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 $j = 1, \dots, 60$). Here $u_1, \dots, u_{60}$ 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 $d_1, \dots, d_{20}$. Describe the distribution of these differences considering the assumptions about the distribution of the original scores in model (1).

Answer

Based on the model assumptions of $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$, for each subject $j = 1, \dots, 20$,

$$\begin{aligned} d_j &= y_{1j} - y_{2j} \\ &= \mu_1 + u_j + e_{1j} - (\mu_2 + u_j + e_{2j}) \\ &= (\mu_1 - \mu_2) + e_{1j} - e_{2j} \end{aligned}$$

$$\begin{aligned} E(d_j) = \mu_1 - \mu_2, &\quad \mathrm{Var}(d_j) = \mathrm{Var}(e_{1j}) + \mathrm{Var}(e_{2j}) = 2\sigma_e^2 \\ \mathrm{Cov}(d_j, d_{j'}) &= \mathrm{Cov}(e_{1j} - e_{2j}, e_{1j'} - e_{2j'}) = 0 \quad \text{for } j \neq j' \end{aligned}$$

Thus, $d_j \overset{iid}{\sim} N(\mu_1 - \mu_2, 2\sigma_e^2)$, which is a constant mean model:

$$\boldsymbol{d} = \mathbf{1}[\mu_1 - \mu_2] + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, 2\sigma_e^2 \mathbf{I})$$

b)

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

Answer

Let $\bar{d} = \frac{1}{20} \sum_{j=1}^{20} d_j$. Then $\bar{d} \sim N(\mu_d, \frac{\sigma_d^2}{20})$, where $\mu_d = \mu_1 - \mu_2$, $\sigma_d^2 = 2\sigma_e^2$.
Under $H_0 : \mu_d = 0$, the test statistic is:

$$t = \frac{\bar{d}}{\sqrt{\mathrm{Var}(\bar{d})}} = \frac{\bar{d}}{\sqrt{\frac{\hat{\sigma}_d^2}{20}}} = \frac{\bar{d}}{\sqrt{\frac{1}{20} \cdot \frac{1}{19} \sum_{j=1}^{20} (d_j - \bar{d})^2}}$$

$$F = t^2 = \frac{380 \cdot \bar{d}^2}{\sum_{j=1}^{20} (d_j - \bar{d})^2}$$

c)

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

Answer

$$t \sim t_{19}\left(\frac{\mu_d}{\sqrt{\sigma_d^2/20}}\right) \stackrel{d}{=} t_{19}\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_e^2/10}}\right)$$

$$F \sim F_{1,19}\left(\frac{5(\mu_1 - \mu_2)^2}{\sigma_e^2}\right)$$

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

Model for all scores:

$$\mathbf{y} = [\mathbf{I}_{2\times 2} \otimes \mathbf{1}_{20\times 1}] \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \boldsymbol{\epsilon}$$

Let: - a_j and b_j be scores from groups 1 and 2 - $\bar{a} = \frac{1}{20}\sum a_j$, $\bar{b} = \frac{1}{20}\sum b_j$

Then:

$$\text{Var}(\bar{a} - \bar{b}) = \text{Var}(\bar{a}) + \text{Var}(\bar{b}) = \frac{1}{10}(\sigma_u^2 + \sigma_e^2)$$

$$\hat{\text{MSE}} = \frac{1}{380}\left(\sum_{j=1}^{20}(a_j - \bar{a})^2 + \sum_{j=1}^{20}(b_j - \bar{b})^2\right)$$

$$\text{CI}_{95\%} = (\bar{a} - \bar{b}) \pm t_{38,0.975} \cdot \sqrt{\frac{1}{380}\left(\sum_{j=1}^{20}(a_j - \bar{a})^2 + \sum_{j=1}^{20}(b_j - \bar{b})^2\right)}$$

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

From part b):

$$\hat{\sigma}_d^2 = 2\hat{\sigma}_e^2 = \frac{1}{19}\sum_{j=1}^{20}(d_j - \bar{d})^2$$

From part d):

$$\sigma_u^2 + \sigma_e^2 = \frac{1}{38}\left(\sum_{j=1}^{20}(a_j - \bar{a})^2 + \sum_{j=1}^{20}(b_j - \bar{b})^2\right)$$

Solving:

$$\begin{cases} \hat{\sigma}_e^2 = \frac{1}{38}\sum_{j=1}^{20}(d_j - \bar{d})^2 \\ \hat{\sigma}_u^2 = \frac{1}{38}\left(\sum_{j=1}^{20}(a_j - \bar{a})^2 + \sum_{j=1}^{20}(b_j - \bar{b})^2\right) - \hat{\sigma}_e^2 \end{cases}$$

