Class Notes for Biostat 250AB

Academic year 2020-2021: the era of COVID-19 pandemic

Scribe: Elvis Cui, Instructor: Weng Kee Wong

May 3, 2021

Contents

Preface

This lecture notes is based on the 2 courses I took from 2020 to 2021 (Biostat 250A and 250B). From 2020 to 2021, all my classmates and colleagues took and taught classes online, which is, I believe, the **VERY FIRST TIME** in history ever. COVID-19 pandemic has changed a lot to our daily life, not only teaching and taking classes, but also working, communicating with other, etc. What stays unchanged is our passion to learn and study.

Note that x and $\mathbf{x}$ can both represent a vector and X and $\mathbf{X}$ can both represent a matrix. Also, I switch between y and Y from time to time. Both $\mathcal{N}$ and $\mathcal{N}_p$ can represent a multivariate distribution. I hope they do not cause too much confusion and I am too lazy to modify them so I apologize. References are given and some key results such as generalized Cauchy-Schwarz inequality and Cook-Weisberg score test are based them.

Finally, a quote from CR Rao,

$$\text{Statistics} = \text{Approximation} + \text{Optimization}$$

and another quote from George Box,

All models are wrong, some are useful.

Elvis Cui
April 5, 2021.

1 Useful Results in Matrix Theory

This section is not based on any single lecture but a collection of all results in matrix theory that I learned in Biostat 250AB and other related materials.

1.1 Diagonalization and g-Inverse

Theorem 1.1 (Simultaneous diagonalization theorem)**.** Let A, M be two symmmetric $n \times n$ matrices, M positive definite. Then $\exists\ C$ and $\det|C| \neq 0$ s.t.

$$C^T MC = I,\ C^T AC = \Lambda$$

where I is identity and Λ is diagonal.

Proof. $M \succ 0 \Rightarrow M = M^{1/2}M^{1/2}$ and let $B := (M^{-1/2})^T A(M^{-1/2})$.

Then $\exists\ P^T\ n' \times n$ orthogonal, i.e., $PP^T = I$ s.t. $PBP^T = \Lambda$ or $P^T M^{-1/2} AM^{-1/2} P = \Lambda$. The diagonal elements of Λ are eigenvalues of B. And let $C = M^{-1/2}P$, then we have $\det|C| \neq 0$ and

$$C^T AC = \Lambda,\ C^T MC = I$$

□

Theorem 1.2 (Existence of g-inverse)**.** Let X be a $n \times p$ matrix and rank$(X) = r < \min(n, p)$, then there exists infinitely many X^-, they are $p \times n$ matrices and satisfy

$$XX^-X = X$$

Proof. Let $X = SDT^T$ be the singular value decomposition and D is a $r \times r$ matrix. Define

$$X^- = TD^{-1}S^T$$

and we have $XX^-X = SDT^T TD^{-1}S^T SDT^T = SDT^T = X$. But there are infinitely many g-inverses:

$$\widetilde{X} = X^- + (I - X^-X)B$$

where B is any $p \times n$ matrix. We have

$$X\widetilde{X}X = XX^-X + X(I - X^-X)BX = X + XBX - XX^-XBX = X$$

□

Example 1.1 (Solution of linear systems)**.** Consider the linear system

$$Ax = b$$

A solution x^* exists iff it is a consistent set of equations, i.e.,

$$\text{rank}(A|b) = \text{rank}(A)$$

and

$$x^* = A^- b$$

is a solution.

Theorem 1.3 (Positive definite relation)**.** If $A \succ B$ and both A, B are positive definite, then we have

$$B^{-1} \succ A^{-1}$$

Proof. By theorem 1.1, there exists a non-singular matrix U such that

$$A^{-1} = U^T U,\ B^{-1} = U^T D^{-1} U$$

where D is diagonal with elements equal to eigenvalues of $A^{-1/2} B A^{-1/2}$. Then,

$$A - B = U^{-1}(I - D)U^{-T} \succ 0$$

suggests that $d_{ii} \leq 1$ for all diagonal elements of D. Thus,

$$A^{-1} - B^{-1} = U^T(I - D^{-1})U \prec 0$$

since $(I - D^{-1}) \prec 0$.

□

1.2 Partitioning

Theorem 1.4 (Determinant formula)**.** Let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},\ B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

Then:

1. If A_{21}=0, then

$$\left| \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right| = |A_{11}||A_{22}|$$

2. If A and B are square matrices of the same dimension, then

$$|I + AB| = |I + BA|$$

3. In the case $A_{21} \neq 0$, we have

$$\left|\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\right| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}|$$

Proof. I only prove 2.

$$\begin{pmatrix} I & A \\ -B & I \end{pmatrix}\begin{pmatrix} I & 0 \\ B & I \end{pmatrix} = \begin{pmatrix} I + AB & A \\ 0 & I \end{pmatrix}$$
$$\begin{pmatrix} I & 0 \\ B & I \end{pmatrix}\begin{pmatrix} I & A \\ -B & I \end{pmatrix} = \begin{pmatrix} I & A \\ 0 & I + BA \end{pmatrix}$$

But the determinants of LHS are identical.

□

Theorem 1.5 (Inverse of block matrices)**.** Let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},\ A^{-1} = B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

Then

$$\begin{aligned} B_{11} &= A_{11\cdot 2}^{-1} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} \\ B_{21} &= -A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1} \\ B_{12} &= -A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1} \\ B_{22} &= A_{22}^{-1} + A_{22}^{-1}A_{21}A_{11\cdot 2}^{-1}A_{12}A_{22}^{-1} \end{aligned}$$

Theorem 1.6 (Diagonal elements of leverage matrix, [1])**.** Let W be an $p \times p$ matrix of rank p with columns w_j, $j = 1, \cdots, p$. Then

$$\left(W^T W\right)_{jj}^{-1} = \left[w_j^T (I - P_{-j}) w_j\right]^{-1} \tag{1.1}$$

where P_{-j} is the projection operator associated with W_{-j}, i.e., W with its j^{th} column omitted.

Proof. By inverse of block matrices 1.5, we have

$$\begin{aligned}(W^TW)^{-1} &= \begin{pmatrix} W_{-p}^TW_{-p} & W_{-p}^Tw_p \\ w_p^TW_{-p} & w_p^Tw_p \end{pmatrix}^{-1} \\ &= \begin{pmatrix} F & g \\ g^T & h \end{pmatrix}\end{aligned}$$

where $h = (W^TW)^{-1}_{pp} = (w_p^Tw_p - w_pP_{-p}w_p)^{-1}$. Let Π be the permutation matrix I with its j^{th} and p^{th} columns interchanged. Then $\Pi^2 = I$, so Π is a symmetric orthogonal matrix, and its own inverse. Hence

$$\begin{aligned}(W^TW)^{-1} &= \Pi(\Pi W^TW\Pi)^{-1}\Pi \\ &= \Pi \begin{pmatrix} F_1 & g_1 \\ g_1^T & h_1 \end{pmatrix} \Pi\end{aligned}$$

where $h_1 = (W^TW)^{-1}_{jj}$. Thus w_p and w_j, W_{-p} and W_{-j} have been effectively interchanged. The result then follows. $\square$

1.3 Extreme values of quadratic forms

Theorem 1.7 (A mini-max theorem [2])**.** Let $A = A^T$ be a $n \times n$ symmetric matrix and the spectral decomposition is $A = P\Lambda P^T$. x is a $n \times 1$ vector and B is a $n \times k$ matrix. Then we have

$$\begin{aligned}\inf_B \sup_{B^Tx=0} \frac{x^TAx}{x^Tx} &= \lambda_{k+1} \\ \sup_B \inf_{B^Tx=0} \frac{x^TAx}{x^Tx} &= \lambda_{n-k}\end{aligned}$$

where λ_i is the i^{th} diagonal element of Λ.

Theorem 1.8 (Sums of quadratic forms [2])**.** Suppose A is a symmetric matrix and $x_1, \cdots, x_k$ are k orthonormal vectors, that is, $x_i^Tx_j = \delta_{ij}$ where δ_{ij} is the Kronecker delta. Then,

$$\sup_{x_1,\cdots,x_k} \sum_{i=1}^k x_i^TAx_i = \sum_{i=1}^k \lambda_i$$

where λ_i is the i^{th} largest eigenvalue of A. The equality is attained when x_i's are the corresponding eigenvectors to λ_i's.

Proof. Let $T\Lambda T^T$ be spectral decomposition of A and $y_i = T^Tx_i$. We have $y_i^Ty_j = \delta_{ij}$. WLOG, suppose y_i is p-dimension. Then LHS without $\sup$ can be written as

$$\sum_{i=1}^{k}\sum_{j=1}^{p}\lambda_j y_{ij}^2 = \sum_{j=1}^{p}\left(\sum_{i=1}^{k} y_{ij}^2\right)\lambda_j$$

But $\left(\sum_{i=1}^{k} y_{ij}^2\right) \leq 1 \ \forall j$ and $\sum_{j=1}^{p}\left(\sum_{i=1}^{k} y_{ij}^2\right) = k$ give us the desired result. □

Theorem 1.9 (Cauchy-Schwarz inequality [2]). Let $A = B^T B$ be a Gram matrix, then

1. $(x^T Ay)^2 \leq (x^T Ax)(y^T Ax)$
2. $(x^T y)^2 \leq (x^T Ax)(y^T A^{-1} y)$

Thus,

$$\sup_x \frac{(u^T x)^2}{x^T Ax} = u^T A^{-1} u$$

where A positive definite and $x = A^{-1}u$.

Proof. The proof is immediately obtained by letting

$$\widetilde{x} = A^{1/2}x,\ \widetilde{y} = A^{1/2}y$$

and

$$\widetilde{x} = A^{1/2}x,\ \widetilde{y} = A^{-1/2}y$$

□

1.4 Projection operator

Consider the classical setting $y = X\beta + \epsilon$ where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$. The projection operator P_X and Q_X are defined as

$$P_X = X(X^T X)^{-1}X^T,\ Q_X = I - P_X$$

In the general case when X is not invertible, then

$$P_X = X(X^T X)^{-}X^T$$

where $(X^T X)^{-}$ is the g-inverse. Moreover, if $\mathcal{V}$ is a vector space, then $P_{\mathcal{V}}$ represents the projection operator onto $\mathcal{V}$. Unless otherwise specified, I use P instead of P_X, Q_X for simplicity. An interesting definition of projection operator in **finite dimension spaces** equipped with special inner product is given below.

Definition 1.1 (Projection operator, [2]). Suppose $\mathbb{R}^n$ is equipped with the inner product

$$\langle x, y\rangle = y^T \Sigma x$$

where $x, y \in \mathbb{R}^n$ and Σ is a positive definite matrix. Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$ spanned by a square matrix P, then we say P is the projection operator on $\mathcal{V}$ if the following holds

1. P id idempotent:
$$P^2 = P$$

2. ΣP is symmetric:
$$(\Sigma P)^T = \Sigma P$$

   In the special case $\Sigma + I$, it is equivalent to say P is symmetric.

Theorem 1.10 (Adding regressors)**.** Let $X = [X_1\ X_2]$ be a $n \times p$ matrix and X_1 is $n \times p_1$ while X_2 is $n \times p_2$ such that $p = p_1 + p_2$. The projection operator onto X can be written as

$$P_X = P_{X_1} + Q_{X_1} X_2 \left(X_2^T Q_{X_1} X_2 \right)^{-1} X_2^T Q_{X_1}$$

Proof. For any vector $\mathbf{m} \in \mathbb{R}^n$, we have

$$\begin{aligned}
P_X \mathbf{m} &= P_X P_{X_1} \mathbf{m} + P_X Q_{X_1} \mathbf{m} \\
&= P_{X_1} \mathbf{m} + P_X Q_{X_1} \mathbf{m} \\
&=_{(*)} P_{X_1} \mathbf{m} + P_{\mathcal{C}(Q_{X_1}) \cap \mathcal{C}(X)} \mathbf{m} \\
&= P_{X_1} \mathbf{m} + P_{\mathcal{C}(Q_{X_1}) \cap [\mathcal{C}(X_1) \cup \mathcal{C}(X_2)]} \mathbf{m} \\
&= P_{X_1} \mathbf{m} + P_{\mathcal{C}(Q_{X_1}) \cap \mathcal{C}(X_2)} \mathbf{m}
\end{aligned}$$

$(*)$ is because $\mathcal{C}(X) = \mathcal{C}(P_X) = \mathcal{C}(P_{X_1}) + \mathcal{C}(Q_{X_1}) \cap \mathcal{C}(X_2)$. And the result follows from

$$P_{\mathcal{C}(Q_{X_1}) \cap \mathcal{C}(X_2)} = Q_{X_1} X_2 \left(X_2^T Q_{X_1} X_2 \right)^{-1} X_2^T Q_{X_1}$$

or equivalently,
$$\mathcal{C}(Q_{X_1}) \cap \mathcal{C}(X_2) = \mathcal{C}\left(P_{(Q_{X_1} X_2)} \right)$$

$\square$

Example 1.2 (Gram-Schmidt process)**.** Let v, u be two p-dimensional vectors, then the projection operator associated with u is
$$P_u = \frac{uu^T}{\|u\|_2^2}$$
The projection of v onto u is
$$\text{proj}_u(v) = P_u(v) = \frac{\langle u, v \rangle}{\|u\|_2^2} u$$
Suppose we have k linearly independent vectors $v_1, \cdots, v_k$ and we want to orthogonalize them so

that they form an orthonormal basis of $\mathbb{R}^k$. Then the **Gram-Schmidt process** works as follows

$$
\begin{aligned}
u_1 &= v_1 & e_1 &= \frac{u_1}{\|u_1\|_2} \\
u_2 &= v_2 - \text{proj}_{u_1}(v_2) & e_2 &= \frac{u_2}{\|u_2\|_2} \\
u_3 &= v_3 - \text{proj}_{u_1}(v_3) - \text{proj}_{u_2}(v_3) & e_3 &= \frac{u_3}{\|u_3\|_2} \\
&\cdots & &\cdots \\
u_k &= v_k - \sum_{j=1}^{k-1} \text{proj}_{u_j}(v_k) & e_k &= \frac{u_k}{\|u_k\|_2}
\end{aligned}
$$

Finally, $\{u_1, \cdots, u_k\}$ forms an orthonormal basis of $\mathbb{R}^k$.

