STAT612 Theory of Linear Model Last update: October 16, 2019

Notes 7: Statistical Inference I

Book reference: Monahan (2008) *Instructor: Raymond Wong*

natural param $\begin{bmatrix} b/\sigma^2 \\ 1/\sigma^2 \end{bmatrix} \in \mathbb{R}^{p+1}$

sufficient stat $\mathbb{R}^{p+1}$ $\begin{bmatrix} X^Ty \\ y^Ty \end{bmatrix} \leftarrow y \in \mathbb{R}^N$

1 Normal Gauss-Markov model

e.g.? N.

In this notes, we assume $\mathbf{y} \sim N_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$, where $\mathbf{b} \in \mathbb{R}^p$ and $\sigma^2 > 0$ are unknown parameters. We call this model the normal Gauss-Markov model. Since the covariance matrix $\sigma^2\mathbf{I}$ is non-singular, $\mathbf{y}$ has a probability density function:

p.

$$f(\mathbf{y}|\mathbf{b}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{b})^\intercal(\mathbf{y} - \mathbf{X}\mathbf{b})\right\}.$$

y: Data

X fixed

Our goal here is to develop several inference tools for the linear model based on the normal Gauss-Markov assumption.

$\exp\{-\frac{1}{2\sigma^2}y^Ty + \frac{y^TXb}{\sigma^2} - \frac{1}{2\sigma^2}b^TX^TXb\}$.

2 Uniformly minimum variance unbiased estimator

Previously, we have concluded that least squares estimators are the best estimators among the class of linear unbiased estimators under the Gauss-Markov model. In fact, the normal Gauss-Markov model allows us to say more about the optimality of least squares estimator. We look into this before we talk about inference.

By Lehmann-Scheffé theorem, we can obtain the following corollary. $\hat{b} = (X^TX)^{-1}X^Ty.$

Corollary 7.1. *Under the normal Gauss-Markov model, the least squares estimator* $\boldsymbol{\lambda}^\intercal\hat{\mathbf{b}}$ *of an estimable function* $\boldsymbol{\lambda}^\intercal\mathbf{b}$ *has the smallest variance among all unbiased estimator.*

The proof is omitted. For those who know Lehmann-Scheffé theorem, refer to Section 6.2 of Monahan (2008) for the proofs of the conditions required to apply Lehmann-Scheffé theorem.

Corollary 7.2. *Let* $\boldsymbol{\Lambda}$ *be a fixed constant such that each element of* $\boldsymbol{\Lambda}^\intercal\mathbf{b}$ *is estimable under the normal Gauss-Markov model. Then the least squares estimator* $\boldsymbol{\Lambda}^\intercal\hat{\mathbf{b}}$*, where* $\hat{\mathbf{b}}$ *is the solution to the normal equations, is the best unbiased estimator of* $\boldsymbol{\Lambda}^\intercal\mathbf{b}$ *such that* $\mathsf{Var}(\mathbf{z}) - \mathsf{Var}(\boldsymbol{\Lambda}^\intercal\hat{\mathbf{b}})$ *is nonnegative definite for any unbiased estimator* $\mathbf{z}$ *such that* $\mathsf{E}(\mathbf{z}) = \boldsymbol{\Lambda}^\intercal\mathbf{b}$.

$\hat{\sigma}^2 = \frac{1}{n-p} y^T(I-P_x)y.$

$= \frac{1}{n-p}(y^Ty - y^TX(X^TX)^{-1}X^Ty).$

3 Maximum likelihood estimation

Under the normal Gauss-Markov model, the likelihood function of the unknown parameters $(\mathbf{b}, \sigma^2)$ is

$$L(\mathbf{b}, \sigma^2) = f(\mathbf{y}|\mathbf{b}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}Q(\mathbf{b})\right\}.$$

Recall that $Q(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$. The log-likelihood function can be expressed as

$$\ell(\mathbf{b}, \sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}Q(\mathbf{b}).$$