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

Both $\bar{d}$ and $\bar{a} - \bar{b}$ are independent unbiased estimators. So:

$$\begin{aligned} \mu_1 - \mu_2 &= \frac{\mathrm{Var}^{-1}(\bar{d})}{\mathrm{Var}^{-1}(\bar{d}) + \mathrm{Var}^{-1}(\bar{a} - \bar{b})} \cdot \bar{d} + \frac{\mathrm{Var}^{-1}(\bar{a} - \bar{b})}{\mathrm{Var}^{-1}(\bar{d}) + \mathrm{Var}^{-1}(\bar{a} - \bar{b})} \cdot (\bar{a} - \bar{b}) \\ &= \frac{\sigma_u^2 + \sigma_e^2}{\sigma_u^2 + 2\sigma_e^2} \cdot \bar{d} + \frac{\sigma_e^2}{\sigma_u^2 + 2\sigma_e^2} \cdot (\bar{a} - \bar{b}) \end{aligned}$$

Q2

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

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

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 $u = [u_1, u_2, \ldots, u_{60}]^\top$. If possible, use Kronecker product notation to simplify your answer.

Answer

Overall, we have 60 subjects, giving our experimental units. We also have 80 responses in total (rows in both our $\mathbf{X}$ and $\mathbf{Z}$ matrices), corresponding to an imbalanced design due to partial replication.

Supposing the responses were sorted first by subject and then by drink into a response vector, and modeled by:

$$y_{ij} = \mu_i + u_j + e_{ij}, \quad i = 1, 2, \quad j = 1, \ldots, 60,$$

Each row of $\mathbf{X}$, fixed treatment effects, corresponds to receiving one of the two sports drinks, 1 in the first column if the observation (not experimental unit) received sports drink 1 and 0 otherwise, and mutually exclusive between columns.

And each row of $\mathbf{Z}$, random effects, corresponds to one observation and contains a 1 in the column corresponding to the experimental unit, and 0 elsewhere, and mutually exclusive between columns.

And it is given that:

$$\beta = [\mu_1, \mu_2]^\top$$

$$\mathbf{u} = [u_1, u_2, ...u_{60}]^\top$$

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

The top block ($\mathbf{1}_{20\times1} \otimes \mathbf{I}_{2\times2}$) corresponds to the 20 subjects who received both drinks, totaling 40 observations.

The bottom block ($\mathbf{I}_{2\times2} \otimes \mathbf{1}_{20\times1}$) corresponds to the remaining 40 observations from subjects who received only one drink, 20 observations for each drink.

Taken together:

Matrix X:

$$X_{80\times2} = \begin{bmatrix} \mathbf{1}_{20\times1} \otimes I_{2\times2} \\ I_{2\times2} \otimes \mathbf{1}_{20\times1} \end{bmatrix}$$

And

Top-left block: $I_{20\times20} \otimes \mathbf{1}_{2\times1}$ corresponds to the 20 subjects who completed both drink trials–this gives 2 rows per subject, totaling 40 rows.

Bottom-right block: $I_{40\times40}$ corresponds to the 40 subjects who completed only one trial, contributing one row each.

Off-Diagonal Elements: The two subject groups are non-overlapping such that they are independent and have zero-valued entries.

Taken together:

Matrix Z:

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

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

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

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

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

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

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

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

Note:

The last step relies upon a few things. To begin with, the cross terms of the square cancel out (turn to zero). This is because:

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

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

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

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

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

We then have:

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

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 $\mathbf{\Sigma} \equiv \text{Var}(\text{y})$ and states that