1.5 Exercises

1. (This formula comes from Hua Zhou's Biostat 257 HW1 for 2020 Spring, question 5) Let U and V be 2 matrices such that UV^T has the same dimension as A, show that

$$|A + UV^T| = |A||I + V^T A^{-1} U|$$

   This formula is useful for evaluating the density of a multivariate normal with covariance matrix $A + UV^T$.

2. (Binomial inversion theorem) Show that if A, B, C and D are conformable matrices, and all indicated inverses exist,

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B\left(C^{-1} + DA^{-1}B\right)^{-1} DA^{-1}$$

   Hence or otherwise, show Sherman-Morrison formula

$$(A + ab^T)^{-1} = A^{-1} - \frac{A^{-1}ab^T A^{-1}}{1 + b^T A^{-1} a}$$

   and Woodbury identity

$$(A + UV)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1} U)^{-1} V^T A^{-1}$$

   This is useful when solving the Henderson's equation in linear mixed models.

3. Show the following:

   (a) If $A \in \mathbb{R}^{n \times m}$, then rank$(A) \leq \min(m, n)$.

   (b) rank$(AB) \leq \min($rank(A), rank$(B))$.

(c) rank$(A+B) \leq$ rank$(A)+$rank(B).

(d) If A, B are invertible, then rank$(C) =$ rank$(AC) =$ rank$(CB) =$ rank(ACB).

(e) If rank$(B) =$ rank(C), then $\mathcal{C}(ACB) = \mathcal{C}(AC)$.

(f) If $A = ABA$, then rank$(A) =$ rank$(AB) =$ rank$(BA) \leq$ rank(B).

4. (Key lemma) Show that

$$\text{rank}(A^T) = \text{rank}(A) = \text{rank}(AA^T) = \text{rank}(A^TA) \tag{1.2}$$

5. (Linear basis) If rank$(A) = r$, then show that there exists B and C such that

$$A = BC$$

where B has full column rank and C has full row rank.

6. (Eigenvalues and eigenvectors of symmetric matrices) The eigenvalues and eigenvectors of a matrix A can be obtained solving

$$Ax = (a+bi)x$$

where $x \neq 0$ and a, b are real numbers and i is an indeterminate satisfying $i^2 = -1$. Show that $b = 0$ if A is symmetric. **Hint**: Let $Z = (A - (a+bi)I)$ and define

$$B = C^T\bar{C}$$

Then use the fact $|B| = 0$. Next, show x's corresponding to distinct eigenvalues are orthogonal to each other.

7. Show that any matrix A can be written as a sum of a symmetric matrix and a skew matrix, and these summands are unique.

8. Prove that if $x^TAx = 0$ for all x and $A = A^T$, then $A = 0$.

9. Use the fact $\mathcal{N}(A^T) = \mathcal{C}(A)^\perp$ to show if

$$PA^TA = QA^TA$$

then $PA^T = QA^T$ for any conformable matrices P and Q.

10. (Simultaneous diagonalization) Let A and B be two $n \times n$ positive definite matrices. If $A - B \succ 0$, show

(a) Tr$(A) >$ Tr(B).

(b) $B^{-1} \succ A^{-1}$, this is the theorem we have proved.

(c) $\det|A| > \det|B|$.

11. Show that the sets of non-zero eigenvalues for AB and AB are the same for any two conformable matrices.

12. Let A be a symmetric matrix and $B = CA^{-1}C$, describe the relationship between eigenvalues and eigenvectors of B and A.

13. Let $\mathbf{1}_n$ be the $n \times 1$ vector with all entries equal to 1 and let B be a $n \times n$ matrix given by

$$B = \frac{2b}{2b-1} I_n - \frac{\mathbf{1}_n \mathbf{1}_n^T}{(2n-1)}$$

and $b > 1/2$.

(a) Is B always positive definite?

(b) Derive Tr(B).

(c) Show that the maximum eigenvalue of B is $\frac{2b}{2b-1}$.

14. (Householder transformation) Let v be a nonzero vector. The **Householder transformation matrix** is defined by

$$H_v = I - \frac{2vv^T}{v^T v}$$

(a) Find the determinant and all eigenvalues of such a matrix.

(b) Show that if $x \neq 0$, then there is a Householder matrix such that

$$Hx = \|x\|_2 e_1$$

where $e_1^T = (1, 0, 0, \cdots, 0)$.

15. (Norm-preserving mapping) Show that $\|x\|_2 = \|y\|_2$ if and only if there is an orthogonal matrix T such that $Tx = y$.

16. (Canonical correlation) Suppose y is a univariate random vector with variance a^2, and X is a $p \times 1$ random vector with covariance matrix V and

$$Cov(y, X) = W$$

where W is a p-dimensional vector. Use the generalized Cauchy-Schwarz inequality 1.9 to answer the following.

(a) If b is any nonzero vector, what is the maximum correlation of $b^T X$ with y?

(b) What choice of b will ensure that the maximum is attained?

2 Distribution Theory

This section is mostly based on the first half part of Biostat 250A where we discussed in depth the non-central χ^2 and t distribution with applications to linear models. Moment generating functions and characteristic functions are also included.

2.1 Multivariate normal distribution

In this subsection, we discuss basic properties of multivariate normal distribution.

Definition 2.1 (Multivariate normal)**.** A p-dimensional random vector Y is said to have a **multivariate normal distribution** with parameters (μ, Σ) iff its moment generating function is

$$\Psi_Y(t) = \mathbb{E}e^{t^T y} = \exp\left(\mu^T t + \frac{1}{2}t^T \Sigma t\right) \tag{2.1}$$

We write

$$Y \sim \mathcal{N}_p(\mu, \Sigma)$$

and its density, if exists, is

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right)$$

Lemma 1 (Alternative definition)**.** *By Cramer-Wold device, we have*

$$Y \sim \mathcal{N}_p(\mu, \Sigma) \;\; \textit{iff} \;\; a^T y \sim \mathcal{N}_1(a^T\mu, a^T\Sigma a) \; \forall a$$

Lemma 2 (Linear transformation)**.** *Let $Y \sim \mathcal{N}(\mu, \Sigma)$, then*

$$AY \perp\!\!\!\perp BY$$

if and only if $A\Sigma B = 0$ where A and B are two matrices of proper dimension.

Example 2.1 (Conditional multivariate normal distribution)**.** Suppose

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}_p\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

Set

$$\begin{pmatrix} Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

Then we have

$$Cov(Y_3, Y_4) = \begin{pmatrix} \Sigma_{11\cdot 2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

where $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Thus, $Y_3 \perp\!\!\!\perp Y_2$ and we have

$$Y_1|Y_2 = y_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11\cdot 2}\right) \tag{2.2}$$

2.2 Non-central distribution

In this subsection, we discuss non-central t, χ^2 and F distribution and their statistical properties. Some results are from [2] and the derivation of the density of non-central χ^2 distribution is from Dr. Wong's 250B notes.

Definition 2.2 (Non-central student t)**.** Let $Y \sim \mathcal{N}(\mu, \sigma^2)$ be a univariate normal variable and $\frac{X}{\sigma^2} \sim \chi_k^2(0)$ where $\chi_k^2(0)$ is the usual central χ^2 distribution. Then we say

$$T = \frac{Y}{\sqrt{X/k}}$$

follows a **non-central student** t distribution with k degrees of freedom and non-centrality parameter $\delta = \mu/\sigma$ and write $T \sim t_k(\delta)$. In the special case when $Y \sim \mathcal{N}(\mu, 1)$, we have $\delta = \mu$ and write $T \sim t_k(\mu)$. The density of T is

$$f_t(t) = \frac{k^{\frac{k}{2}}}{\Gamma(\frac{k}{2})} \frac{\exp(-\frac{\delta^2}{2})}{(k+t^2)^{\frac{k+1}{2}}} \sum_{s=0}^{\infty} \left(\Gamma(\frac{k+s+1}{2})(\frac{\delta^s}{s!})(\frac{2t^2}{k+t^2})^{\frac{s}{2}} \right)$$

Note there is a multivariate version of t distribution, see definition 7.3.

Definition 2.3 (Non-central χ^2)**.** Let $X_i \sim \mathcal{N}(\mu_i, 1), i = 1, 2, \cdots, k$ and $X_i \perp\!\!\!\perp X_j$ for $i \neq j$. Define

$$\delta^2 = \sum_{i=1}^{n} \mu_i^2 = \|\mu\|_2^2$$

Then we say

$$Y = \sum_{i=1}^{n} X_i^2$$

has a non-central χ^2 distribution with n degrees of freedom and non-centrality parameter δ^2 and write

$$Y \sim \chi_n^2(\delta^2)$$

Example 2.2 (Density of non-central χ^2)**.** Suppose $X \sim \mathcal{N}_n(\mu, I)$ where $\mu^T = (\mu_1, \cdots, \mu_n)$. Let A be an orthogonal matrix derived from Gram-Schmidt process with the first row equals to $\frac{\mu^T}{\|\mu\|_2}$. Define

$$W = AX, \; Y = X^T X$$

Then

$$W \sim \mathcal{N}_n\left(\begin{pmatrix}\|\mu\| \\ 0 \\ \dots \\ 0\end{pmatrix}, I\right)$$

and

$$\begin{aligned}
Y &= X^T X \\
&= X^T A^T A X \\
&= W^T W \\
&= \underbrace{w_1^2}_{\chi_1^2(\|\mu\|_2^2)} + \underbrace{\sum_{i=2}^{n} w_i^2}_{\chi_{n-1}^2(0)}
\end{aligned}$$

Thus, we can re-write a $\chi_n^2(\delta^2)$ variable as

$$\begin{aligned}
\chi_n^2(\delta^2) &= \chi_1^2(\delta^2) + \chi_{n-1}^2(0) \\
&= V + U
\end{aligned} \tag{2.3}$$

and also $V \perp\!\!\!\perp U$. Let $Z \sim \mathcal{N}(\delta^2, 1)$, then by definition,

$$\begin{aligned}
F_V(v) = \mathbb{P}(V \leq v) &= \mathbb{P}(-\sqrt{v} \leq Z \leq \sqrt{v}) \\
&= \int_{-\sqrt{v}}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\delta^2)^2} dz
\end{aligned}$$

Thus, the density of V is

$$\begin{aligned}
f_V(v) &= \frac{d}{dv} F_V(v) \\
&= \frac{1}{\sqrt{2\pi}} \left(\frac{e^{-\frac{1}{2}(\sqrt{v}-\delta^2)^2}}{2\sqrt{v}} + \frac{e^{-\frac{1}{2}(-\sqrt{v}-\delta^2)^2}}{2\sqrt{v}} \right) \\
&= \frac{1}{2\sqrt{2\pi v}} e^{-\frac{1}{2}(v+\delta^4)} \left(e^{\delta^2\sqrt{v}} + e^{-\delta^2\sqrt{v}} \right) \\
&= \frac{1}{2\sqrt{2\pi v}} e^{-\frac{1}{2}(v+\delta^4)} \left\{ \sum_{i=0}^{\infty} \frac{(\delta^2\sqrt{v})^{2i}}{(2i)!} \right\}
\end{aligned} \tag{2.4}$$

and $\frac{(\delta^2\sqrt{v})^{2i}}{(2i)!}$'s are derived from Taylor's expansion of e^x. The joint density of V and U is

$$f_{V,U}(v, u) = f_V(v) f_U(u)$$

with

$$f_U(u) = \frac{e^{-\frac{u}{2}} u^{\frac{n-3}{2}}}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})}$$

Next, let $Z_1 = U + V$ and $Z_2 = \frac{U}{U+V}$ or equivalently, $U = Z_1 Z_2$ and $V = Z_1(1 - Z_2)$. The Jacobian for this transformation is given by

$$J = \left\| \begin{pmatrix} Z_2 & Z_1 \\ 1 - Z_2 & -Z_1 \end{pmatrix} \right\| = Z_1$$

So,

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &= f_V(z_1(1 - z_2)) f_U(z_1 z_2) \times z_1 \\ &= \text{const.} \times \exp(-\frac{1}{2}(z_1 + \delta^4)) z_1^{\frac{n-2}{2}} (1 - z_2)^{-\frac{1}{2}} \left\{ \sum_{i=0}^{\infty} (1 - z_2)^i \frac{\delta^{4i} z^i}{(2i)!} \right\} \end{aligned}$$

Since $\int_0^1 (1-x)^{i-\frac{1}{2}} = (i + \frac{1}{2})!$, thus,

$$\begin{aligned} f_{Z_1}(z) &= \int_0^1 f_{Z_1,Z_2}(z, z_2) d_{z_2} \\ &= \text{const.} \times e^{-\frac{1}{2}(z+\delta^4)} z^{\frac{n-2}{2}} \left\{ \sum_{i=0}^{\infty} \frac{\delta^{4i} z^i}{(2i)!(i + \frac{1}{2})!} \right\} \end{aligned}$$

Recall $\int_0^\infty x^{a-1} e^{-bx} dx = \Gamma(a)/b^a$, so we get

$$f_{Z_1}(z) = \sum_{i=0}^{\infty} \underbrace{\frac{e^{-\lambda/2} (\frac{\lambda}{2})^i}{i!}}_{q_i} \underbrace{\left(\frac{1}{\Gamma(\frac{n+2i}{2})} (\frac{1}{2})^{\frac{n+2i}{2}} z^{\frac{n+2i}{2}-1} e^{-\frac{z}{2}} \right)}_{\text{density of a } \chi^2_{n+2i}(0)} \tag{2.5}$$

In conclusion, we can re-write the density of $Y \sim \chi^2_n(\delta^2)$ as

$$f_Y(y) = \sum_{i=0}^{\infty} q_i f_{\chi^2_{2n+i}(0)}(y) \tag{2.6}$$

and q_i's are called **Poisson weights**.