$\hat{\sigma}^2 = \frac{SSE}{N-p}$

Take partial derivatives with respect to σ^2 and $\hat{\mathbf{b}}$, and set them to zero:

$$\begin{array}{l} \frac{\partial}{\partial\sigma^2} \\ \frac{\partial}{\partial b} \end{array} \begin{cases} -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} Q(\mathbf{b}) & = 0 \\ \frac{1}{2\sigma^2} \frac{\partial Q(\mathbf{b})}{\partial \mathbf{b}} & = \mathbf{0} \qquad \text{(normal equations)} \end{cases}$$

Therefore, the maximum likelihood estimator (MLE) of $\mathbf{b}$ and σ^2 are the least squares estimator $\hat{\mathbf{b}}$, and $Q(\hat{\mathbf{b}})/N = SSE/N$ respectively.

Note that the MLE of σ^2 is not unbiased.

Due to the invariance property of MLE, we have the following corollary.

Corollary 7.3. *Under the normal Gauss-Markov model, the maximum likelihood estimator of estimable functions* $\mathbf{\Lambda}^\intercal \mathbf{b}$ *is* $\mathbf{\Lambda}^\intercal \hat{\mathbf{b}}$, *where* $\hat{\mathbf{b}}$ *is the least squares estimator.*

4 Testing general linear hypothesis

In this section, we develop tools for testing general linear hypothesis such as:

1. $H_0 : b_j = 0$
2. $H_0 : b_1 = b_2 = b_3 = 0$
3. $H_0 : b_1 + b_3 = 1$, $b_2 = 3$
4. $H_0 : b_2 = b_3 = b_4$
5. $H_0 : \mathbf{b} \in \mathcal{C}(\mathbf{B})$

The general form of hypotheses of interest is

$$H_0 : \mathbf{K}^\intercal \mathbf{b} = \mathbf{c} \quad \text{versus} \quad H_1 : \mathbf{K}^\intercal \mathbf{b} \neq \mathbf{c},$$

where $\mathbf{K} \in \mathbb{R}^{p \times s}$ with full column rank and $\mathbf{c} \in \mathbb{R}^s$. The full-column-rank assumption of $\mathbf{K}$ follows from the avoidance of any redundancies in writing these hypotheses. Note that, since usually $s > 1$, we only consider two-sided tests here.

To test H_0, we must have each component of $\mathbf{K}^\intercal \mathbf{b}$ estimable, which is equivalent to that every column of $\mathbf{K}$ belongs to $\mathcal{C}(\mathbf{X}^\intercal)$.

Definition 7.4. *The general linear hypothesis* $H_0 : \mathbf{K}^\intercal \mathbf{b} = \mathbf{c}$ *is testable if and only if* $\mathbf{K}$ *has full-column rank and each component of* $\mathbf{K}^\intercal \mathbf{b}$ *is estimable. If any of* $\mathbf{K}^\intercal \mathbf{b}$ *are not estimable, this hypothesis is consdiered as nontestable.*

$\Leftrightarrow K \subset \mathcal{C}(X^T)$

Example 1

Consider $H_0 : \mathbf{b} \in \mathcal{C}(\mathbf{B})$. To write this as $\mathbf{K}^\intercal \mathbf{b} = \mathbf{c}$, we have to find the corresponding $\mathbf{K}$ and $\mathbf{c}$. One can first construct a basis $\{\mathbf{a}_1, \ldots, \mathbf{a}_s\}$ of $\mathcal{N}(\mathbf{B}^\intercal)$. Take $\mathbf{K} = [\mathbf{a}_1, \ldots, \mathbf{a}_s]$ and $\mathbf{c} = \mathbf{0}$. If $\mathcal{N}(\mathbf{B}^\intercal) \subseteq \mathcal{C}(\mathbf{X}^\intercal)$, H_0 is testable.