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

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

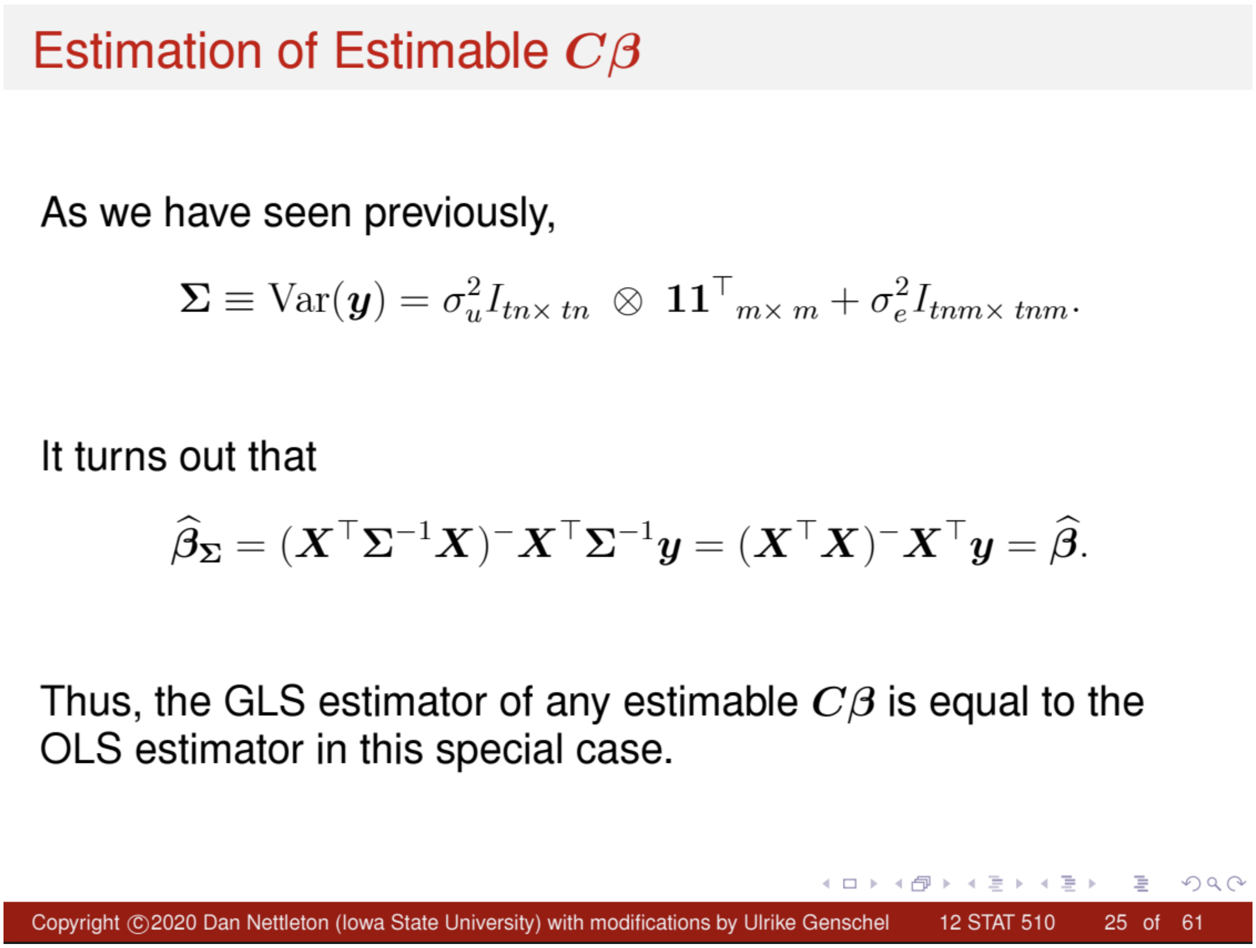

Figure 1: Slide 25: CocoMelon

---

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 first condition is multi-faceted, but I would say falls under the perview of the "Suppose…" from Slide 2 of the Chapter 12 Slides. That is, we assume:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_u^2\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \sigma_e^2\mathbf{I}\end{bmatrix}\right), \text{ where } \sigma_u^2, \sigma_e^2 \in \mathbb{R}^+ \text{ are unknown variance components.}$$

This means:

- The variance-covariance matrix $\mathbf{\Sigma}$ of the response vector $\mathbf{y}$ must satisfy $\mathbf{\Sigma} = \sigma^2\mathbf{I}$
- The errors are iid with constant variance σ^2.

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

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

This ensures that the normal equations have a unique solution for $\boldsymbol{\beta}$, and thus both the GLS and OLS estimators are uniquely defined.

Without full column rank, $(\mathbf{X}^\top\mathbf{X})^{-1}$ does not exist, and we would need to use a generalized inverse to define the estimators, in which case the equivalence in (2) may not hold.

3. For estimable $\mathbf{C}\beta$, $\mathbf{C}\hat{\beta}_{\hat{\mathbf{\Sigma}}} = \mathbf{C}\hat{\beta}$

This is to say that:

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

Meaning the inverse of the variance-covariance 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.

---

b)

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

Answer

Assume that the assumptions from part a) all hold.

Then, for:

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

Given:

$$\hat{\beta}_{GLS} = (\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{y}$$

Via substitution we have:

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

We may treat $\frac{1}{\sigma^2}$ as a scalar for the purposes of matrix algebra, such that we have:

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

Noting that $(aA)^{-1} = \frac{1}{a}A^{-1}$, applying to the above and simplifying:

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

Noting that by definition, noting the normal equations, OLS is given by:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

So, the primary condition required for the above to work is for $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$.