Definition 2.4 (Non-central F)**.** Let $X \sim \chi^2_n(\delta^2)$ and $Y \sim \chi^2_m(0)$ and assume $X \perp\!\!\!\perp Y$, then

$$F = \frac{X/n}{Y/m}$$

is said to have a non-central F distribution with parameter (δ^2, n, m) and we write

$$F \sim F_{n,m}(\delta^2)$$

2.3 Basic results for quadratic forms

Suppose $y \sim \mathcal{N}_p(0, I)$ and let

$$Q = y^T A y$$

where $A^T = A$. What is the distribution of Q?

Theorem 2.1 (Fundamental theorem)**.** WLOG, let A be symmetric.

$$Q \sim \chi^2_r(0) \text{ iff } A^2 = A, \text{rank}(A) = r$$

Proof. It can be shown (exercise) the moment generating function (mgf) of Q is

$$\begin{aligned}\Psi_Q(t) &= \mathbb{E}e^{tQ} \\ &= \det|I - 2tA|^{-\frac{1}{2}} \\ &= \det|I - 2tD|^{-\frac{1}{2}} \\ &= \prod_{i=1}^{p}(1 - 2t\lambda_i)^{-\frac{1}{2}}\end{aligned} \tag{2.7}$$

where $A = TDT^T$ is the spectral decomposition of A. If $A^2 = A$, then $\lambda_i = 0$ or 1, which implies

$$\Psi_Q(t) = (1 - 2t)^{-r/2}$$

But this is the mgf of $\chi^2_r(0)$. On the other hand, if the above holds, then

$$\prod_{i=1}^{p}(1 - 2\lambda_i t)^{-1/2} = (1 - 2t)^{-r/2} \ \forall t$$

This means $\lambda_i = 0$ for $(p - r)$ different i's and $\lambda_i = 1$ for r different i's.

$\square$

Example 2.3 (Sample variance)**.** Suppose $y \sim \mathcal{N}_p(0, \sigma^2 I)$. Let

$$Q = y^T(I - \frac{\mathbf{1}\mathbf{1}^T}{n})y$$

Then by fundamental theorem,

$$Q/\sigma^2 \sim \chi^2_r$$

where $r = \text{Tr}(I - \frac{\mathbf{1}\mathbf{1}^T}{n}) = n - 1$.

Example 2.4 (General multivariate normal)**.** Suppose $y \sim \mathcal{N}_p(0, \Sigma)$, let

$$W = \Sigma^{-\frac{1}{2}} y \sim \mathcal{N}_p(0, I)$$

and

$$Q = y^T A y = W^T \Sigma^{1/2} A \Sigma^{1/2} W$$

Then by fundamental theorem,

$$Q \sim \chi^2_r \text{ iff } A\Sigma A = A, \text{rank}(A) = r$$

Lemma 3 (Craig's independence lemma)**.** *Let* $y \sim \mathcal{N}(0, I)$ *and* $A^2 = A, B^2 = B$*. Then*

$$y^T A y \perp\!\!\!\perp y^T B y \textit{ iff } AB = 0$$

Proof. If $AB = 0$, then $Ay \perp\!\!\!\perp By$. Since A, B are idempotent, we have

$$y^T A y = y^T A^2 y \perp\!\!\!\perp y^T B^2 y = y^T B y$$

If $y^T A y \perp\!\!\!\perp y^T B y$, then

$$y^T (A + B) y \sim \chi^2$$

By fundamental theorem,

$$(A + B)^2 = A + B$$

Thus, $2AB = 0$. □

Theorem 2.2 (Hogg-Craig)**.** Let $y \sim \mathcal{N}(\mu, I)$ and $Q_i = y^T P_i y$ for $i = 1, 2$. If $Q_i \sim \chi^2_{r_i}(0)$ and $Q_1 - Q_2 \geq 0$, $r_1 > r_2$, then

$$Q_1 - Q_2 \perp\!\!\!\perp Q_2,\ Q_1 - Q_2 \sim \chi^2_{r_1 - r_2}(0)$$

Proof. $0 \leq Q_1 - Q_2 = y^T (P_1 - P_2) y\ \forall y$. In particular, if $y \in \mathcal{N}(P_1)$,

$$0 \leq y^T (-P_2) y \leq 0$$

or

$$y^T P_2^T P_2^T y = 0$$

Thus, $\mathcal{N}(P_1) \subset \mathcal{N}(P_2)$. For any y, $y^T P_2 (I - P_1) y = 0$ since $(I - P_1) \in \mathcal{N}(P_1)$. Therefore,

$$(P_1 - P_2)^2 = P_1^2 - P_1 P_2 - P_2 P_1 + P_2^2 = P_1 - P_2$$

and rank$(P_1 - P_2) =$ Tr$(P_1 - P_2) = r_1 - r_2$. The result follows from fundamental theorem and independence lemma.

□

2.4 General results for quadratic forms

Lemma 4. *If $y \sim \mathcal{N}(0, \Sigma)$, then the mgf of $Q = y^T A y$ is*

$$\frac{1}{\det|I - 2t\Sigma A|^{1/2}} \text{ or } \frac{1}{\det|I - 2t\Sigma^{-1/2} A \Sigma^{1/2}|^{1/2}} \text{ or } \frac{1}{\det|I - 2tA\Sigma|^{1/2}}$$

Proof. Use the fact $|I - AB| = |I - BA|$. □

Theorem 2.3 (Independence of quadratic forms)**.** Let $y \sim \mathcal{N}(0, \Sigma)$, $Q_i = y^T A_i y, i = 1, 2$. Then

$$Q_1 \perp\!\!\!\perp Q_2 \text{ iff } A_1 \Sigma A_2 = 0$$

Proof. If $A_1 \Sigma A_2 = 0$, then

$$\begin{aligned}
\Psi_{Q_1,Q_2}(t_1, t_2) &= \mathbb{E}e^{t_1 Q_1 + t_2 Q_2} \\
&= \int e^{t_1 Q_1 + t_2 Q_2} \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2} y^T \Sigma^{-1} y} dy \\
&= \frac{1}{2\pi|\Sigma|^{1/2}} \int e^{-\frac{1}{2} y^T [\Sigma^{-1} - 2t_1 A_1 - 2t_2 A_2] y} dy \\
&= \frac{\left|\Sigma^{-1} - 2t_1 A_1 - 2t_2 A_2\right|^{-1/2}}{|\Sigma|^{1/2}} \\
&= \frac{1}{|I - 2t_1 A_1 \Sigma - 2t_2 A_2 \Sigma|^{1/2}} \\
&=_{(*)} \frac{1}{|I - 2t_1 A_1 \Sigma|^{1/2} |I - 2t_2 A_2 \Sigma|^{1/2}} \\
&= \Psi_{Q_1}(t_1) \Psi_{Q_2}(t_2)
\end{aligned}$$

where $(*)$ follows from the fact $A_1 \Sigma A_2 = 0$. The other direction follows similarly.

□

Now let $Y \sim \mathcal{N}_p(\mu, \Sigma),\ \Sigma \succ 0$ and we want to find the distribution of $Q = y^T A y$, given that rank$(A) = p$. Write

$$Q = y^T \Sigma^{-1/2} T T^T \Sigma^{1/2} A \Sigma^{1/2} T T^T \Sigma^{-1/2} y$$

where $TT^T = T^T T = I$ and $T^T \Sigma^{1/2} A \Sigma^{1/2} T = \Lambda$, i.e., the spectral decomposition of $\Sigma^{1/2} A \Sigma^{1/2}$. Hence,

$$Q = w^T \Lambda w = \sum_{i=1}^{p} \lambda_i w_i^2$$

where λ_i's are eigenvalues of $\Sigma^{1/2} A \Sigma^{1/2}$ and

$$w \sim \mathcal{N}_p(T^T \Sigma^{-1/2} \mu, I)$$

Thus, $Q = \sum_{i=1}^{r} \lambda_i w_i^2$ is a weighted sum of non-central $\chi_1^2(\delta_i^2)$ variables where

$$\delta_i^2 = (t_i^T \Sigma^{-1/2} \mu)^2$$

and t_i is the i^{th} column of T, r is the rank of A.

Example 2.5 ($\mu = 0, \Sigma = I, A^2 = A$)**.** Then

$$Q =_d \sum_i \lambda_i \chi_1^2(0)$$

where λ_i's are non-zero eigenvalues of A. In other words,

$$Q \sim \chi_r^2(0)$$

Example 2.6 ($\mu = 0, A = \Sigma^{-1}$)**.** Then

$$\begin{aligned} Q &= y^T A y \\ &= y^T \Sigma^{-1/2} \Sigma^{1/2} A \Sigma^{1/2} \Sigma^{-1/2} y \\ &= \left(\Sigma^{-1/2} y\right)^T \left(\Sigma^{-1/2} y\right) \end{aligned}$$

Hence,

$$Q \sim \chi_p^2(0)$$

Example 2.7 ($\mu \neq 0, A = \Sigma^{-1}$)**.** In this case,

$$\delta_i^2 = \mu^T \Sigma^{-1/2} t_i t_i^T \Sigma^{-1/2} \mu$$

and thus,

$$\sum_{i=1}^{p} \delta_i^2 = \mu^T \Sigma^{-1/2} (\sum_{i=1}^{p} t_i t_i^T) \Sigma^{-1/2} \mu$$

and $\sum_{i=1}^{p} t_i t_i^T = TT^T = I$. Since $T^T \Sigma^{1/2} A \Sigma^{1/2} T = I$, $\lambda_1 = \cdots = \lambda_p = 1$. Hence,

$$Q \sim \chi_p^2(\mu^T \Sigma^{-1} \mu)$$

Example 2.8 ($\mu \neq 0, \Sigma = I, A^2 = A$)**.** In this case, we have

$$Q \sim \chi_r^2(\sum_{i=1}^{r} \delta_i^2)$$

where

$$\sum_{i=1}^{r} \delta_i^2 = \mu^T A \mu$$

2.5 Fisher-Cochran's theorem

Fisher-Cochran's theorem is a powerful tool to prove many fundamental results in least square theory. To show Fisher-Cochran's theorem, we need the following two lemmas: Loynes lemma and Marsaglia-Garaybill lemma.

Lemma 5 (Loynes). *If $M^2 = M = M^T$, $P^T = P \succeq 0$ and $I - M - P \succeq 0$. Then*

$$PM = MP = 0$$

Proof. Let $y \in \mathbb{R}^n$ and $z = My$, then $z^T = y^T M^T My$ or

$$z^T z = y^T MMMy = z^T Mz$$

Thus,

$$z^T (I - M)z = 0$$

Next, by assumption,

$$0 \leq z^T (I - M - P)z = -z^T Pz \leq 0$$

Thus,

$$z^T Pz = 0 \Rightarrow Pz = 0 \text{ or } PMy = 0 \, \forall y$$

which suggests

$$PM = MP = 0$$

□

The next lemma was firstly proved by Garabill and Marsaglia in 1957 [3], then in 1964, K. S. Banerjee provided a cleaner proof.

Lemma 6 (Marsaglia-Garaybill, [3] and [4]). *Suppose for $1 \leq i \leq k$,*

$$D_i = D_i^T$$

*Then **any two** of the following statements imply the third:*

1. $D_i^2 = D_i, i = 1, 2, \cdots, k.$
2. $D_i D_j = 0 \; \forall i \neq j.$
3. $D = \sum_{i=1}^k D_i$ *is idempotent.*

Proof. The proof has 3 parts.

- $(1)(2) \Rightarrow (3)$: trivial.

- $(3)(1) \Rightarrow (2)$: Since for any i, D_i is symmetric, then $D_i^2 = D_i$ implies

$$D_i \succeq 0 \; \forall i$$

Thus,

$$D - D_i - D_j = \sum_{k \neq i,j} D_k \succeq 0$$

Next, by third condition,

$$D^2 = D, D^T = D \Rightarrow I - D \succeq 0$$

Hence

$$I - D_i - D_j = (I - D) + (D - D_i - D_j) \succeq 0$$

By Loynes' lemma where

$$M = D_i, P = D_j$$

we have

$$D_i D_j = D_j D_i = 0$$

for any $i \neq j$.

- $(2)(3) \Rightarrow (1)$: By definition of eigenvalues,

$$D_i x = \lambda x \Rightarrow D D_i x = \lambda D x$$

or by second condition,

$$\lambda^2 x = D_i^2 x = \lambda D x$$

Thus, by third condition, $\lambda = 0$ or 1, which means $D_i^2 = D_i$ is idempotent.

□

Theorem 2.4 (Fisher-Cochran). Suppose

$$y \sim \mathcal{N}_p(0, I)$$

and

$$y^T y = \sum_{i=1}^{k} Q_i$$

where $Q_i = y^T A_i y$, rank$(A_i) = r_i$ and each A_i is, of course, symmetric. Then T.F.A.E.:

1. $Q_i \perp\!\!\!\perp Q_j \; \forall i \neq j$.
2. $Q_i \sim \chi^2_{r_i}(0), i = 1, 2, \cdots, k$.
3. $\sum_{i=1}^{k} r_i = p$.

Proof. I will show that $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (i)$.