Example 2

Consider a one-way ANOVA model

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, 2, 3,$$

and test the hypothesis of linear trend, i.e. $\alpha_i = \beta i$ for some β. A common way to express the hypothesis is $H_0 : \alpha_2 - \alpha_1 = \alpha_3 - \alpha_2$ or $H_0 : \alpha_1 - 2\alpha_2 + \alpha_3 = 0$. If $\mathbf{b} = [\mu, \alpha_1, \alpha_2, \alpha_3]^\intercal$, we can write $\mathbf{K} = [0, 1, -2, 1]^\intercal$ and $\mathbf{c} = 0$. From Example 2 of Notes 4, we know that this is testable.

4.1 F test

Assume the normal Gauss-Markov model. If $\mathbf{K}^\intercal\mathbf{b}$ is estimable, we know that its BLUE is the least squares estimator $\mathbf{K}^\intercal\hat{\mathbf{b}} = \mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{X}^\intercal\mathbf{y}$, which is a linear transformation of a multivariate normal vector $\mathbf{y} \sim N_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$. Due to unbiasedness, $\mathsf{E}(\mathbf{K}^\intercal\hat{\mathbf{b}}) = \mathbf{K}^\intercal\mathbf{b}$. Similarly as in Example 2 of Notes 5, we can show that $\mathsf{Var}(\mathbf{K}^\intercal\hat{\mathbf{b}}) = \sigma^2\mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K}$ (a symmetric matrix). Therefore

$$\mathbf{K}^\intercal\hat{\mathbf{b}} \sim N_s(\mathbf{K}^\intercal\mathbf{b}, \sigma^2\mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K}).$$

Proposition 7.5. *Under the normal Gauss-Markov model $\mathbf{y} \sim N_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$, if $\mathbf{K}^\intercal\mathbf{b}$ is estimable with $\mathbf{K} \in \mathbb{R}^{p\times s}$ such that* $\operatorname{rank}(\mathbf{K}) = s$, *then $\mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K}$ is non-singular.*

As the covariance matrix of $\mathbf{K}^\intercal\hat{\mathbf{b}}$ is non-singular, we can adopt Theorem 6.16 to construct the distribution of the quadratic form:

$$(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})^\intercal(\sigma^2\mathbf{H})^{-1}(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c}) \sim \chi^2_s\left(\frac{1}{2}(\mathbf{K}^\intercal\mathbf{b} - \mathbf{c})^\intercal(\sigma^2\mathbf{H})^{-1}(\mathbf{K}^\intercal\mathbf{b} - \mathbf{c})\right),$$

where $\mathbf{H} = \mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K}$.

Recall the residual vector is $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y}$ and $SSE = \hat{\mathbf{e}}^\intercal\hat{\mathbf{e}} = \mathbf{y}^\intercal(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y}$. Note that

$$(\mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{X}^\intercal)(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{P}_\mathbf{X}) = \mathbf{0}$$

as $\mathbf{X}^\intercal(\mathbf{I} - \mathbf{P}_\mathbf{X}) = [(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{X}]^\intercal = \mathbf{0}$. Hence, by Theorem 6.24, $\mathbf{K}^\intercal\hat{\mathbf{b}}$ and SSE are independent. By definition of (noncentral) F distribution,

$$F = \frac{(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})^\intercal\mathbf{H}^{-1}(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})/s}{SSE/(N-r)} \sim F_{s,N-r}\left(\frac{1}{2}(\mathbf{K}^\intercal\mathbf{b} - \mathbf{c})^\intercal(\sigma^2\mathbf{H})^{-1}(\mathbf{K}^\intercal\mathbf{b} - \mathbf{c})\right).$$

Under the null hypothesis $H_0 : \mathbf{K}^\intercal\mathbf{b} = \mathbf{c}$, the noncentrality parameter is zero and therefore we end up having a central F distribution. Under the alternative hypothesis $H_1 : \mathbf{K}^\intercal\mathbf{b} \neq \mathbf{c}$, we have a noncentral F distribution. By Theorem 6.20, we expect a larger value for F when noncentrality parameter increases. Therefore, one reasonable idea is to reject H_0 if F is too large. It corresponds to a rejection rule for a test with significance level α:

$$\text{Reject } H_0 \text{ if } F = \frac{(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})^\intercal\mathbf{H}^{-1}(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})/s}{SSE/(N-r)} > F_{s,N-r,\alpha},$$

where $\alpha = \mathsf{P}(f > F_{s,N-r,\alpha})$ with $f \sim F_{s,N-r}$.

