

# Partial Methods 1 Exam

2025-03-25

## Outline

- 4: d), e), f) for additional edits
- 5: b) onwards
- 6: all but b)

### 4.

Consider a one-way ANOVA model with two levels and two observations at each level,

$$E(y_{ij}) = \mu + \alpha_i, \quad i, j = 1, 2$$

Using the given information, we may represent it in matrix form to aid in solving the following questions:

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

(a) Is  $\alpha_1$  estimable? Show work to justify why it is or is not estimable.

- If  $C$  is any  $q \times p$  matrix, we say that the linear function of  $\boldsymbol{\beta}$  given by  $C\boldsymbol{\beta}$  is **estimable** if and only if  $C = A\mathbf{X}$  for some matrix  $q \times n$  matrix  $A$ .

Figure 1: Ch. 2: Some Key Linear Models Results

The definition of estimability is given above as first defined in Chapter 2 of lecture notes.

For  $\alpha_1$ : Can we find a matrix  $A$  to satisfy the above?

We have:

$$c^T = [0 \quad 1 \quad 0]$$

There in fact does not exist a matrix  $A$  such that:

$$C = A\mathbf{X}$$

This means that  $\alpha_1$  is not estimable (this is also descriptively found by noting that the matrix  $C$  is not in row space of  $\mathbf{X}$ ).

Generally, parameters (elements of the  $\beta$  vector) are only estimable if they can be expressed as a linear combination of the rows of the design matrix,  $\mathbf{X}$ . For our purposes, the model matrix does not have full rank, so we know there are elements of the parameter vector that are not estimable, though we still have linear combinations that are (more on that in the following).

(b) Provide a quantity that is estimable.  $\alpha_1 - \alpha_2$  or  $\mu + \alpha_1$ .

We again utilize the above definition to determine estimability. We find there are more than one quantity (linear combination of parameters) that are estimable, despite our design matrix not being full rank. Two examples:

For  $\alpha_1 - \alpha_2$ , an estimable quantity:

$$c^T = [0 \quad 1 \quad -1] = 1 \cdot [1 \quad 1 \quad 0] + (-1) \cdot [1 \quad 0 \quad 1]$$

The corresponding matrix  $A$  (a row vector) satisfying  $C = A\mathbf{X}$  is:

$$A = [1 \quad 0 \quad -1 \quad 0]$$

Another estimable quantity,  $\mu + \alpha_1$  has  $C$  matrix:

$$c^T = [1 \quad 1 \quad 0]$$

But this is just the first row of  $\mathbf{X}$ ! So the corresponding  $A$  matrix is:

$$A = [1 \quad 0 \quad 0 \quad 0]$$

The remaining questions refer to any general linear model as discussed in class. Thus, provide answers for a general  $\mathbf{X}$  instead of referring to the particular  $\mathbf{X}$  defined above.

Additional Note: The following are related to the “Prerequisite Knowledge” Slides and an extension of the OLS Slides (Chapters 2 and 3) from lectures. Of note:

- The set of all possible linear combinations of the columns of  $\mathbf{A}$  is called the column space of  $\mathbf{A}$  and is written as

$$\mathcal{C}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x} = \mathbf{A}\mathbf{c} \text{ for some } \mathbf{c} \in \mathbb{R}^n\}.$$

Any member of the set  $\mathcal{C}(\mathbf{A})$  is an  $m \times 1$  vector, so  $\mathcal{C}(\mathbf{A})$  is a subset of  $\mathbb{R}^m$ .

Figure 2: Linear Algebra Overview

Suppose  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ,  $E(\boldsymbol{\epsilon}) = 0$ ,  $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$

- $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$  with  $\boldsymbol{\beta}$  unknown,  $\mathbf{X}$  is full-rank

Figure 3: Column Rank

(c) **True or False** Circle the appropriate choice. The expected value of any observation is only estimable when  $\mathbf{X}$  has full column rank.

False.

The expected value of any observation is estimable as long as it is a linear combination of the columns of  $\mathbf{X}$ . Having full column rank just ensures that every element of the parameter vector,  $\beta$ , is estimable. There are examples of observations able to be estimated without having full column rank, such as the second estimable quantity in the above part b).