- $(i) \Rightarrow (ii)$:

  $$Q_i \perp\!\!\!\perp Q_j \Rightarrow Q_1 \perp\!\!\!\perp \sum_{i=2}^{k} Q_i$$

  or eequivalently,

  $$y^T A_1 y \perp\!\!\!\perp y^T (\sum_{i=2}^{k} A_i) y$$

  Then by Craig's independence lemma 3 we have

  $$A_1 (\sum_{i=2}^{k} A_i) = 0$$

  Since $\sum_{i=1}^{k} A_i = I$,

  $$A_1(I - A_1) = 0 \Rightarrow A_1^2 = A_1 \Rightarrow A_i^2 = A_i, i = 1, \cdots, k$$

  Hence by fundamental theorem 2.1,

  $$Q_i \sim \chi^2_{r_i}(0)$$

- $(ii) \Rightarrow (iii)$: Note by fundamental theorem 2.1,

  $$Q_i \sim \chi^2_{r_i}(0) \Rightarrow A_i^2 = A_i$$

  Thus,

  $$\sum_{i=1}^{k} r_i = \sum_{i=1}^{k} \text{Tr}(A_i) = \text{Tr}(\sum_{i=1}^{k} A_i) = \text{Tr}(I) = p$$

- $(iii) \Rightarrow (i)$: This is the part that we need **Marsaglia-Garaybill's lemma** (or Loynes' lemma). Let $A = \sum_{i=2}^{k} A_i$, then

  $$A_1 + A = I$$

  Let T be such that $T^T = TT^T = I$ and

  $$T^T A_1 T = \Lambda$$

  where $T \Lambda T^T$ is the spectral decomposition of A_1. Then

  $$T^T A_1 T + T^T A T = T^T T = I$$

or

$$\Lambda + T^T A T = I$$

KEY STEP: but

$$\text{rank}(T^T A T) = \text{rank}(A) = \text{rank}(\sum_{i=2}^{p} A_i) \leq \sum_{i=2}^{p} \text{rank}(A_i) = p - r_1 \tag{2.8}$$

where the last equality follows from the third condition. Denote the non-zero diagonal entries of Λ as $\lambda_1, \cdots, \lambda_{r_1}$, and the rest are $0's$. Then the corresponding elements of $T^T A T$ are $1 - \lambda_1, \cdots, 1 - \lambda_{r_1}$ and $1's$ respectively. However, inequality 2.8 means

$$1 - \lambda_i = 0 \ \forall i = 1, 2, \cdots, r_1$$

or equivalently,

$$\lambda_i = 1 \ \forall i = 1, 2, \cdots, r_1$$

Therefore,

$$A_1^2 = A_1 \Rightarrow A_i^2 = A_i, i = 1, 2, \cdots, k$$

which means $A_i \succeq 0$ for $i = 1, \cdots, k$. Next,

$$I - A_i - A_j = \sum_{k \neq i,j} A_k \succeq 0$$

and hence by Loynes' lemma

$$A_i A_j = A_j A_i = 0 \ \forall 1 \leq i \neq j \leq k$$

and these two steps can be derived directly by Marsaglia-Garaybill's lemma. But by Craig's independence lemma,

$$Q_i \perp\!\!\!\perp Q_j \text{ iff } A_i \Sigma A_j = 0$$

Thus,

$$Q_i \perp\!\!\!\perp Q_j \ \forall 1 \leq i \neq j \leq k$$

□

2.6 Applications of Fisher-Cochran's theorem

Example 2.9. Suppose $y \sim \mathcal{N}(0, I)$. We have

$$y^T y = y^T \frac{\mathbf{1}\mathbf{1}^T}{n} y + y^T (I - \frac{\mathbf{1}\mathbf{1}^T}{n}) y$$

By Fisher-Cochran, we have

$$y^T(I - \frac{\mathbf{1}\mathbf{1}^T}{n}) \sim \chi_r^2(0),\ r = n-1$$

$$y^T\frac{\mathbf{1}\mathbf{1}^T}{n}y \sim \chi_1^2(0)$$

Example 2.10 (Multiple regression)**.** In regression problem:

$$y = X\beta + \epsilon$$

where $\beta = (\beta_0, \beta_1, \cdots, \beta_{p-1})$. We want to test (**Omnibus test**)

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$$

and note that $\widehat{\beta} = (X^TX)^{-1}X^Ty,$ rank$(X) = p$. Then

$$y^Ty = \underbrace{y^T(I-P)y}_{\text{SSE}} + \underbrace{y^T(P - \frac{\mathbf{1}\mathbf{1}^T}{n})y}_{\text{SSReg}} + y^T\frac{\mathbf{1}\mathbf{1}^T}{n}y \tag{2.9}$$

Then by Fisher-Cochran, the test statistic is

$$\frac{\text{SSReg}}{\text{SSE}}\frac{n-p}{p-1} \sim_{H_0} F_{p-1,n-p}$$

Example 2.11 (One-way ANOVA)**.** Suppose for $i = 1, \cdots, k$ and $j = 1, \cdots n$,

$$\begin{aligned} y_{ij} &= \mu + \tau_i + \epsilon_{ij} \\ \epsilon_{ij} &\sim_{iid} \mathcal{N}(0, \sigma^2) \end{aligned}$$

How do we test $\tau_1 = \cdots = \tau_k$? Note that

$$\sum_{i=1}^k\sum_{j=1}^n(y_{ij} - \bar{y})^2 = \underbrace{\sum_{i=1}^k\sum_{j=1}^n(y_{ij} - \bar{y}_{i\cdot})^2}_{\text{SSE}} + \underbrace{n\sum_{i=1}^k(\bar{y}_{i\cdot} - \bar{y})^2}_{\text{SSTr}} \tag{2.10}$$

By Fisher-Cochran, we have

$$\frac{\text{SSE}}{\sigma^2} = \sum_{i=1}^k\sum_{j=1}^n(\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2/\sigma^2 \sim_{H_0} \sum_{i=1}^k\chi_{n-1}^2$$

and $\mathbb{E}(\frac{\text{SSE}}{(n-1)k}) = \sigma^2$,

$$\frac{\text{SSTr}}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\cdot} - \bar{y})^2/\sigma^2 \sim_{H_0} \chi^2_{k-1}$$

and $\mathbb{E}(\frac{\text{SSTr}}{k-1}) = \sigma^2 + n\frac{\sum_{i=1}^{k}(\tau_i-\bar{\tau})^2}{k-1}$. For more on these, see **ANOVA mixed models**.

2.7 Exercises

1. (Reproducing property) Let $X_i, i = 1, \cdots, n$ be a sequence of normal variables with mean μ_i. Let $Y_1 = \sum_{i=1}^{r} X_i^2$ and $Y_2 = \sum_{i=r+1}^{n} X_i^2$. Show that for $0 < t < 0.5$,

   (a) The mgf of Y_1 is
   $$\Psi_{Y_1}(t) = \frac{e^{\delta_1^2 t/(1-2t)}}{(1-2t)^{r/2}}$$

   (b) The mgf of $Y = Y_1 + Y_2$ is
   $$\Psi_Y(t) = \frac{e^{\lambda t/(1-2t)}}{(1-2t)^{n/2}}$$
   where $\lambda = \delta_1^2 + \delta_2^2$, $\delta_1^2 = \sum_{i=1}^{r} \mu_i^2$ and $\delta_2^2 = \sum_{i=r+1}^{n} \mu_i^2$.

2. Prove formula 2.7.

3. (MVN when covariance is singular) Suppose Y has a p-dimensional multivariate normal distribution with mean m and covariance V. Let Y be partitioned into Y_1 and Y_2 so that $Y^T = (Y_1^T, Y_2^T)$ and the dimension of Y_1 and Y_2 are, respectively, p_1 and p_2 with $p_1 + p_2 = p$. Find the distribution of Y_1 given Y_2 if V is **singular** and express the conditional distribution in terms of only m, p_1 and the four submatrices in the partitioned matrix V and an appropriate g-inverse.

4. (Non-central χ^2) Let m be a p-dimensional vector, $y \sim \mathcal{N}_p(m, I)$ and A is idempotent of rank k. What is the distribution of $(y-a)^T A(y-a)$?

5. (Density of non-central χ^2 and its mgf) Let p_m be the density of a central χ^2 distribution with m degrees of freedom, and for each non-negative s, let
   $$q_j = \frac{(s/2)^j}{j!}\exp(-s/2)$$
   for $j = 0, 1, \cdots$. A random variable with density
   $$h(z) = \sum_{j=0}^{\infty} q_j p_{m+2j}(z),\ z > 0$$

is said to have a non-central χ^2 distribution with m degrees of freedom and non-centrality parameter s. Use the **power series expansion** of $\exp(x)$ and **monotone convergence theorem** to find the moment generating function of such a random variable.

6. (Non-central F) Find the expectation of a non-central F distribution with numerator and denominator degrees of freedom n and m, respectively, and non-centrality parameter Φ.

7. Let $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$ be ind random samples from $\mathcal{N}(\mu_1, v_1^2)$ and $\mathcal{N}(\mu_2, v_2^2)$ respectively. Let $\bar{X}, \bar{Y}, S_1^2, S_2^2$ denote the respective sample means and variances, and let c be a fixed constant. Identify the distribution of each of the following statistics by finding a suitable value of k:

   (a) $\frac{\sqrt{n_2}(\bar{Y}-c)}{kv_2}$.
   (b) $\frac{k\sqrt{n_1}(\bar{X}-c)}{S_1}$.
   (c) $\frac{k(X_1+X_2)}{|Y_1-Y_2|}$.
   (d) $\frac{k[(X_1-c)^2+(X_2-c)^2]}{S_2^2}$.

8. Let $(X_j, Y_j), j = 1, 2, \cdots, n$ be a random sample from the bi-variate normal distribution with parameters m_1, m_2, v_1^2, v_2^2 and correlation r.

   (a) If d is a fixed constant, find a constant k s.t.

   $$T = \frac{k(\bar{X} - \bar{Y} - d)}{\sqrt{\sum_{i=1}^{n}(X_i - Y_i - \bar{X} + \bar{Y})^2}}$$

   has a non-central $t_m(s)$ distribution.

   (b) Express m and s as a function of the parameters and the constant d.

   (c) What is the expectation of T?

9. (250A Midterm) (The ANOVA Theorem) Let V be a $n \times n$ positive definite matrix, let $y \sim \mathcal{N}(m, V)$, let A be a $n \times n$ symmetric matrix of rank r and let

   $$A = A_1 + A_2 + \cdots + A_k$$

   where each A_i is symmetric of rank $r_i, i = 1, 2, \cdots, k$. The following conditions may be defined:

   - (i) $A_iV = (A_iV)^2, i = 1, 2, \cdots, k$;
   - (ii) $A_iVA_j = 0, 1 \leq i \neq j \leq k$;
   - (iii) $AV = (AV)^2$.

   Show that the following statements are all simultaneously true if and only if **any two** of (i), (ii) and (iii) are true:

(a) $y^T A_i y \sim \chi^2_{r_i}(m^T A_i m), i = 1, 2, \cdots, k$.

(b) $y^T A_i y \perp\!\!\!\perp y^T A_j y, 1 \leq i \neq j \leq k$.

(c) $y^T A y \sim \chi^2_r(m^T A m)$.

3 Theory of Least Square Estimation

In 250A, we talked about basics of linear models, properties of LS and G-LS estimators, constrained LSE, adding regressors, more on orthogonal projections, rank-deficient case and some hypothesis testing problems.

3.1 Basics of linear models

3.2 Properties of LS and G-LS estimators

3.3 Adding regressors

3.4 More on projections

3.5 Constrained least square estimation

3.6 When X has less than full column rank

3.7 Some testing problems

3.8 Exercises

4 Multiple and Partial Correlation Coefficient

In this section, I study the basic properties of LS coefficient and multiple and partial correlation coefficient (MCC & PCC).

4.1 Multiple correlation coefficient (MCC)

The multiple correlation coefficient is denoted as r and is defined as

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\widehat{y}_i - \bar{\widehat{y}})}{\sqrt{(\sum_i (y_i - \bar{y})^2)(\sum_i (\widehat{y}_i - \bar{\widehat{y}})^2)}}$$

Theorem 4.1 (Relation between r and R^2)**.** If $\mathbf{1} \in \mathcal{C}(X)$, then $r^2 = R^2$.