This F test that we have derived enjoys an invariance property. Say, we want to test $H_0 : \mathbf{K}^\intercal\mathbf{b} = \mathbf{c}$ with a full-column-rank $\mathbf{K}$ and the elements of $\mathbf{K}^\intercal\mathbf{b}$ being estimable. This hypothesis can also be represented by

$H_0' : 2\mathbf{K}^\intercal\mathbf{b} = 2\mathbf{c}$. The F test would be undesirable if the corresponding tests derived from testing H_0 and H_0' are different.

Example 3

In Example 2, if we want to test the equality of all group means. There are different options for H_0. For examples,

$$H_0 : \alpha_1 - \alpha_2 = 0, \quad \alpha_2 - \alpha_3 = 0 \quad \left(\mathbf{K}_1^\intercal = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \mathbf{c}_1 = \mathbf{0}\right)$$
$$H_0' : \alpha_1 - \alpha_2 = 0, \quad \alpha_1 - \alpha_3 = 0 \quad \left(\mathbf{K}_2^\intercal = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \mathbf{c}_2 = \mathbf{0}\right)$$

if $\mathbf{b} = [\mu, \alpha_1, \alpha_2, \alpha_3]^\intercal$.

We want the F test to be invariant under certain "equivalent" changes in $\mathbf{K}$ and $\mathbf{c}$. More precisely, suppose we have two different expressions of the test, $\mathbf{K}^\intercal\mathbf{b} = \mathbf{c}$ and $\mathbf{K}_*^\intercal\mathbf{b} = \mathbf{c}_*$ such that

$$\mathcal{S} = \{\mathbf{b} : \mathbf{K}^\intercal\mathbf{b} = \mathbf{c}\} = \mathcal{S}_* = \{\mathbf{b} : \mathbf{K}_*^\intercal\mathbf{b} = \mathbf{c}_*\},$$

where $\mathbf{K} \in \mathbb{R}^{p\times s}$ and $\mathbf{K} \in \mathbb{R}^{p\times s^*}$ both have full column rank. We want to show that they both give the same numerator in the test statistic of the F test.

dim S = dim S^*.

Lemma 7.6. *Let $\mathcal{S} = \{\mathbf{b} : \mathbf{K}^\intercal\mathbf{b} = \mathbf{c}\}$ and $\mathcal{S}_* = \{\mathbf{b} : \mathbf{K}_*^\intercal\mathbf{b} = \mathbf{c}_*\}$ where $\mathbf{K} \in \mathbb{R}^{p\times s}$ and $\mathbf{K} \in \mathbb{R}^{p\times s^*}$ both have full column rank. The two spaces $\mathcal{S} = \mathcal{S}_*$ if and only if $s = s_*$ and there exists a non-singular matrix $\mathbf{Q}$ such that $\mathbf{K}_* = \mathbf{K}\mathbf{Q}^\intercal$ and $\mathbf{c}_* = \mathbf{Q}\mathbf{c}$.*

By this lemma, if $\mathcal{S} = \mathcal{S}_*$, then the numerator in the test statistic of the F test for testing $\mathbf{K}_*^\intercal\mathbf{b} = \mathbf{c}_*$ is

$$\begin{aligned}(\mathbf{K}_*^\intercal\hat{\mathbf{b}} - \mathbf{c}_*)^\intercal(\mathbf{K}_*^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K}_*)^{-1}(\mathbf{K}_*^\intercal\hat{\mathbf{b}} - \mathbf{c}_*)/s_* &= (\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})^\intercal\mathbf{Q}^\intercal(\mathbf{Q}\mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K}\mathbf{Q}^\intercal)^{-1}\mathbf{Q}(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})/s \\ &= (\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})^\intercal(\mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K})^{-1}(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})/s,\end{aligned}$$

which implies that the F tests are equivalent.