(d) The set of vectors  $\mathbf{c}$  for which  $\mathbf{c}^T \beta$  is estimable forms a vector space. Specify the vector space.

Answer: **The column space of  $\mathbf{X}$ .**

**Note: Add image specifically for vector spaces here from Linear Algebra Overview**

(e) **Fill in the blank.**

The column rank of a model matrix  $\mathbf{X}$  is always **equal to** the number of linearly independent vectors that span the vector space in part (d).

**Note: Add connection of linearly independent vectors to estimability**

(f) What is the relationship/connection between the column rank of  $\mathbf{X}$  and the estimability of  $\beta$ ? Answer using a short sentence.

The higher the column rank of  $\mathbf{X}$ , the more parameters in  $\beta$  that are estimable. The column rank of  $\mathbf{X}$  determines the number of estimable parameters, specifically.

## 5.

Consider the following linear model with  $n = 5$  observations:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

Note that the columns of  $\mathbf{X}$  are perpendicular so that  $\mathbf{X}^T \mathbf{X}$  is diagonal.

(a) In a Gauss-Markov version of this model, which of the parameters,  $\beta_1, \beta_2, \beta_3$ , can be estimated with greatest precision? Explain carefully.

In a Gauss-Markov version of this model,  $\beta_1$  can be estimated with the greatest precision.

$\beta_1$  corresponds to the first column of the design matrix,  $\mathbf{X}$ , which has the largest diagonal entry in  $\mathbf{X}^T \mathbf{X}$ , corresponding to the 5 rows we can use in estimating  $\beta_1$ . This means we have more information to estimate  $\beta_1$  which in turn corresponds to a smaller variance in estimation.

Calculating explicitly:

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

By contrast, the other two parameter vectors,  $\beta_2$  and  $\beta_3$  can only be estimated using 4 of the rows, so they use comparatively less information and would have greater variance, less precision.

(b) Suppose  $\mathbf{y}$  is such that  $\text{SSE} = 3$  and  $\hat{\beta} = (5 \ 6 \ 2)^T$ . Consider an analysis under the Gauss-Markov model with Normal errors and the following two null hypotheses:

$$H_{0,1} : E(y_1) = E(y_2) \quad \text{and} \quad H_{0,2} : E(y_1) = E(y_5)$$

- i. Write  $H_{0,1}$  and  $H_{0,2}$  in testable form  $H_0 : \mathbf{C}\beta = 0$  by identifying an appropriate matrix  $\mathbf{C}$ . (**Hint:** Start out by expressing each expected value as a function of  $\beta$  given  $\mathbf{X}$  and  $\beta$  as defined above.)

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & -1 \end{pmatrix}$$

**Note: Add More Here**

- ii. Based on  $\mathbf{C}$  compute an  $F$  statistic for testing  $H_0$  (you need not do the arithmetic, but plug correct numbers into a correct formula).

The formula we use for this problem is given by:

$$F = \frac{(\mathbf{C}\hat{\beta})^T [\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T]^{-1} (\mathbf{C}\hat{\beta})/q}{\text{MSE}}$$

where  $q$  is the number of linear constraints.

**Note: Add More Here**

- iii. Specify the reference distribution of  $F$  under the null hypothesis.

The reference distribution of  $F$  under the null hypothesis is an  $F$ -distribution with  $q$  and  $n - p$  degrees of freedom, where  $p$  is the number of parameters.

**Note: Add More Here, Particularly Non-Centrality Parameter**

## 6.

Consider a completely randomized experiment in which a total of 10 freshly cut Gerber daisies were placed into 10 vases (one daisy per vase). The Gerber daisies were randomly assigned to five treatment groups with two Gerber daisies in each treatment group. The treatment corresponds to the amount of a chemical compound added to the water in each vase. Of interest is the longevity of the Gerber daisies measured in days.