Proof. If $\mathbf{1} \in \mathcal{C}(X)$, then $\widehat{y} = Py \Rightarrow \mathbf{1}^T\widehat{y} = \mathbf{1}^T Py = (P\mathbf{1})^T y = \mathbf{1}^T y \Rightarrow \bar{\widehat{y}} = \bar{y}$. Thus,

$$\sum_{i=1}^{n} \bar{y}(\widehat{y}_i - \bar{\widehat{y}}) = y^T P_{\mathbf{1}}(Py - P_{\mathbf{1}}y) = 0$$

Thus I re-write $r = \frac{y^T(P-P_{\mathbf{1}})y}{\sqrt{y^T Q_{\mathbf{1}} y y^T (P-P_{\mathbf{1}})y}}$. Therefore, R^2 is

$$R^2 = \frac{\text{SSR}}{\text{TSS}} = \frac{y^T(P - P_{\mathbf{1}})y}{y^T Q_{\mathbf{1}} y} = r^2$$

□

Example 4.1. We want to test $\beta_1 = \cdots = \beta_{p-1} = 0$ in a LM. This boils down to a F-statistic:

$$\begin{aligned}
F &= \frac{\text{SSE}_0 - \text{SSE}}{\text{SSE}} \frac{n-p}{p-1} \\
&= \frac{\text{TSS} - \text{SSR}_0 - (\text{TSS} - \text{SSR})}{\text{TSS} - \text{SSR}} \frac{n-p}{p-1} \\
&= \frac{1 - R_0^2 - (1 - R^2)}{1 - R^2} \frac{n-p}{p-1} \\
&= \frac{R^2 - R_0^2}{1 - R^2} \frac{n-p}{p-1} \\
&= \frac{R^2}{1 - R^2} \frac{n-p}{p-1}
\end{aligned}$$

The last equality is due to $R_0^2 = \frac{y^T(P_1 - P_1)y}{\text{TSS}} = 0$. Thus, test statistic for H_0 is

$$
\begin{aligned}
F &= \frac{R^2}{1-R^2}\frac{n-p}{p-1} \\
R^2 &= \frac{(p-1)F}{(n-p)+(p-1)F}
\end{aligned}
\tag{4.1}
$$

Note (exercise) $X \sim F_{d_1,d_2} \Rightarrow \frac{d_1 X/d_2}{1+d_1X/d_2} \sim B(\frac{d_1}{2}, \frac{d_2}{2})$. So

$$
R^2 \sim B(\frac{p-1}{2}, \frac{n-p}{2}),\ \mathbb{E}(R^2) = \frac{p-1}{n-1}
$$

4.2 Geometry of LSE

Let $\mathcal{V}_{-k} = \mathcal{L}(x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p)$ and $\theta = \mathbb{E}(y) = X\beta = \sum_{i=1}^p \beta_i \mathbf{x}_i$ and for $k = 1, \cdots, p$,

$$
\begin{aligned}
\widehat{\mathbf{x}}_k &= P_{\mathcal{V}_{-k}}(\mathbf{x}_k) := P_{-k}\mathbf{x}_k \\
\mathbf{x}_k^{\perp} &= \mathbf{x}_k - \widehat{\mathbf{x}}_k = Q_{\mathcal{V}_{-k}}(\mathbf{x}_k) := Q_{-k}\mathbf{x}_k
\end{aligned}
$$

and $\mathbf{x}_k^{\perp}$ provides information other than these provided by $\{x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_p\} \subset \mathcal{V}_{-k}$.

Example 4.2 (Geometry of LSE)**.** Observe that for $j = 1, \cdots, p$,

$$
\langle \theta, \mathbf{x}_j^{\perp} \rangle = \langle \sum_{i=1}^p \beta_i \mathbf{x}_i, \mathbf{x}_j^{\perp} \rangle = \beta_j \|\mathbf{x}_j^{\perp}\|_2^2
$$

Replace θ with y and β_i with $\widehat{\beta}_i$, then for $i, j = 1, \cdots, p$, we have

$$
\begin{aligned}
\widehat{\beta}_i &= \frac{\langle y, \mathbf{x}_i^{\perp} \rangle}{\|\mathbf{x}_i^{\perp}\|_2^2} \\
&= \frac{y^T Q_{-i}\mathbf{x}_i}{\mathbf{x}_i^T Q_{-i}\mathbf{x}_i} && (4.2) \\
Var(\widehat{\beta}_i) &= \frac{\sigma^2}{\|\mathbf{x}_i^{\perp}\|_2^2} \\
&= \frac{\sigma^2}{\mathbf{x}_i^T Q_{-i}\mathbf{x}_i} && (4.3) \\
Cov(\widehat{\beta}_i, \widehat{\beta}_j) &= \frac{\sigma^2 \langle \mathbf{x}_i^{\perp}, \mathbf{x}_j^{\perp} \rangle}{\|\mathbf{x}_i^{\perp}\|_2^2 \|\mathbf{x}_j^{\perp}\|_2^2} \\
&= \cos(\mathbf{x}_i^{\perp}, \mathbf{x}_j^{\perp}) \frac{\sigma^2}{\|x_i\|_2 \|x_j\|_2} && (4.4) \\
\sigma^2 (X^T X)_{ij}^{-1} &= \cos(\mathbf{x}_i^{\perp}, \mathbf{x}_j^{\perp}) \frac{\sigma^2}{\|x_i\|_2 \|x_j\|_2} && (4.5)
\end{aligned}
$$

Lemma 7 (F-test)**.** *Define*

$$R^2 = \frac{\mathit{SSR}}{\mathit{TSS}},\ R_{k-1}^2 = \frac{\mathit{SSR}(\mathbf{x}_1, \cdots, \mathbf{x}_{k-1})}{\mathit{TSS}},\ R_k^2 = \frac{\mathit{SSR}(\mathbf{x}_1, \cdots, \mathbf{x}_k)}{\mathit{TSS}}$$

Then we have

$$\frac{R_k^2 - R_{k-1}^2}{1 - R_{k-1}^2} = \frac{t^2}{n - k + t^2} \tag{4.6}$$

where t is the test statistic for $\beta_k = 0$.

4.3 Partial correlation coefficient (PCC)

Definition 4.1 (MCC)**.** Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{x}_1, \cdots, \mathbf{x}_k \in \mathbb{R}^n$, then the partial correlation coefficient (PCC) of $\mathbf{v}_1$ and $\mathbf{v}_2$ with the linear effects of $\mathbf{x}_1, \cdots, \mathbf{x}_k$ removed is

$$r_{\mathbf{v}_1\mathbf{v}_2\cdot\mathbf{x}_1\cdots\mathbf{x}_k} = \frac{\langle \mathbf{v}_1^{\perp}, \mathbf{v}_2^{\perp} \rangle}{\|\mathbf{v}_1^{\perp}\|_2 \|\mathbf{v}_2^{\perp}\|_2}$$

where $\mathbf{v}_i^{\perp} = Q\mathbf{v}_i, i = 1, 2$. When $k = 1$, $r_{\mathbf{v}_1\mathbf{v}_2\cdot\mathbf{x}_1}$ is just Pearson correlation.

Now we express PCC of order k in terms of oder $k-1$. For example, if $k = 4$, then

$$r_{14\cdot23} = \frac{r_{14} - r_{13\cdot2}r_{34\cdot2}}{\sqrt{1 - r_{13\cdot2}^2}\sqrt{1 - r_{34\cdot2}^2}}$$

Consider the general case in the definition and let $\mathcal{V}_J = \mathcal{L}(x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_k)$ and $\mathcal{V}_I = \mathcal{L}(x_1, \cdots, x_k)$. Define $P_J = P_{\mathcal{V}_J}$, $Q_J = I - P_J$, $P_I = P_{\mathcal{V}_I}$, $Q_I = I - P_I$ and

$$\widehat{v}_i := P_J\mathbf{v}_i,\ v_i^{\perp} := Q_J\mathbf{v}_i,\ \widehat{x}_i := P_J\mathbf{x}_i,\ x_i^{\perp} := Q_J\mathbf{x}_i$$

By 1.10 we have $P_I = P_J + \frac{Q_J\mathbf{x}_j\mathbf{x}_j^TQ_J}{\mathbf{x}_j^TQ_J\mathbf{x}_j}$. Thus,

$$\begin{aligned} w_i &= \mathbf{v}_i - P_I\mathbf{v}_i \\ &= Q_J\mathbf{v}_i - \frac{Q_Jx_jx_j^TQ_J}{x_j^TQ_Jx_j}\mathbf{v}_i \\ &= v_i^{\perp} - \frac{\langle x_j^{\perp}, \mathbf{v}_i^{\perp} \rangle}{\|x_j^{\perp}\|_2^2}x_j^{\perp} \end{aligned}$$

Thus, for $i = 1, 2$,

$$\|w_i^{\perp}\|_2^2 = \|v_i^{\perp}\|_2^2 - \frac{\langle x_j^{\perp}, \mathbf{v}_i^{\perp} \rangle^2}{\|x_j^{\perp}\|_2^2} = \|v_i^{\perp}\|_2^2(1 - r_{ij\cdot J}^2)$$

Where $r_{ij\cdot J} = r_{\mathbf{v}_1\mathbf{v}_2\cdot\mathbf{x}_1\cdots\mathbf{x}_{j-1}\mathbf{x}_{j+1}\cdots\mathbf{x}_k}$. Besides,

$$\langle w_1, w_2 \rangle = \|v_1^{\perp}\|_2\|v_2^{\perp}\|_2(r_{12\cdot J} - r_{13\cdot J}r_{23\cdot J})$$

To sum up, we showed

$$r_{12\cdot I} = \frac{r_{12\cdot J} - r_{13\cdot J}r_{23\cdot J}}{\sqrt{1-r_{1j\cdot J}^2}\sqrt{1-r_{2j\cdot J}^2}} \tag{4.7}$$

4.4 Exercises

1. If $F \sim F_{a,b}$, then show that
   (a) $\frac{aF/b}{1+aF/b} \sim B(a/2, b/2)$.
   (b) The expectation of this Beta RV is $\frac{a}{a+b}$.
2. ([1], p113) Consider a general full rank linear regression model. Show that R^2 and F for testing $H_0 : \beta_j = 0, j \neq 0$ are independent of the units in which y_i and x_{ij} are measured.
3. Let $y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$. Show $r_{yx_k\cdot x_1\cdots x_{k-1}}$ is a function of the test statistic for $H_0 : \beta_k = 0$.

5 Violation of Assumptions and Remedies

This section talks about outlier detection, consequences of under-fitting and over-fitting, transformation and collinearity.

5.1 Outlier detection

Consider a general linear model:

$$\begin{aligned} y &= X\beta + \epsilon \\ \mathbb{E}y &= X\beta \\ \widehat{y} &= X\widehat{\beta} = Py \\ e &= y - \widehat{y} = Qy \end{aligned}$$

We have $\mathbb{E}e = 0$ and $Var(e) = \sigma^2 Q, Cov(e, \widehat{y}) = Cov(Qy, Py) = 0$.

Definition 5.1 (Internally and externally studentized residual)**.** The internally studentized residual is defined as

$$r_i = \frac{e_i}{s\sqrt{(1-h_{ii})}}$$

and the externally studentized residual is defined as

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

where h_{ii} is the i^{th} diagonal element of P and $s, s_{(i)}$ are the usual estimates of σ^2 with and without the i^{th} case.

Recall that we have

$$\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{(X^TX)^{-1}x_i e_i}{1-h_{ii}}, \ e_i = y_i - x_i^T\widehat{\beta}$$

where x_i is the i^{th} row of X. We can use this formula to show the relationship between $s_{(i)}^2$ and s^2. Besides, we have two useful results:

$$\begin{aligned} t_i^2 &=_d \frac{B}{1-B}(n-p-1) \\ r_i^2 &=_d B(n-p) \end{aligned} \tag{5.1}$$

where $B \sim Beta(1/2, (n-p-1)/2)$. For proof, see exercises.

5.2 Under-fitting and over-fitting

In this subsection, we study the linear model when the working model is over and under fitted.

Example 5.1 (Underfitting)**.** Suppose the working model is (full rank)

$$\mathbb{E}(y) = X\beta$$

and the true model is

$$\mathbb{E}(y) = X\beta + Z\gamma$$

Let $\widehat{\beta} = (X^TX)^{-1}X^Ty$ be the usual LSE and $Cov(\widehat{\beta}) = \sigma^2(X^TX)^{-1}$. Then,

$$\mathbb{E}\widehat{\beta} = \beta + (X^TX)^{-1}X^TZ\gamma$$

Thus, $\widehat{\beta}$ is unbiased iff $X^TZ = 0$. Next, SSE is

$$s^2 = \frac{y^TQy}{n-p}$$

Thus,

$$\begin{aligned}\mathbb{E}s^2 &= \frac{1}{n-p}\left((\mathbb{E}y)^TQ\mathbb{E}y + Tr(QVar(y))\right)\\ &= \sigma^2 + \frac{1}{n-p}\gamma^TZ^TQZ\gamma\end{aligned}$$

So s^2 overestimates σ^2 unless $QZ = 0$. In addition, let $\widehat{y} = X\widehat{\beta}$,

$$\mathbb{E}\widehat{y} = X\beta + P_XZ\gamma$$

Finally, let $e = y - \widehat{y}$, then

$$\mathbb{E}e = \mathbb{E}y - \mathbb{E}X\widehat{\beta} = Q_XZ\gamma$$

$$Var(e) = \sigma^2Q$$

So $Var(e)$ stays the same. Next, let $\widetilde{\beta}$ be the LSE under the true model, then

$$\begin{aligned}\widetilde{y} &= (X\ Z)\widetilde{\beta}\\ &= Py - QZ(Z^TQZ)^{-1}Z^TQy\end{aligned}$$

Thus, $Var(\widetilde{y}) - Var(\widehat{y})$ is positive semi-definite.

Example 5.2 (Overfitting)**.** The true model is $\mathbb{E}y = X_1\beta_1$ and the working model is $\mathbb{E}y = X_1\beta_1 + X_2\beta_2$. Let $X = (X_1\ X_2)$, then the expectation of LSE is

$$\mathbb{E}\widehat{\beta} = (X^TX)^{-1}X^T(X_1\ X_2)(\beta_1^T\ 0)^T$$

So $\widehat{\beta}_1$ is unbiased.

Example 5.3 (Mis-specification of covariance)**.** The working model is $Cov(\epsilon) = \sigma^2I$ while the true model is $Cov(\epsilon) = \sigma^2 V$. Then $\widehat{\beta}$ is still unbiased but

$$Cov(\widehat{\beta}) = \sigma^2 (X^T X)^{-1} X^T V X (X^T X)^{-1}$$

But is $\widehat{\sigma}^2 = \frac{y^T Q y}{n-p}$ biased for σ^2?

Theorem 5.1 (Bounding the ratio)**.** If $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ are eigenvalues of V, then

$$\frac{1}{n-p} \sum_{i=1}^{n-p} \mu_i \leq \mathbb{E} \frac{\widehat{\sigma}^2}{\sigma^2} \leq \frac{1}{n-p} \sum_{i=n-p+1}^{n} \mu_i$$

To prove theorem 5.1, we need the following lemma.