4.2 The likelihood ratio test

In this section, we derive the likelihood ratio test. Unlike the above derivation which is motivated specifically from the distributional theory of linear model, the likelihood ratio test is derived from a general idea that could be utilized in other scenarios.

Again, we assume the normal Gauss-Markov model: $\mathbf{y} \sim N_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$, and we want to test the hypothesis $H_0 : \mathbf{K}^\intercal\mathbf{b} = \mathbf{c}$ with the elements of $\mathbf{K}^\intercal\mathbf{b}$ being estimable. Let

K_*b C_*

$$\Omega = \{(\mathbf{b}, \sigma^2) : \mathbf{b} \in \mathbb{R}^p, \sigma^2 > 0\} \quad \text{and} \quad \Omega_0 = \{(\mathbf{b}, \sigma^2) : \mathbf{K}^\intercal\mathbf{b} = \mathbf{c}, \sigma^2 > 0\}$$

be the (unrestricted) parameter space and the parameter space restricted by the null hypothesis H_0 respectively. Recall that the likelihood function is

$$L(\mathbf{b}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}Q(\mathbf{b})\right\}.$$

The likelihood ratio test statistic is

$\Omega_0 = \Omega_0'$

$$\phi(\mathbf{y}) = \frac{\max_{(\mathbf{b},\sigma^2)\in\Omega_0} L(\mathbf{b}, \sigma^2)}{\max_{(\mathbf{b},\sigma^2)\in\Omega} L(\mathbf{b}, \sigma^2)}.$$

When H_0 is incorrect, one would expect that the restriction of the parameter space could lead to a small $\phi(\mathbf{y})$. Therefore, the test rejects H_0 when $\phi(\mathbf{y})$ is small, i.e. $\phi(\mathbf{y}) < c$. The question is how to find a c such that the test has the desired type I error.

To understand $\phi(\mathbf{y})$, we should derive the maximized likelihoods in both the denominator and numerator in the formulation of $\phi(y)$. Similarly as before, we resort to differentiation. When $\mathbf{b}$ is fixed, we take the derivative with respect to σ^2 and set it zero

$$-\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} Q(\mathbf{b}) = 0 \quad \Rightarrow \sigma^2 = Q(\mathbf{b})/N$$

This results in a profile likelihood of $\mathbf{b}$:

$b \quad \sigma^2$

$$L(b) = \quad L(\mathbf{b}, Q(\mathbf{b})/N) = (2\pi/N)^{-N/2} Q(\mathbf{b})^{-N/2} \exp\left\{-\frac{N}{2}\right\}.$$

To maximize the profile likelihood, we have to minimize $Q(\mathbf{b})$. Let $(\hat{\mathbf{b}}_H)$ denote the value of $\mathbf{b}$ that minimizes $Q(\mathbf{b})$ over Ω_0. Therefore

$$\phi(\mathbf{y}) = \frac{L(\hat{\mathbf{b}}_H, Q(\hat{\mathbf{b}}_H)/N)}{L(\hat{\mathbf{b}}, Q(\hat{\mathbf{b}})/N)} = \frac{Q(\hat{\mathbf{b}}_H)^{-N/2}}{Q(\hat{\mathbf{b}})^{-N/2}}$$

Hence

$>c.$

$$\begin{aligned} \phi(\mathbf{y}) < c &\iff Q(\hat{\mathbf{b}}_H)/Q(\hat{\mathbf{b}}) > c^{-2/N} \\ &\iff (Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}))/Q(\hat{\mathbf{b}}) > c^{-2/N} - 1 \qquad F_2. \\ &\iff \boxed{\frac{(Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}))/s}{Q(\hat{\mathbf{b}})/(N-r)}} > \frac{N-r}{s}(c^{-2/N} - 1) \end{aligned}$$

$Q(\hat{b}) = \|y - X\hat{b}\|^2 = y^T(I-P_X)y.$

Write $\tilde{c} = s(c^{-2/N} - 1)/s$. Note that the denominator is exactly the same as the denominator of the test statistic F in F test. If the numerator is the same as the numerator of F, we can show that the two tests are equivalent. The following theorem establishes this result.