Treatment	1	2	3	4	5
Amount of compound (g)	0	2	4	10	16

Suppose for  $i = 1, \dots, 5$  and  $j = 1, 2$ ,  $y_{ij}$  denotes the longevity in days of the study of the  $j^{th}$  Gerber daisy from treatment group  $i$ . Furthermore, suppose

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where the  $\mu_i$  are unknown parameters and the  $\varepsilon_{ij}$  terms are  $\mathcal{N}(0, \sigma^2)$  for some unknown  $\sigma^2 > 0$ .

Use the R code and partial output provided with this exam to answer the following questions.

(a) For the first model fit in R, called M1, specify the model matrix  $\mathbf{X}$  used by R.

The model matrix  $\mathbf{X}$  used by R is a design matrix with an intercept and indicator columns for each level of **amt**. By default R produces a design matrix of full rank using the baseline constraint where the first level of the factor **amt** is used as the baseline. This corresponds to:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(b) Consider the following information from the output associated with model M1:

**F-statistic:** 30.84 on 4 and 5 DF,    **p-value:** 0.001019

Specify the null and alternative hypothesis associated with this test:

$H_0$ : All treatment means (average response per level of **amt**) are equal ( $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ ).

$H_a$ : At least one treatment mean differs from the treatment mean of the other groups (levels of the factor **amt**).

(c) Provide the BLUE of  $\mu_2$ :

By default the estimate of  $\beta_2 = \mu_2 - \mu_1$  due to the R baseline constraint, as noted more in part a). Additionally, R by default and as specified provides the OLS estimated effects, which are BLUE as they correspond to the normal equations under GMM. That is all to say:

The BLUE of  $\mu_2$  is the estimated expression using “intercept + coefficient”, which corresponds to  $\hat{\mu}_2 = 2.32 + 2.41 = 4.73$ .

(d) What is the standard error of the BLUE of  $\mu_2$ ?

The standard error of the BLUE of  $\mu_2$  is 0.8747, as given in the R output.

## Here on down

(e) Provide the BLUE of  $\mu_1 - \mu_2$ :

Under R’s default treatment contrast coding, the model fitted with `lm(y ~ factor(amt))` uses the first level of the factor (i.e., `amt = 0`) as the baseline. Therefore, the model is:

$$y_{ij} = \beta_0 + \beta_2 \cdot \mathcal{K}_{\text{amt}=2} + \beta_3 \cdot \mathcal{K}_{\text{amt}=4} + \cdots + \varepsilon_{ij},$$

where:

- $\beta_0 = \mu_1$  (mean of treatment 1),
- $\beta_2 = \mu_2 - \mu_1$

Then:

$$\mu_1 - \mu_2 = \mu_1 - (\mu_1 + \beta_2) = -\beta_2$$

So, the BLUE of  $\mu_1 - \mu_2$  is:

$$-\hat{\beta}_2 = -2.41$$

(f) What is the standard error of the BLUE of  $\mu_1 - \mu_2$ ?

Since  $\mu_1 - \mu_2 = -\beta_2$ , this is a direct transformation of a single coefficient estimate. Thus, the standard error of the BLUE is simply:

$$\text{SE}(\mu_1 - \mu_2) = \text{SE}(-\hat{\beta}_2) = \text{SE}(\hat{\beta}_2) = 0.8747$$

The earlier expression using  $\sqrt{0.8747^2 + 0.8747^2}$  would be appropriate **only if** the estimate were a difference of **two independent** parameters, which is not the case here. Since we are using just one coefficient ( $\beta_2$ ), we use its standard error directly.

(g) What is the value of  $\mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y}$ , where  $\mathbf{y}$  denotes the vector containing the values of longevity?

The quantity  $\mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y}$  is the **residual sum of squares (RSS)** from the fitted model. It measures the variation in the data not explained by the model.