Lemma 8. *Let $H^T = H = H^2$ and rank$(H) = r$. If $A^T = A$, then*

$$\sum_{i=1}^{r} \lambda_i \leq Tr(HA) \leq \sum_{i=n-r+1}^{n} \lambda_i$$

where $\lambda_1 \leq \cdots \leq \lambda_n$ are eigenvalues of A.

Proof. By spectral decomposition theorem, $\exists\ P$ and $P^T P = P^T P = I$ such that

$$P^T H P = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} := D_r$$

Note $D_r^2 = D_r$. Thus,

$$\begin{aligned} Tr(HA) &= Tr(P D_r P^T A) \\ &= Tr((P D_r)^T (A P D_r)) \\ &= \sum_{i=1}^{r} p_i^T A p_i \end{aligned}$$

where p_i is the i^{th} column of A. Then the lemma follows from theorem 1.8. □

5.3 Transformation

Consider transform x by $g(x)$, then

$$\mathbb{E}(y) = \alpha + \beta^T g(x)$$

Consider simple transformations like

$$g_\lambda(x) = \begin{cases} \ln x, & \lambda = 0 \\ \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \end{cases}$$

In practice, run the regression with

$$\lambda = \{-2, , -1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{2}, 1, 2\}$$

Do the same procudure on the response ($y > 0$) and ad-hoc. Alternatively, use variance stabilizing transformation so that $Var(g(y))$ is a constant. By Taylor's expansion,

$$g(y) = g(\mu) + g'(\mu)(y - \mu) + o_p(y - \mu)$$
$$Var(g(y)) \approx Var(y)[g'(\mu)]^2 = \sigma^2[g'(\mu)]^2$$

Thus, to make $Var(g(y)) = c$ a constant, $g'(\mu) = \sqrt{\frac{c^2}{\sigma^2}}$, or,

$$g(\mu) = \int \frac{c}{\sqrt{Var(y)}} d\mu$$

In case of Poisson, $\mu = Var(y), c = 1$,

$$g(y) = 2\sqrt{y}$$

5.4 Collinearity

Recall that in a classical linear model, we have least square estimate $\widehat{\beta}^T = (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)$ and

$$Var(\widehat{\beta}_k) = \frac{\sigma^2}{\|x_k^{\perp}\|_2^2}$$

where $x_k^{\perp} = x_k - P_{-k}x_k$ and P_{-k} is the projection operator without the k^{th} column x_k. If x_k can be approximated by linear combinations of the other $x's$ (collinearity), then $\widehat{\beta}_k$ will be estimated unreliably. How to detect collinearity?

Definition 5.2 (Variance inflation factor)**.** Let R_j^2 be the usual R^2 obtained when regressing x_j on the other covariates, say X_{-j}. Then variance inflation factor is defined as

$$VIF_j = \frac{1}{1 - R_j^2}$$

If VIF_j is large, say > 10, then x_j is almost linearly related with the rest of x's. What to do?

1. Omit the variable or variables with $VIF > 10$.

2. $\mathbb{E}y = X\beta = XR^{-1}R\beta$ where R is an upper triangular non-singular matrix and is obtained from the QR decomposition of X:

$$X = QR = \begin{pmatrix} q_1 & q_2 & \cdots & q_p \end{pmatrix} R$$

$$q_i^T q_j = \delta_{ij},\ q_1 = \frac{x_1}{\|x_1\|_2}$$

Let $\gamma = R\beta$ then

$$\widehat{\gamma} = \widehat{R\beta} = (Q^T Q)^{-1} Q^T y = Q^T y = (XR^{-1})^T y$$

The regression sum of squares is

$$\text{SSReg} = y^T P_{\mathcal{C}(Q)} y = \sum_{i=1}^{p} (q_i^T y)^2$$

where q_i is the i^{th} column of Q.

5.5 Exercises

1. (Mean shift model) We want to test whether the i^{th} case has outlying x-values using the mean-shift mode given by

$$\mathbb{E}y = X\beta + \Theta\phi_i$$

where ϕ_i is the vector with all its components equal to 0 except for the i^{th} element, which is euqal to 1. Derive a test whether the i^{th} case is outlying and show that this test statistic is the i^{th} externally studentized statistic and its square has a $F_{1,n-p-1}$ distribution.

2. ([1], p270) Prove equations 5.1.

3. ([1], p268 and p311) Show that

$$t_i^2 = \frac{(n-p-1)r_i^2}{n-p-r_i^2}$$

4. (Cook's distance) The Cook's distance of the i^{th} case is

$$\text{Cook}_i = \frac{(\widehat{y} - \widehat{y}_{(i)})^T (\widehat{y} - \widehat{y}_{(i)})}{ps^2}$$

where $\widehat{y}_{(i)} = X\beta_{(i)}$ and $s^2 = \frac{\text{SSE}}{n-p}$. Express it in terms of p, h_{ii} and r_i.

5. (Added variable plot) Suppose $\mathbb{E}y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and the standard assumptions hold. An added variable plot is constructed by first regressing y on $x_1, \cdots, x_k$ and regressing x_k on $x_1, \cdots, x_{k-1}$, and then regressing the first set of residuals e_1 on the second set of residuals e_2.

   (a) What are the fitted regression coefficients when you regress e_1 on e_2 and relate your answers to those fitted coefficients from regressing y on $x_1, \cdots, x_k$.

   (b) What is correlation e_1 and e_2 and describes its relationship to the partial correlation between y and x_k controlling for $x_1, x_2, \cdots, x_{k-1}$.

6. Prove theorem 5.1.

7. (A new metric for detecting outlier)To measure the influence of the i^{th} observation on the i^{th} predicted value, it was proposed that a distance measure based on a scale difference $\widehat{y}_i - \widehat{y}_{(i)}$ be used:

$$\frac{|\widehat{y}_i - \widehat{y}_{(i)}|}{s_{(i)} h_{ii}^{1/2}}$$

Here the usual notation holds, eg. h_{ii} is the leverage for the i^{th} case, etc.

   (a) Show that this statistic can be written as

   $$|t_i| \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

   where t_i is the externally studentized residual for the i^{th} case.

   (b) Use this statistic and suggest a rule to assess whether the i^{th} observation is influential.

8. (VIF and LSE) Suppose we regress Y on p covariates $x_1, \cdots, x_p$ and assume further that the model has an intercept and all covariates are centered and scaled so that each has sample mean 0 and sample variance equal to unity. Prove or disprove that under standard assumptions, the variance of the LSE for the coefficient of j^{th} covariate is

$$\frac{\sigma^2}{1 - R_j^2}$$

where R_j is the sample multiple correlation coefficient obtained by regressing x_j on the rest of the other x_i's. Why is this result useful?

9. (Final exam, covariance ratio) Consider the model $y = X\beta + \epsilon$, X is a $n \times p$ matrix of full column rank and $\mathbb{E}\epsilon = 0$, $Var(\epsilon) = \sigma^2 I_n$. Let b be the LSE of β and let VR_i be the ratio of the determinants of the estimated covariance matrices of b with and without the i^{th} case. Express this statistic only in terms of n, p, h_{ii} and r_i, the i^{th} internally studentized residue.

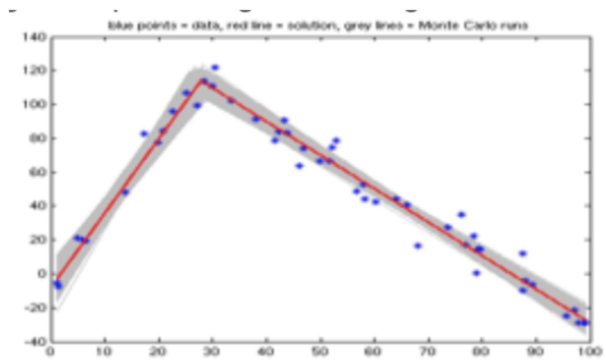
6 Hypothesis Testing and Simultaneous Inference

In this section, topics include Fieller's theorem, lack of fitness test, Scheffe's method, Tukey's q. Besides, various tests of homogeneity of variance are discussed, including Hartley's test, Levene's test, Brown-Forsythe's test and Cook's likelihood ratio test.

6.1 Fieller's theorem

Consider a two phase regression problem:

Figure 1: Two phase regression.



$$\mathbb{E}y = \begin{cases} \alpha_1 + \beta_1 x, & x \le \gamma \\ \alpha_2 + \beta_2 x, & x \ge \gamma \end{cases}$$

The continuity assumption

$$\alpha_1 + \beta_1 \gamma = \alpha_2 + \beta_2 \gamma$$

generates the point estimation of γ:

$$\widehat{\gamma} = \frac{\widehat{\alpha}_1 - \widehat{\alpha}_2}{\widehat{\beta}_2 - \widehat{\beta}_1}$$

The question is: can we construct a confidence interval for γ based on $\widehat{\gamma}$? Fieller gave the most complete discussion in 1954.

Theorem 6.1 (Fieller's theorem)**.** Suppose $\begin{pmatrix} X \\ Y \end{pmatrix}$ is the mean of a random sample of bi-variate normal random variables with distribution $\mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \sigma^2 \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}\right)$, and s^2 is an unbiased estimator of σ^2 such that

$$\frac{rs^2}{\sigma^2} \sim \chi^2_r$$

Let $\theta = \frac{\mu_X}{\mu_Y}$ be the parameter of interest. Then a $(1-\alpha)$ confidence interval (θ_L, θ_U) for θ is

$$(\theta_L, \theta_U) = \frac{1}{1-g}\left[\frac{X}{Y} - \frac{gv_{12}}{v_{22}} \mp \frac{t_{r,\alpha}s}{Y}\sqrt{v_{11} - 2\frac{X}{Y}v_{12} + \frac{X^2}{Y^2}v_{22} - g\left(v_{11} - \frac{v_{12}^2}{v_{22}}\right)}\right] \tag{6.1}$$

where

$$g = \frac{t_{r,\alpha}^2 s^2 v_{22}}{Y^2}$$

and $t_{r,\alpha}$ is the α-level deviate from the student's t-distribution based on r degrees of freedom.

Proof. Let $Z = X - \theta Y$ and we have $Z \perp\!\!\!\perp s^2$. Thus,

$$\frac{Z/\sqrt{v_{11} - 2v_{12}\theta + v_{22}\theta^2}}{s} \sim t_r$$

where t_r is a student t variable with r degrees of freedom. Thus,

$$\mathbb{P}(\left(\frac{Z}{s\sqrt{v_{11} - 2v_{12}\theta + v_{22}\theta^2}}\right)^2 \leq t_{r,\alpha}^2) = 1 - \alpha$$

Solving for the quadratic euqation

$$Z^2 = t_{r,\alpha}^2 s^2(v_{11} - 2v_{12}\theta + v_{22}\theta^2)$$

It boils down to

$$(1-g)\theta^2 - (\frac{2X}{Y} - \frac{2gv_{12}}{v_{22}})\theta + X^2 - \frac{gv_{11}}{v_{22}} = 0$$

and the solution gives the upper and lower bound. □

Example 6.1 (x-intercept estimation)**.** Consider a simple linear model:

$$y - \bar{y} = \beta_0 + \beta(x - \bar{x})$$

To estimate the x-intercept, we let $y = y_0$ and plug-in the usual LSE for β_0, β:

$$\widehat{x}_0 = \frac{y_0 - \bar{y}}{\widehat{\beta}} + \bar{x}$$

We apply Fieller's theorem to construct a confidence interval for the true x_0:

$$\begin{aligned}
&X \leftarrow y_0 - \bar{y}\\
&Y \leftarrow \widehat{\beta}\\
&v_{11} = \sigma^2(1 + \frac{1}{n})\\
&v_{22} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\\
&v_{12} = 0
\end{aligned}$$

Note $v_{12} = 0$ follows from $P_{\mathbf{1}} \perp P_x$. Then Fieller's theorem gives

$$(x_L, x_U) = \bar{x} + \frac{1}{1-g}\left[\frac{y_0 - \bar{y}}{\widehat{\beta}} \mp \frac{t_{r,\alpha} s}{\widehat{\beta}}\sqrt{\sigma^2(1-g)(1+\frac{1}{n}) + \frac{(y_0 - \bar{y})^2 \sigma^2}{\widehat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2}}\right]$$

Example 6.2 (Clinical trial)**.** We are performing a prognosting factors interaction assessment in a 2-treatment trial with treatment T and C, along with a prognostic factor. The model is

$$\mathbb{E}y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where x_2 is a prognostic factor and

$$x_1 = \begin{cases} 1 & \text{if treatment } T \\ 0 & \text{if treatment } C \end{cases}$$

Thus, $\mathbb{E}y_T = \beta_0 + \beta_1 + \beta_2 x_2 + \beta_3 x_2, \mathbb{E}y_C = \beta_0 + \beta_2 x_2$ and

$$\mathbb{E}(y_T - y_C) = \beta_1 + \beta_3 x_2$$

We want to know x_2 for which $\mathbb{E}(y_T - y_C) > 0$. That is, we want to estimate $x_2 = -\frac{\beta_1}{\beta_3}$. Let $\widehat{\beta}$ be the LSE and define

$$\widetilde{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_3 \end{pmatrix};\ Cov(\widetilde{\beta}) = V = \begin{pmatrix} v_{11} & v_{13} \\ v_{31} & v_{33} \end{pmatrix}$$