Theorem 7.7. *Assume a normal Gauss-Markov model. Suppose the elements of $\mathbf{K}^\intercal\mathbf{b}$ are estimable, and $\mathbf{K}$ has full column rank. Let $\hat{\mathbf{b}}_H$ be the part of a solution to the restricted normal equations with constraint $\mathbf{K}^\intercal\mathbf{b} = \mathbf{c}$. Then*

$$Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}) = (\hat{\mathbf{b}}_H - \hat{\mathbf{b}})^\intercal \mathbf{X}^\intercal\mathbf{X}(\hat{\mathbf{b}}_H - \hat{\mathbf{b}}) = (\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})^\intercal[\mathbf{K}^\intercal(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{K}]^{-1}(\mathbf{K}^\intercal\hat{\mathbf{b}} - \mathbf{c})$$

where $\hat{\mathbf{b}}$ is a solution to the normal equations.

Example 4

The following is one of the most common testing situations. Under the normal Gauss-Markov model $\mathbf{y} \sim N_N(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I})$, we want to test $H_0 : \mathbf{b}_1 = \mathbf{0}$, where

$$\mathbf{X}\mathbf{b} = \begin{bmatrix}\mathbf{X}_0 & \mathbf{X}_1\end{bmatrix}\begin{bmatrix}\mathbf{b}_0 \\ \mathbf{b}_1\end{bmatrix}$$

with $\mathbf{X} \in \mathbb{R}^{N\times p}$, $\mathbf{X}_1 \in \mathbb{R}^{N\times s}$, $\text{rank}(\mathbf{X}) = r$, $\text{rank}((\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1) = s$. We can write

$$\mathbf{K} = \begin{bmatrix}\mathbf{0}_{(p-s)\times s} \\ \mathbf{I}_s\end{bmatrix}, \quad \mathbf{c} = \mathbf{0}.$$

Clearly $\mathbf{K}$ has linearly independent columns. Next, we show that each element of $\mathbf{K}^\intercal\mathbf{b}$ is estimable.

$$\mathcal{C}(\mathbf{X}^\intercal) = \mathcal{C}\left(\begin{bmatrix}\mathbf{X}_0^\intercal \\ \mathbf{X}_1^\intercal\end{bmatrix}\right) \supseteq \mathcal{C}\left(\begin{bmatrix}\mathbf{X}_0^\intercal \\ \mathbf{X}_1^\intercal\end{bmatrix}(\mathbf{I}-\mathbf{P}_{\mathbf{X}_0})\right) = \mathcal{C}\left(\begin{bmatrix}\mathbf{0}_{(p-s)\times N} \\ [(\mathbf{I}-\mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1]^\intercal\end{bmatrix}\right) = \mathcal{C}\left(\begin{bmatrix}\mathbf{0}_{(p-s)\times s} \\ \mathbf{I}_s\end{bmatrix}\right),$$

where last equality is due to that $[(\mathbf{I}-\mathbf{P}_{\mathbf{X}_0})\mathbf{X}_1]^\intercal$ has only s rows and is of rank s. So all columns of $\mathbf{K}$ belong to $\mathcal{C}(\mathbf{X}^\intercal)$ and so all elements of $\mathbf{K}^\intercal\mathbf{b}$ are estimable.