From the `anova(M1)` output, the residual sum of squares is:

$$\mathbf{y}^T(\mathbf{I} - \mathbf{P}_1)\mathbf{y} = \text{RSS} = 3.826$$

(h) Provide the value of the F-statistic, numerator and denominator df, and the p-value associated with the following ANOVA table:

From the ANOVA table of the model M1, we have:



- **F-statistic:** 30.84
- **Numerator degrees of freedom:** 4 (number of groups – 1 = 5 – 1)
- **Denominator degrees of freedom:** 5 (total n – number of groups = 10 – 5)
- **p-value:** 0.001019

The p-value is calculated as the probability that an  $F_{4,5}$  random variable exceeds the observed value of 30.84. You can compute it directly in R using:

```
pf(30.84, df1 = 4, df2 = 5, lower.tail = FALSE)
```

```
## [1] 0.001018976
```

```
OLD
```

```
> anova(M1)
```

```
Analysis of Variance Table
```

```
Response: longevity
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
amt	---	94.398	---	---	---
Residuals	---	3.826			

```
NEW
```

```
> anova(M1)
```

```
Analysis of Variance Table
```

```
Response: longevity
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
amt	4	94.398	23.5995	30.84	0.001019
Residuals	5	3.826	0.7652		

For the general case, consider the test statistic

$$F = \frac{\mathbf{y}^\top (\mathbf{P}_X - \mathbf{P}_{X_0}) \mathbf{y} / [\text{rank}(\mathbf{X}) - \text{rank}(\mathbf{X}_0)]}{\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_X) \mathbf{y} / [n - \text{rank}(\mathbf{X})]}.$$

Figure 4: Matrix Notation: Form of Partial F-Test

Thus, the  $F$  statistic has the familiar form

$$\frac{(SSE_{\text{REDUCED}} - SSE_{\text{FULL}}) / (DFE_{\text{REDUCED}} - DFE_{\text{FULL}})}{SSE_{\text{FULL}} / DFE_{\text{FULL}}}.$$

Figure 5: Familiar Form of Partial F-Test

(i) Look at the output associated with Model 2, M2.

i. Fill in the missing entries in the ANOVA table produced by the R command `anova(M2, M1)`.

```
OLD
> anova(M2, M1)
Analysis of Variance Table

Model 1: longevity ~ amount
Model 2: longevity ~ amt

  Res.Df  RSS    Df Sum of Sq  F      Pr(>F)
1     ---   ---    ---      ---    ---
2     ---   ---    ---  ---(*)    ---  0.006412 **
```

```
NEW
> anova(M2, M1)
Analysis of Variance Table

Model 1: longevity ~ amount
Model 2: longevity ~ amt

  Res.Df  RSS    Df Sum of Sq  F      Pr(>F)
1      9  37.757    1      33.931  2.02  0.006412 **
2      5   3.826    4      33.931  2.02  0.006412 **
```

ii. Provide an interpretation of Sum of Squares in part (i). This is the value denoted by (\*).

The sum of squares (\*) represents the additional variability explained by adding the **amount** factor to the model.

iii. Provide a conclusion in the context of the data about the null hypothesis that is tested in part (i).

Conclusion: The small p-value (0.006412) suggests we reject the null hypothesis, concluding that the amount of compound has a statistically significant effect on longevity.

R code and output for Question 6:

Name \_\_\_\_\_

```
> amount<-c(0, 0, 2, 2, 4, 4, 10, 10, 16, 16)
> plot(amount, longevity, ylim = c(0,14))
> amt=as.factor(amount)
> M1<-lm(longevity~amt)
> summary(M1)
```

Call:

```
lm(formula = longevity ~ amt)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.3200	0.6185	3.751	0.013281 *
amt2	2.4100	0.8747	2.755	0.040063 *
amt4	6.0850	0.8747	6.957	0.000943 ***
amt10	8.2400	0.8747	9.420	0.000227 ***
amt16	7.0300	0.8747	8.037	0.000482 ***

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

Residual standard error: 0.8747 on 5 degrees of freedom

Multiple R-squared: 0.9611, Adjusted R-squared: 0.9299

F-statistic: 30.84 on 4 and 5 DF, p-value: 0.001019

```
> anova(M1)
```

Analysis of Variance Table

Response: longevity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
amt		94.398			
Residuals		3.826			

---

```
> is.numeric(amount)
```

```
[1] TRUE
```

```
> M2<-lm(longevity~amount)
```

```
> anova(M2)
```

Analysis of Variance Table

Response: longevity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
amount		60.467			
Residuals		37.757			

Figure 6: CocoMelon