Let $w = \widetilde{\beta}_1 + \theta\widetilde{\beta}_3, \theta = -\beta_1/\beta_3$, then$\mathbb{E}w = 0$ and $Var(w) = (v_{11} + \theta^2 v_{33} + 2\theta v_{13})\sigma^2$. We have

$$\frac{w}{\sqrt{Var(w)}} \sim \mathcal{N}(0,1)$$

If σ^2 is known, then solving for

$$(\widehat{\beta}_1 + \theta\widehat{\beta}_3)^2 = z_{\alpha/2}^2(v_{11} + 2\theta v_{13} + \theta^2 v_{33})\sigma^2$$

gives the $1-\alpha$ confidence interval for θ:

$$(\theta_L, \theta_U) = \frac{-w_2 \mp \sqrt{w_2^2 - 4w_1 w_2}}{2w_2}$$

where

$$\begin{aligned} w_1 &= \widehat{\beta}_1^2 - v_{11} z_{\alpha/2}^2 \\ w_2 &= 2\widehat{\beta}_1 \widehat{\beta}_3 - 2v_{13} z_{\alpha/2}^2 \\ w_3 &= \widehat{\beta}_3^2 - v_{33} z_{\alpha/2}^2 \end{aligned}$$

6.2 Lack of fitness test

Suppose we want to test

$$H_0 : \mathbb{E}Y = X\beta \text{ vs } H_a : \mathbb{E}Y \neq X\beta$$

with the following assumptions:

$$\begin{aligned} &n_1 \text{ obs on } y \text{ at } x_1^T = (x_{11}, \cdots, x_{1k}) \\ &n_2 \text{ obs on } y \text{ at } x_2^T = (x_{21}, \cdots, x_{2k}) \\ &\qquad\cdots\cdots \\ &n_g \text{ obs on } y \text{ at } x_g^T = (x_{g1}, \cdots, x_{gk}) \\ &n = \sum_{i=1}^{g} n_i \text{ is the total number of obs} \end{aligned}$$

In matrix form, we have

$$\mathbb{E}Y = WX\beta$$

where

$$W = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_g} \end{pmatrix} \; X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \cdots \\ \mathbf{x}_g^T \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \end{pmatrix}$$

Regressing y on $x_1, \cdots, x_k$ gives error sum of square **SSE** $= y^T Q_{WX} y$, pure error sum of square **SSPE** $= y^T Q_W y$ and lack of fitness sum of square **SSLoF** $= y^T(Q_{WX} - Q_W)$ where

$$Q_{WX} = I_n - WX\left(X^T W^T W X\right)^{-1} X^T W^T$$

$$Q_W = \begin{pmatrix} I_{n_1} - \frac{\mathbf{1}\mathbf{1}^T}{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & I_{n_2} - \frac{\mathbf{1}\mathbf{1}^T}{n_2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_{n_g} - \frac{\mathbf{1}\mathbf{1}^T}{n_g} \end{pmatrix}$$

Then

$$\begin{aligned} \text{SSE} &= y^T(Q_{WX} - Q_W + Q_W)y \\ &= y^T(Q_{WX} - Q_W)y + y^T Q_W y \\ &= \text{SSLoF} + \text{SSPE} \end{aligned} \tag{6.2}$$

Since $\mathcal{C}(WX) \subset \mathcal{C}(X)$,

$$\begin{aligned} (Q_{WX} - Q_W)Q_W &= Q_W - Q_W = 0 \\ (Q_{WX} - Q_W)(Q_{WX} - Q_W) &= Q_{WX} - Q_W \end{aligned}$$

This implies

$$y^T(Q_{WX} - Q_W)y \perp\!\!\!\perp y^T Q_W y$$

So **SSLoF** $\perp\!\!\!\perp$ **SSPE** and

$$\begin{aligned} \textbf{SSLoF}/\sigma^2 &\sim \chi^2_{g-k} \\ \textbf{SSPE}/\sigma^2 &\sim \chi^2_{n-g} \end{aligned}$$

The test statistic for $H_0 : \mathbb{E}Y = X\beta$ is

$$F = \frac{\textbf{SSLoF}/(g-k)}{\textbf{SSPE}/(n-g)} \sim F_{g-k,n-g} \tag{6.3}$$

6.3 Scheffe's method

Let $A \in \mathbb{R}^{q \times p}$ and rank$(A) = q$. We want to find an interval estimate for $\Phi = A\beta$. For LSE, we have

$$\widehat{\Phi} = A\widehat{\beta} \sim \mathcal{N}_q(A\beta, \sigma^2 L)$$

where $L = A(X^T X)^{-1} A^T$. Thus,

$$\frac{1}{\sigma^2}(\widehat{\Phi} - \Phi)^T L^{-1}(\widehat{\Phi} - \Phi) \sim \chi^2_q$$

Besides, $\frac{s^2(n-p)}{\sigma^2} = \frac{\widehat{\sigma}^2(n-p)}{\sigma^2} \sim \chi^2_{n-p}$. So,

$$\frac{(\widehat{\Phi}-\Phi)^T L^{-1}(\widehat{\Phi}-\Phi)}{qs^2} \sim F_{q,n-p}$$

and

$$1-\alpha = \mathbb{P}\left(\frac{(\widehat{\Phi}-\Phi)^T L^{-1}(\widehat{\Phi}-\Phi)}{qs^2} \leq F_{q,n-p,\alpha}\right) \tag{6.4}$$

Solving the inequality in 6.4 for Φ is non-trivial. However, Scheffe in 1953 provided a beautiful method to obtain confidence intervals for $h^T\Phi$ where h is any q-dimensional vector.

Theorem 6.2 (Scheffe's theorem)**.** Let $m = qs^2 F_{q,n-p,\alpha}$ and $b = \widehat{\Phi} - \Phi$. The folloing holds:

$$\mathbb{P}\left(b^T L^{-1} b \leq m\right) = 1-\alpha = \mathbb{P}\left(\forall h : h^T\Phi \in h^T\widehat{\Phi} \pm \sqrt{m \cdot h^T L h}\right) \tag{6.5}$$

Proof.

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(b^T L^{-1} b \leq m\right) \\
&=_{(*)} \mathbb{P}\left(\max_{h\neq 0} \frac{(h^T b)^2}{h^T L h} \leq m\right) \\
&= \mathbb{P}\left(\forall h, (h^T b)^2 \leq m \cdot h^T L h\right) \\
&= \mathbb{P}\left(\forall h, |h^T b| \leq \sqrt{m \cdot h^T L h}\right) \\
&= \mathbb{P}\left(\forall h : h^T\Phi \in h^T\widehat{\Phi} \pm \sqrt{m \cdot h^T L h}\right)
\end{aligned}$$

$(*)$ is due to the generalized Cauchy-Schwarz inequality 1.9. ☐

6.4 Tukey's q

This subsection includes Tukey's q distribution and multiple comparison tests for one-way ANOVA.

Definition 6.1 (Tukey's q)**.** Let $Z_1, \cdots, Z_k$ and U be independent random variables with $Z_i \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_m(0)$. Define

$$q = \max_{i\neq j} \frac{|Z_i - Z_j|}{\sqrt{U/m}}$$

We call q has a studentized range distribution with k and m degrees of freedom and write

$$q \sim q_{k,m}$$

Example 6.3 (Tukey's pairwise comparison)**.** In a one-way ANOVA, for $j = 1, \cdots, n$ and $i =$

$1, \cdots, k$,

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\
\bar{y}_{i\cdot} &= \mu + \alpha_i + \bar{\epsilon}_{i\cdot} \\
\bar{y}_{j\cdot} &= \mu + \alpha_j + \bar{\epsilon}_{j\cdot}
\end{aligned}
$$

We have (exercise)

$$
\max_{i \neq j} \frac{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\alpha_i - \alpha_j)|}{\widehat{\sigma}} \sim q_{k,k(n-1)} \tag{6.6}
$$

where $\widehat{\sigma}^2$ is the usual estimate of σ^2. It follows that

$$
\begin{aligned}
&\mathbb{P}\left(\alpha_i - \alpha_j \in \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm \frac{\widehat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \forall i \neq j\right) \\
&= \mathbb{P}\left(\frac{\sqrt{n}(\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\alpha_i - \alpha_j))}{\widehat{\sigma}} \leq q_{k,k(n-1),\alpha} \forall i \neq j\right) \\
&= 1 - \alpha
\end{aligned}
$$

Next step is to construct sets of confidence intervals for **contrasts**

$$
\sum_i c_i \alpha_i \ \forall c \text{ s.t. } c^T \mathbf{1} = 0
$$

in a one-way ANOVA. We have

$$
\begin{aligned}
&\mathbb{P}\left(\sum_i c_i \alpha_i \in \sum_i c_i \bar{y}_{i\cdot} \pm \frac{\widehat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \left[\sum_i |c_i|/2\right], \forall c\right) \\
&= \mathbb{P}\left(\left|\sum_i c_i (\bar{y}_{i\cdot} - \alpha_i)\right| \leq \frac{\widehat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \left[\sum_i |c_i|/2\right], \forall c\right) \\
&=_{(*)} \mathbb{P}\left(|\bar{y}_{i\cdot} - \alpha_i - (\bar{y}_{j\cdot} - \alpha_j)| \leq \frac{\widehat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha}, \forall\, i, j\right) \\
&= 1 - \alpha
\end{aligned}
$$

where $(*)$ comes from the following lemma.

Lemma 9. *Let $a_1, a_2, \cdots, a_k$ be real numbers. Then the following are equivalent*

- $$
|a_i - a_j| \leq b, \forall i, j
$$

- $$
|\sum_{i=1}^k c_i a_i| \leq \frac{b}{2}\left(\sum_{i=1}^k |c_i|\right), \forall c \textit{ s.t. } \sum_{i=1}^k c_i = 0
$$

Proof. One direction is trivial: set $c_i = 1$, $c_j = -1$ and $c_m = 0$ for $m \neq i, j$. The other direction is non-trivial but there is a simpler case in the exercise.

$\square$

6.5 Test of homogeneity

In ANOVA, k groups may have variances σ_i^2, $i = 1, \cdots, k$.

$$
\begin{aligned}
&H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma \\
&H_1 : \sigma_i \neq \sigma_j, i \neq j
\end{aligned}
$$

Example 6.4 (Hartley's test)**.** Let $\widehat{s}_i^2$ be the unbiased estimate of $\sigma_i^2, i = 1, \cdots, k$. The ratio

$$
F_{\max} = \frac{\max_{1 \leq i \leq k} \widehat{s}_i^2}{\min_{1 \leq i \leq k} \widehat{s}_i^2} \tag{6.7}
$$

is close to 1 if H_0 holds, otherwise we reject H_0. This is Hartley's $F_{\max}$ statistic.

Example 6.5 (Levene's test)**.** Given $\{y_{ij} : i = 1, \cdots, k; j = 1, \cdots, n_i\}$, we define

$$
\begin{aligned}
z_{ij} &= |y_{ij} - \bar{y}_{i\cdot}| \\
z_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij} \\
z_{\cdot\cdot} &= \frac{1}{n} \sum_{i}^{k} \sum_{j}^{n_i} z_{ij}
\end{aligned}
$$

Then use

$$
W = \frac{\sum_i^k n_i (z_i - z_{\cdot\cdot})^2 / (k-1)}{\sum_i \sum_j (z_{ij} - z_{i\cdot})^2 / (n-k)} \tag{6.8}
$$

as the test statistic. This is called Levene's test.

Example 6.6 (Brown-Forsythe test)**.** If $\bar{y}_{i\cdot}$ is replaced by $\widetilde{y}_{i\cdot}$: the median in i^{th} group, then it is Brown-Forsythe test.

Example 6.7 (Bartlett's test)**.** In 1937, Bartlett proposed a modification of the likelihood ratio test (LRT) totest the homogeneity problem. Let s_i^2 be the sample variance in i^{th} group, $n = \sum_{i=1}^k n_i$ be the total number of observations and

$$
s^2 = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2
$$

The Bartlett's test statistic is

$$\chi^2 = \frac{(n-k)\ln(s^2) - \sum_{i=1}^{k}(n_i-1)\ln(s_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(\frac{1}{n_i-1}) - \frac{1}{n-k}\right)} \tag{6.9}$$

The test statistic has approximately a χ^2_{k-1} distribution.

6.6 Cook-Weisberg score test

This subsection is based on question 2 of the final exam of 250B. Suppose we would like to test whether the variance of a continuous response depends on the levels of several independent variables in a regression model, then what kind of statistical procedure should we use? I think the paper written by Cook and Weisberg [5] in 1983 provided an alternative and it deserves my effort of writing a subsection.

Example 6.8 (Cook-Weisberg test, [5])**.** Assume the regression model is

$$\begin{aligned} Y &= \beta_0\mathbf{1} + X\beta + \epsilon \\ X_0 &= (\mathbf{1}, X) \text{ is fixed.} \\ \epsilon &\sim \mathcal{N}_n(0, \sigma^2 W) \end{aligned}$$

where $X \in \mathbb{R}^{n\times(p-1)}$, Y is the continuous response vector and

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

and

$$w_i = w(\lambda, z_i)$$

where z_i's are q_z-dimensional independent variables and λ is a q_λ-dimensional parameter to be tested. For example, w_i can be taken as

$$\begin{aligned} w_i &= \exp(\sum_{j=1}^{q}\lambda_j z_{ij}) \\ w_i &= \prod_{j=1}^{q} z_{ij}^{\lambda_j} \end{aligned}$$

We want to test

$$H_0 : \lambda = 0 \text{ vs } H_1 : \lambda \neq 0 \tag{6.10}$$

WLOG, assume $q_z = q_\lambda = q$ and

$$w(0, z_i) = 1,\ W =_{H_0} I$$

Let $\widehat{\beta}_0$ and $\widehat{\beta}$ be the maximum likelihood estimation of β_0 and β under H_0. Define

$$e_i = (y_i - \widehat{\beta}_0 - \widehat{\beta})$$
$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p}$$

Next, define

$$U = \begin{pmatrix} e_1^2/\widehat{\sigma}^2 \\ e_2^2/\widehat{\sigma}^2 \\ \cdots \\ e_n^2/\widehat{\sigma}^2 \end{pmatrix} \quad w_i' = \begin{pmatrix} \frac{\partial w_i}{\partial \lambda_1} \\ \frac{\partial w_i}{\partial \lambda_2} \\ \cdots \\ \frac{\partial w_i}{\partial \lambda_q} \end{pmatrix} \quad D = \begin{pmatrix} (w_1')^T \\ (w_2')^T \\ \cdots \\ (w_n')^T \end{pmatrix}$$

so D is a $n \times q$ matrix. Finally, let

$$\bar{D} = D - \mathbf{1}\mathbf{1}^T D/n$$

be the $n \times q$ matrix obtained from D subtracting its column means.