In this situation, we have

$$SSE(reduced) = \mathbf{y}^\intercal(\mathbf{I}-\mathbf{P}_{\mathbf{X}_0})\mathbf{y} = Q(\hat{\mathbf{b}}_H)$$
$$SSE = \mathbf{y}^\intercal(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{y} = Q(\hat{\mathbf{b}})$$

Therefore $Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}) = SSE(reduced) - SSE = \mathbf{y}^\intercal(\mathbf{P}_{\mathbf{X}}-\mathbf{P}_{\mathbf{X}_0})\mathbf{y}$. Hence

$$F = \frac{(Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}}))/s}{Q(\hat{\mathbf{b}})/(N-r)} = \frac{(SSE(reduced) - SSE)/s}{SSE/(N-r)}$$

is the test statistic in the F test due to Theorem 7.7, and so this follows $F_{s,N-r}$ under H_0.

Instead, the distribution of F can also be derived directly from Cochran's Theorem: From Cochran's Theorem, we know that $\mathbf{y}^\intercal\mathbf{P}_{\mathbf{X}_0}\mathbf{y}$, $\mathbf{y}^\intercal(\mathbf{P}_{\mathbf{X}}-\mathbf{P}_{\mathbf{X}_0})\mathbf{y}$ and $\mathbf{y}^\intercal(\mathbf{I}-\mathbf{P}_{\mathbf{X}_0})\mathbf{y}$ are independent with noncentral chi-square distributions, with noncentralities $(\mathbf{X}\mathbf{b})^\intercal\mathbf{P}_{\mathbf{X}_0}(\mathbf{X}\mathbf{b})/(2\sigma^2)$ $(\mathbf{X}\mathbf{b})^\intercal(\mathbf{P}_{\mathbf{X}}-\mathbf{P}_{\mathbf{X}_0})(\mathbf{X}\mathbf{b})/(2\sigma^2)$ and 0. Hence

$$F \sim F_{s,N-r}$$

under H_0.

4.3 Computing the test statistic via the ANOVA table

Example 5 (Test of homogeneity.)

Consider a one-way ANOVA model,

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, a \quad j = 1, \ldots, n_i$$

and the hypothesis,

$$H_0 : \alpha_1 = \cdots = \alpha_a \,. \tag{7.1}$$

This null hypothesis can be written in the form of general linear hypothesis $\mathbf{K}^\intercal\mathbf{b} = \mathbf{0}$, where $\mathbf{b} = [\mu, \alpha_1, \ldots, \alpha_a]^\intercal$. However, there are more than one $\mathbf{K}$ that can express H_0. For examples,

$$\mathbf{K}_1^\intercal = \begin{bmatrix} 0 & 1 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & \ddots & & \vdots \\ 0 & \cdots & \cdots & & & 1 & -1 \end{bmatrix}. \Rightarrow$$

$\alpha_1 = \alpha_2$

$\alpha_2 = \alpha_3$

and

$$\mathbf{K}_2^\intercal = \begin{bmatrix} 0 & 1 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \ddots & & \vdots \\ 0 & 1 & 0 & & \cdots & 0 & -1 \end{bmatrix}.$$

$\alpha_1 = \alpha_2$

$\alpha_1 = \alpha_3$

$\alpha_1 = \alpha_4$

Both $\mathbf{K}_1^\intercal\mathbf{b} = \mathbf{0}$ and $\mathbf{K}_2^\intercal\mathbf{b} = \mathbf{0}$ lead to (7.1). Furthermore, both $\mathbf{K}_1$ and $\mathbf{K}_2$ imply that H_0 is testable, because, as shown in Example 1 of Notes 4 that $\sum_i d_i\alpha_i$ is estimable if $\sum_i d_i = 0$.

$1-1$

Definition 7.8 (Contrasts). *Let* $\mathbf{s} = [s_1, \ldots, s_k]^\intercal$ *be a set of variables, either parameters or statistics, and let* $\mathbf{a} = (a_1, \ldots, a_k)^\intercal$ *be known constants. The linear function* $\sum_{i=1}^k a_i s_i$ *is called a contrast if* $\sum_{i=1}^k a_i = 0$.

