink was received in the second trial. Throughout the remainder of this problem, please assume that this information is not important. In other words, you may assume that the subjects would have scored the same for drinks 1 and 2 regardless of the order the trials were performed.

Suppose the following model is appropriate for the data.

$$y_{ij} = \mu_i + u_j + e_{ij}, \tag{1}$$

where y_{ij} is the score for drink i and subject j, μ_i is the unknown mean score for drink i, u_j is a random effect corresponding to subject j, and e_{ij} is a random error corresponding to the score for drink i and subject

j ($i = 1, 2$ and $j = 1, \ldots, 60$). Here $u_1, \ldots, u_{60}$ are assumed to be independent and identically distributed as $\mathcal{N}(0, \sigma_u^2)$ and independent of the e_{ij}'s, which are assumed to be independent and identically distributed as $\mathcal{N}(0, \sigma_e^2)$.

a)

For each of the subjects who received both drinks, the difference between the scores (drink 1 score $-$ drink 2 score) was computed. This yielded 20 score differences denoted $d_1, \ldots, d_{20}$. Describe the distribution of these differences considering the assumptions about the distribution of the original scores in model (1).

Answer

b)

Suppose you were given only the differences $d_1, \ldots, d_{20}$ from part (a). Provide a formula for a test statistic (as a function of $d_1, \ldots, d_{20}$) that could be used to test $H_0 : \mu_1 = \mu_2$.

Answer

c)

Fully state the exact distribution of the test statistic provided in part (b).

Answer

d)

Let $a_1, \ldots, a_{20}$ be the scores of the subjects who received only drink 1. Let $b_1, \ldots, b_{20}$ be the scores of the subjects who received only drink 2. Suppose you were given only these 40 scores. Provide a formula for a 95% confidence interval for $\mu_1 - \mu_2$ (as a function of $a_1, \ldots, a_{20}$ and $b_1, \ldots, b_{20}$).

Answer

e)

Suppose you were given $d_1, \ldots, d_{20}$ from part (a) and $a_1, \ldots, a_{20}$ and $b_1, \ldots, b_{20}$ from part (d). Provide formulas for unbiased estimators of σ_u^2 and σ_e^2 as a function of these observations.

Answer

f)

Suppose you were given $\bar{d}. = \sum_{i=1}^{20} d_i/20$, $\bar{a}. = \sum_{i=1}^{20} a_i/20$, and $\bar{b}. = \sum_{i=1}^{20} b_i/20$; where $d_1, \ldots, d_{20}$ are from part (a) and $a_1, \ldots, a_{20}$ and $b_1, \ldots, b_{20}$ are from part (d). Furthermore, suppose σ_e^2 and σ_u^2 are known. Provide a simplified expression for the estimator of $\mu_1 - \mu_2$ that you would use. Your answer should be a function of $\bar{d}., \bar{a}., \bar{b}., \sigma_u^2, \sigma_e^2$.

Answer

Q2

Suppose the responses in problem 2 were sorted first by subject and then by drink into a response vector y; i.e.,

$$y = [45, 52, 69, 73, \dots, 29, 46, 35, 81, \dots, 55, 17, 54, \dots, 61]^\top.$$

Provide X and Z matrices so that the model in equation (1) may be written as $y = X\beta + Zu + e$, where $\beta = [\mu_1, \mu_2]^\top$ and $u = [u_1, u_2, \dots, u_{60}]^\top$. If possible, use Kronecker product notation to simplify your answer.

Answer

As given, the response vector would be given in vector notation as:

$$y = [y_{11}, y_{21}, \cdots, y_{1,20}, y_{2,20}, y_{1,21}, \cdots, y_{1,40}, y_{2,41}, \cdots, y_{2,60}]'.$$

And modelled by:

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$,

The Kronecker product notation for $\mathbf{X}$ and $\mathbf{Z}$ are then given by:

$$X_{80\times 2} = \begin{bmatrix} \mathbf{1}_{20\times 1} \otimes I_{2\times 2} \\ I_{2\times 2} \otimes \mathbf{1}_{20\times 1} \end{bmatrix}$$

And

$$Z_{80\times 60} = \begin{bmatrix} I_{20\times 20} \otimes \mathbf{1}_{2\times 1} & 0_{40\times 40} \\ 0_{40\times 20} & I_{40\times 40} \end{bmatrix}$$

Q3

The following question refers to the slide set 12 titled The ANOVA Approach to the Analysis of Linear Mixed-Effects Models.

Derive the expected mean square for xu(trt) for the ANOVA table on slide 9 using the technique illustrated on slides 15 through 17.

