Partial Methods 1 Exam

2025-03-25

4.

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

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

Overall Setup

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

(a) Is α_1 estimable? Show work to justify why it is or is not estimable.

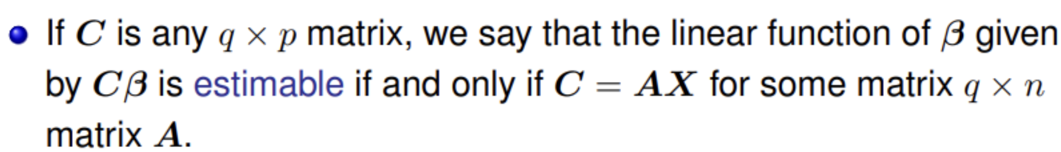

Figure 1: Ch. 2: Some Key Linear Models Results

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

For α_1:

$$c^T = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

There does **not** exist a matrix A such that:

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

Hence, α_1 is **not estimable** (this is not in row space of $\mathbf{X}$ and is not estimable).

Using the above definition, no, α_1 is not estimable.

Parameters are only estimable if they can be expressed as a linear combination of the rows of the design matrix. Additionally, the model matrix does not have full rank, which limits what is estimable in this scenario.

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

For $\alpha_1 - \alpha_2$:

$$c^T = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} = 1 \cdot \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} + (-1) \cdot \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}$$

Then, the corresponding matrix A (a row vector in $\mathbb{R}^{1\times 4}$) satisfying $C = A\mathbf{X}$ is:

$$A = \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix}$$

For $\mu + \alpha_1$:

$$c^T = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$$

This is exactly the first row of $\mathbf{X}$, so:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

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.

Noting the "Prerequisite Knowledge" Slides and an extension of the OLS Slides used for the following.

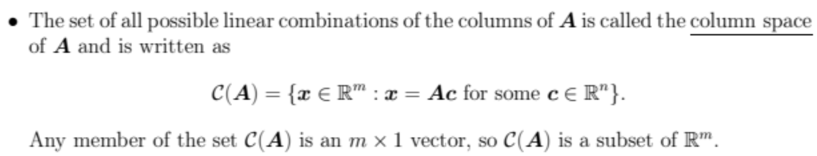

Figure 2: Linear Algebra Overview

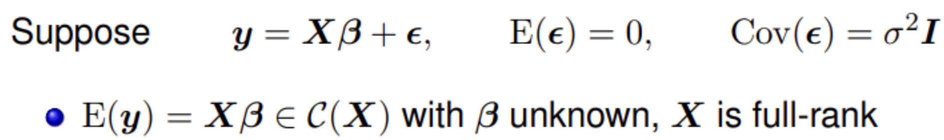

Figure 3: Column Rank

(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 $\mathbf{X}$, not necessarily requiring full column rank.

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

Answer: **The column space of X**.

(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).

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

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

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, β_1 can be estimated with the greatest precision. This is because β_1 corresponds to the first column of $\mathbf{X}$, which has the largest diagonal entry in $\mathbf{X}^T\mathbf{X}$ (indicating more information and smaller variance).

(b) Suppose $\mathbf{y}$ is such that SSE $= 3$ and $\hat{\boldsymbol{\beta}} = (5\ 6\ 2)^T$. 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 $H_0 : \mathbf{C}\boldsymbol{\beta} = 0$ by identifying an appropriate matrix $\mathbf{C}$. (**Hint:** Start out by expressing each expected value as a function of $\boldsymbol{\beta}$ given $\mathbf{X}$ and $\boldsymbol{\beta}$ as defined above.)

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

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).

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

where q is the number of linear constraints.

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.

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 μ_i are unknown parameters and the ε_{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.

For the first model fit in R, called M1, the model matrix $\mathbf{X}$ used by R is a design matrix with an intercept and indicator columns for each level of `amt`.

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

- H_0: All treatment means are equal ($\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$).
- H_a: At least one treatment mean differs.

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

Specify the null and alternative hypothesis associated with this test:

(c) Provide the BLUE of μ_2:

The BLUE of μ_2 is the intercept + coefficient for `amt2` = 2.32 + 2.41 = 4.73.

(d) What is the standard error of the BLUE of μ_2?

The standard error of the BLUE of μ_2 is 0.8747.

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

The BLUE of $\mu_1 - \mu_2$ is $2.32 - 4.73 = -2.41$.

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

The standard error of the BLUE of $\mu_1 - \mu_2$ is $\sqrt{0.8747^2 + 0.8747^2} = 1.237$ (calculated as the square root of the sum of variances).

(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 value of $\mathbf{y}^T(\mathbf{I} - \mathbf{P_1})\mathbf{y}$ is 3.826 (the residual sum of squares).

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

The F-statistic is 30.84 with numerator df = 4, denominator df = 5, and p-value = 0.001019.

p-value: probability that a random variable with distribution $F_{q,n-r}$ matches or exceeds the observed value of the test statistic F=30.84.

```
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

$$F = \frac{\boldsymbol{y}^\top(\boldsymbol{P_X} - \boldsymbol{P_{X_0}})\boldsymbol{y}/\left[\operatorname{rank}(\boldsymbol{X}) - \operatorname{rank}(\boldsymbol{X}_0)\right]}{\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}/\left[n - \operatorname{rank}(\boldsymbol{X})\right]}.$$

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
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