Contrast hypotheses are important because they can be used to compare parameters such as comparing treatment means in Example 5.

The form of F test statistic inspired by the likelihood ratio test (due to Theorem 7.7) for testing general linear hypotheses is constructed based on the function of the sum of squared residuals from the restricted (or reduced) model under the null H_0 and that of the unrestricted (or full) model.

For the one-way ANOVA model, both the SSE under the reduced and full models can often be easily obtained from the ANOVA table because it is constructed by fitting submodels sequentially.

Example 6 (Continuation of Example 5)

$X = [1_n \quad 1_{n_1} \quad 1_{n_2} \quad \cdots \quad 1_{n_a}]$

X_1 $\quad$ X_2

The one-way ANOVA table is given by Example 6 of Notes 6:

$y^T y$

$y^T(P_X - P_1)y = \sum n_i(\bar{y}_i - \bar{y})^2$

Source	df	Projection	Sum of squares	Noncentrality
Mean	1	$\mathbf{P_1}$	$N\bar{y}^2$	$\frac{\boldsymbol{\eta}^\intercal \mathbf{P_1} \boldsymbol{\eta}}{2\sigma^2}$
Group	$a-1$	$\mathbf{P_X} - \mathbf{P_1}$	$\sum_{i=1}^a n_i \bar{y}_{i.}^2 - N\bar{y}^2$	$\frac{\boldsymbol{\eta}^\intercal (\mathbf{P_X} - \mathbf{P_1}) \boldsymbol{\eta}}{2\sigma^2}$
Error	$N-a$	$\mathbf{I} - \mathbf{P_X}$	SSE	0

The sum of squared residuals under the full model is fit $\mu, \alpha_1 \cdots \alpha_a$

$$Q(\mathbf{b})/\sigma^2 = \mathbf{y}^\intercal(\mathbf{I} - \mathbf{P_X})\mathbf{y}/\sigma^2 = \text{SSE}/\sigma^2 \sim \chi^2_{N-a}$$

full $y = \mu X_1 + X_2\alpha + \varepsilon$

Under the null hypothesis (7.1), the reduced model is $(\mu-1)+(\alpha+1)$

$$y_{ij} = \mu + \alpha + \varepsilon_{ij} = \tilde{\mu} + \varepsilon_{ij}\,.$$

reduced $y = \mu X_1 + \varepsilon$

That is, the reduced model has only one parameter, the overall mean, and so the design matrix could be regarded as $\mathbf{1}$. Thus the sum of squared residuals after fitting the overall mean is $Q(\hat{\mathbf{b}}_H) = \mathbf{y}^\intercal(\mathbf{I} - \mathbf{P_1})\mathbf{y}$. We get

$$SS_{\boldsymbol{\alpha}} := \left\{Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}})\right\}/\sigma^2 = \mathbf{y}^\intercal(\mathbf{P_X} - \mathbf{P_1})\mathbf{y}/\sigma^2 \sim \chi^2_{a-1}(v)\,,$$

where the noncentrality parameter

$$v = \frac{1}{2\sigma^2}\sum_{i=1}^a n_i(\alpha_i - \bar{\alpha})^2\,, \quad \bar{\alpha} = \frac{1}{n}\sum_{k=1}^a n_k\alpha_k$$

and thus

$$\phi(\mathbf{y}) = \frac{\{Q(\hat{\mathbf{b}}_H) - Q(\hat{\mathbf{b}})\}/(a-1)}{Q(\hat{\mathbf{b}})/(N-a)} = \frac{\text{SS}_{\boldsymbol{\alpha}}/(a-1)}{SSE/(N-a)} \sim F_{a-1,N-a}(v).$$

The noncentrality $v = 0$ under the null hypothesis in (7.1).

$$Xb = \begin{bmatrix} \mu + \alpha_1 \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu + \alpha_2 \\ \mu \end{bmatrix}$$