Answer

$$E\left(MS_{xu(trt)}\right) = \frac{1}{df_{xu(trt)}} E(SS_{xu(trt)})$$

$$= \frac{1}{tn-t} E\left(\sum_{i=1}^{m}\sum_{j=1}^{t}(y_{ij.} - \bar{y}_{i..})^2\right)$$

$$= \frac{1}{tn-t} E\left(\sum_{i=1}^{m}\sum_{j=1}^{t}\left([\mu + \tau_i + u_{ij} + \bar{\epsilon}_{ij.}] - [\mu + \tau_i + \bar{u}_{i}. + \bar{\epsilon}_{i..}]\right)^2\right)$$

$$= \frac{m}{tn-t} \sum_{i=1}^{t}\sum_{j=1}^{n} E\left((u_{ij} - \bar{u}_{i.}) + (\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..})\right)^2$$

$$= \frac{m}{tn-t} \sum_{i=1}^{t}\sum_{j=1}^{n} \left\{E(u_{ij} - \bar{u}_{i.})^2 + E(\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..})^2\right\}$$

Note:

$$E\left\{(u_{ij} - \bar{u}_{i.}) + (\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..})\right\}^2 = \mathrm{Var}\left((u_{ij} - \bar{u}_{i.}) + (\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..})\right)$$

$$= \mathrm{Var}(u_{ij} - \bar{u}_{i.}) + \mathrm{Var}(\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..})$$

$$= E(u_{ij} - \bar{u}_{i.})^2 + E(\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..})^2$$

Since:

$$E(u_{ij} - \bar{u}_{i.}) = E(\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..}) = 0$$

Because we suppose:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_u^2\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \sigma_e^2\mathbf{I}\end{bmatrix}\right)$$

Continuing:

$$E\left(MS_{xu(trt)}\right) = \frac{m}{tn-t}\sum_{i=1}^{t}\left[E\left\{\sum_{j=1}^{n}(u_{ij}-\bar{u}_{i.})^2\right\} + E\left\{\sum_{j=1}^{n}(\bar{\epsilon}_{ij.}-\bar{\epsilon}_{i..})^2\right\}\right]$$

$$= \frac{m}{tn-t}\sum_{i=1}^{t}\left\{(n-1)\sigma_u^2 + (n-1)\frac{\sigma_e^2}{m}\right\}$$

And, since

$$u_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$$

And

$$\bar{\epsilon}_{ij.} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_e^2}{m}\right)$$

$$EMS_{xu(trt)} = \frac{m}{tn-t}\left\{t(n-1)\sigma_u^2 + t(n-1)\frac{\sigma_e^2}{m}\right\}$$

Giving us

$$E\left(MS_{xu(trt)}\right) = m\sigma_u^2 + \sigma_e^2$$

As given in Lecture Slides 12, around slide 18 (Slide 9 as mentioned provides df and Sums of Squares).

Q4

The following question refers to the slide 25 of slide set 12 titled The ANOVA Approach to the Analysis of Linear Mixed-Effects Models.

The slide addresses the estimation of estimable $\mathbf{C}\beta$ and provides an expression for the variance $\mathbf{\Sigma} \equiv \text{Var(y)}$ and states that

$$\hat{\beta}_{GLS} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y = (X^\top X)^{-1} X^\top y = \hat{\beta}_{OLS}. \tag{2}$$

Thus, the GLS estimator of any estimable $\mathbf{C}\beta$ is equal to the OLS estimator in this special case.

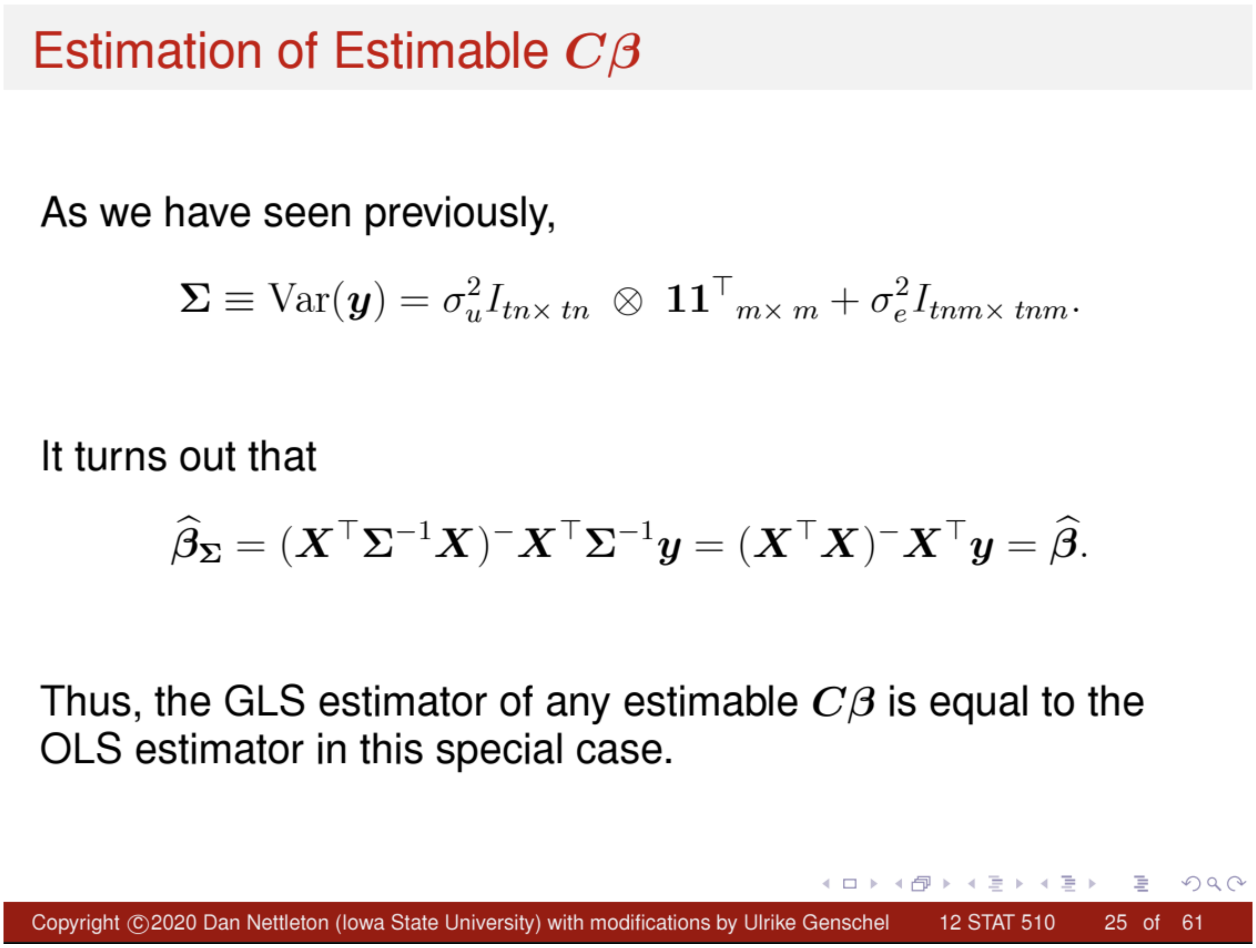

Figure 1: Slide 25: CocoMelon

---

a)

What conditions have to be fulfilled for the result in (2) to hold?

Answer

What conditions have to be fulfilled for the result in (2) to hold?

1. The variance-covariance matrix $\mathbf{\Sigma}$ of the response vector $\mathbf{y}$ must satisfy:

$$\mathbf{\Sigma} = \sigma^2 \mathbf{I}$$

This implies:

- Homoscedasticity: constant variance across observations.
- No correlation between observations.
- The errors $\boldsymbol{\varepsilon}$ are i.i.d. with variance σ^2.

2. It is sufficient that:

$$\mathbf{X}^\top \mathbf{\Sigma}^{-1} \mathbf{X} = \mathbf{X}^\top \mathbf{X}$$

That is, the matrix $\mathbf{\Sigma}^{-1}$ acts like the identity on the column space of $\mathbf{X}$, even if $\mathbf{\Sigma}$ is not a scalar multiple of the identity on the full space.

3. The design matrix $\mathbf{X}$ must be of full column rank so that:

$$(\mathbf{X}^\top \mathbf{X})^{-1}$$

exists.

This ensures that both GLS and OLS estimators are uniquely defined.

---

b)

Verify the result in (2) assuming the conditions are met.

Answer

Assume that the assumptions from the prior part hold.

Then:

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\mathbf{I}.$$

Substitute this into the expression for the GLS estimator:

$$\hat{\boldsymbol{\beta}}_{\mathbf{\Sigma}} = \left(\mathbf{X}^\top \mathbf{\Sigma}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{\Sigma}^{-1} \mathbf{y} = \left(\mathbf{X}^\top \left(\frac{1}{\sigma^2}\mathbf{I}\right) \mathbf{X}\right)^{-1} \mathbf{X}^\top \left(\frac{1}{\sigma^2}\mathbf{I}\right) \mathbf{y}$$

Factor out $\frac{1}{\sigma^2}$:

$$= \left(\frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{y}\right)$$

Now use the identity $(aA)^{-1} = \frac{1}{a}A^{-1}$ to simplify:

$$= \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \cdot \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{y} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} = \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}.$$

So, under the assumption that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, the GLS estimator reduces exactly to the OLS estimator. This verifies the result in Equation (2).