Then the **Cook-Weisberg score statistic** id defined as (exercise)

$$S = \frac{1}{2} U^T \bar{D} (\bar{D}^T \bar{D})^{-1} \bar{D}^T U = \frac{1}{2} U^T P_{\bar{D}} U \tag{6.11}$$

with the assumption that $\bar{D}$ is of full rank. The asymptotic distribution of S is

$$S \to_d \chi_q^2$$

Thus, reject H_0 if

$$S > \chi_{q,\alpha}^2$$

where $\chi_{q,\alpha}^2$ is the upper α quantile of a χ_q^2 random variable.

In the special case when $q = 1$, i.e., z_i and λ are both scalar, an alternative to Cook-Weisberg score statistic is (exercise)

$$\widetilde{S} = \frac{Y^T (I - V) A (I - V) Y}{Y^T (I - V) Y} \tag{6.12}$$

where $V = P_{X_0} = X_0 (X_0^T X_0)^{-1} X_0^T$ and $A = \text{diag}(w_i')$. By simultaneous diagonalization theorem 1.1, we have

$$S' = \frac{\sum_{i=1}^{n-p} \mu_i \chi_i^2}{\sum_{i=1}^{n-p} \chi_i^2} \tag{6.13}$$

where χ_i^2's are iid χ_1^2 variables and μ_i's are, at most, $n-p$ nonzero eigenvalues of $(I-V)A(I-V)$. Denote the null distribution of $\widetilde{S}$ as $\mathbb{P}_{\widetilde{S}}$, then reject null if

$$\widetilde{S} > \mathbb{P}_{\widetilde{S},\alpha}$$

where $\mathbb{P}_{\widetilde{S},\alpha}$ is the upper α quantile of $\mathbb{P}_{\widetilde{S}}$.

6.7 Exercises

1. (Final exam) Consider the standard linear model where $\mathbb{E}y = X\beta$ and X is of full rank. Find a 95% confidence interval for a ratio of two given linear combinations of the model parameters $\frac{c^T\beta}{d^T\beta}$ where c and d are given vectors.

2. Find a 95% confidence interval for the unknown change point in a two phase linear regression.

3. (Hoadley's formula, [1]) Consider a simple linear regression and let $F = \widehat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 / s^2$, the F-statistic for testing $H_0 : \beta_1 = 0$ for a straight line. Using the notation of Section 6.1.5 in Seber and Lee [1], prove that

$$\widetilde{x}_0 - \bar{x} = \frac{F}{F + (n-2)} (\widehat{x}_0 - \bar{x})$$

4. Prove the claim 6.6.

5. Suppose we have a standard one-way ANOVA with k groups and n observations per group with the usual notation.

   (a) Show that the statistic

   $$T_{k,n} = \max_{1 \le i \le j \le k} \sqrt{n} |\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - (\mu_i - \mu_j)| / \widehat{\sigma}$$

   is distributed as the studentized range distribution and identify its degrees of freedom.

   (b) Find the asymptotic distribution of $T_{k,n}$ as n goes to ∞.

6. (Studentized maximum modulus distribution, [6]) Let $Z_1, \cdots, Z_k$ be i.i.d. $\mathcal{N}(0,1)$ and $U \sim \chi_m^2(0)$ and they are independent. Define

$$M = \max_{1 \le i \le k} \frac{|Z_i|}{\sqrt{U/m}}$$

We say that M has a studentized maximum modulus distribution and write

$$M \sim M_{k,m}$$

(a) Find simultaneous confidence intervals for the set of all $\mu_i = \theta + \alpha_i$ for an one-way ANOVA model with k groups and n observations in each group.

Hint: If $Z = (Z_1, \cdots, Z_k)$ has distribution

$$Z \sim \mathcal{N}_k(0, \Sigma)$$

Then we still have [7]

$$\mathbb{P}\left(\max_{1\le i\le k} \frac{|Z_i|}{\sqrt{U/m}} \ge M_{k,m,\alpha}\right) \le \alpha$$

where $M_{k,m,\alpha}$ is the upper α quantile of $M_{k,m}$.

(b) Let $a_1, a_2, \cdots, a_k$ be a set of numbers. Is it true that $\max_i |a_i| \le c$ iff $|\sum_{i=1}^k d_i a_i| \le c\sum_{i=1}^k |d_i|$ for all numbers $d_1, d_2, \cdots, d_k$?

(c) Use (b) and the studentized maximum modulus distribution to find a simultaneuous set of confidence intervals for $\sum_{i=1}^k d_i\mu_i$.

7. Refer to the Cook and Weisberg's paper **Diagnostics for Heteroscedasticity in Regression** [5]. Derive the test statistic for testing heteroscedasticity using the score test for the situation described in the paper and see whether our results agree with equations (8), (9) and (10).

8. Let $X \sim \chi^2_n$. Find $\mathbb{E}\sqrt{X}$ and hence find an unbiased estimate of σ in the standard linear model $\mathbb{E}y = X\beta$ where X is $n \times p$ with full column rank and error terms are independent, each with mean 0 and variance σ^2.

9. Suppose we have n data points with independent observations $y_1, \cdots, y_n$ from the simple linear model defined on a compact interval X:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + e_i, x \in X \\ e_i &\sim \mathcal{N}(0, \sigma^2), i = 1, \cdots, n \end{aligned}$$

(a) Find an unbiased estimator for β_0, β_1 and σ.

(b) Show that if $n > 2$, a $100(1-\alpha)\%$ confidence interval for $\xi = \sigma/\beta_1$ is

$$\widehat{\xi} \mp z_{\frac{\alpha}{2}}\widehat{\xi}\sqrt{\frac{1}{xn-4} + \frac{\widehat{\xi}^2}{s_x^2}}$$

where $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$, $\widehat{\xi} = \widehat{\sigma}/\widehat{\beta}_1$, and $z_{\frac{\alpha}{2}}$ is the upper α percentile of the standard normal.

7 Shrinkage and Bayes Estimation

In this section, we study different alternatives to the ordinary least square estimation. Principal component regression, ridge regression, LASSO, James-Stein estimator and Bayes estimator are included.

7.1 Principal component regression

Consider the linear model

$$\mathbb{E}y = X\beta$$

and X is a $n \times p$ matrix but the rank is $r < p$. In other words, we have a collinearity issue. Suppose the spectral decomposition of $X^T X$ is

$$U^T X^T X U = D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

and $\lambda_{r+1}, \cdots, \lambda_p = 0$. Re-write

$$\mathbb{E}y = X\beta = XUU^T\beta = XU_1\gamma_1$$

where U_1 is the first r columns of U and $\gamma_1 = U_1^T\beta$. $Z = XU_1$ are called the principal components of X. So we regress y on $z_1, \cdots, z_r$ and note that

$$z_i^T z_j = \delta_{ij} u_i^T X^T X u_j$$

where u_i is the i^{th} column of U_1 and $z_i = Xu_i$. A screen plot is the plot of

$$\frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{p} \lambda_p}, j = 1, \cdots, p$$

7.2 Ridge estimator

The mean square error (MSE) of $\widetilde{\beta}$ of β is

$$\text{MSE}(\widetilde{\beta}) = \mathbb{E}\left(\widetilde{\beta} - \beta\right)\left(\widetilde{\beta} - \beta\right)^T = Cov(\widetilde{\beta} + bb^T)$$

where $b = \mathbb{E}\left(\widetilde{\beta} - \beta\right)$ is called bias. Clearly, for LSE we have $b(\widehat{\beta}) = 0$ and MSE$(\widehat{\beta}) = \sigma^2(X^T X)^{-1}$.

Definition 7.1 (Ridge estimator)**.** The ridge estimator with parameter κ of β is defined as

$$b_\kappa = \left(X^T X + \kappa I\right)^{-1} X^T y, \ \kappa > 0 \tag{7.1}$$

Let $G_\kappa = (X^TX + \kappa I)^{-1}$, then

$$\mathbb{E}b_\kappa = G_\kappa X^TX\beta = \left(I + \kappa(X^TX)^{-1}\right)^{-1}\beta$$

Thus, by Woodbury identity,

$$\begin{aligned}\text{bias}(b_\kappa) &= -\kappa\left(I + \kappa(X^TX)^{-1}\right)^{-1}\left(X^TX\right)^{-1}\beta \\ &= -\kappa\left(\kappa I + (X^TX)\right)^{-1}\beta \\ &= -\kappa G_\kappa\beta\end{aligned}$$

and

$$Var(b_\kappa) = \sigma^2 G_\kappa X^TXG_\kappa$$

Thus,

$$\text{MSE}(b_\kappa) = G_\kappa\left(\sigma^2X^TX + \kappa^2\beta\beta^T\right)G_\kappa \tag{7.2}$$

It can be shown (exercise)

$$Tr\left(\text{MSE}(\widehat{\beta}) - \text{MSE}(b_\kappa)\right) \geq 0 \tag{7.3}$$

for some κ. Let the SVD of X be

$$X_{n\times p} = U_{n\times p}D_{p\times p}V^T_{p\times p}$$

Let u_i be the i^{th} column of U, then

$$\begin{aligned}\widehat{\theta} &= X\widehat{\beta} \\ &= UU^Ty \\ &= \sum_{i=1}^{p} u_iu_i^Ty \\ \widetilde{\theta} &= Xb_\kappa \\ &= X(X^TX + \kappa I)^{-1}X^Ty \\ &= \sum_{i=1}^{p}\frac{\lambda_i^2}{\lambda_i^2 + \kappa}u_iu_i^Ty\end{aligned}$$

Since $\frac{\lambda_i^2}{\lambda_i^2+\kappa} < 1$, so Xb_κ shrinks $X\widehat{\beta}$.

7.3 James-Stein estimator

For technical details, please see Chapter 11 of [8]. Suppose

$$y \sim \mathcal{N}_n(\mu, \sigma^2 I),\ \mu \in \mathcal{V}, \sigma^2 > 0$$

where $\mathcal{V}$ is a p-dimensional vector space of $\mathbb{R}^n$. Only in this subsection, I use P to denote the projection operator $\mathcal{P}_{\mathcal{V}}$ onto $\mathcal{V}$ and Q is the abbreviation for $\mathcal{Q}_{\mathcal{V}}$. The LSE is

$$\widehat{\mu} = Py$$

and

$$\widehat{\sigma}^2 = y^T Qy/(n-p)$$

However, $\widehat{\mu}$ over-estimates the square of the norm of μ:

$$\|\widehat{\mu}\|_2^2 = \|\mu\|_2^2 + \sigma^2 \mathrm{Tr}(P) = \|\mu\|_2^2 + p\sigma^2 > \|\mu\|_2^2$$

Consider shrinking $\widehat{\mu}$ and consider estimators of the form

$$\widehat{\widehat{\mu}} = \left(1 - c\frac{\widehat{\sigma}^2}{\|\mu\|_2^2}\right)\widehat{\mu}$$

where c is a constant to be determined. In 1960, James and Stein provided one choice of c:

Definition 7.2 (James-Stein (JS) estimator)**.** The JS estimator of μ is

$$\widehat{\widehat{\mu}} = \left(1 - \frac{(p-2)(n-p)}{n-p+2}\frac{\widehat{\sigma}^2}{\|\mu\|_2^2}\right)\widehat{\mu} \tag{7.4}$$

That is,

$$c = \frac{(p-2)(n-p)}{n-p+2}$$

They also provided the following theorem:

Theorem 7.1 (Risk of JS estimator, [8])**.** The estimator $\widehat{\widehat{\mu}}$ is better than $\widehat{\mu}$. Its risk function is

$$R(\widehat{\widehat{\mu}}; (\mu, \sigma^2)) = p - \frac{(p-2)^2(n-p)}{n-p+2}\mathbb{E}\left(\frac{1}{p-2+2K}\right) < p = R(\widehat{\mu}; (\mu, \sigma^2)) \tag{7.5}$$

where K has a Poisson distribution with mean $\frac{\|\mu\|_2^2}{2\sigma^2}$. In particular,

$$R(\widehat{\widehat{\mu}}; (0, \sigma^2)) = \frac{2n}{n-p+2}$$

For proof of theorem 7.1, see [8].

7.4 Bayes estimator

As Dr. Wong mentioned, in 1970s and 1980s, there is a huge debate between Frequentism and Bayesianism. But starting at 1990s, it seems that Bayesianism beats Frequentism. Therefore, in this subsection we talk about Bayes estimation in linear models. Let $f(y|\theta)$ be the conditional density of y and $g(\theta)$ be the prior density of θ. Then the likelihood function of θ is

$$L(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

The posterior distribution of θ is

$$f(\theta|y) = \frac{f(y|\theta)g(\theta)}{\int_\Theta f(y|\theta)g(\theta)d\theta} \propto f(y|\theta)g(\theta)$$

Consider a linear model:

$$y = X\beta + \epsilon,\ \epsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

In the following examples, we put different priors on the parameter $\theta^T = (\beta^T, \log\sigma)$. Recall

$$f(\theta|y) \propto f(y|\theta)g(\theta),\ g(\theta) = g_1(\beta|\sigma)g_2(\sigma)$$

and

$$y|\theta \sim \mathcal{N}(X\beta, \sigma^2 I)$$

Example 7.1 (Non-informative prior)**.** We need formulas 7.6, 7.7 and 7.8 (exercises).

The non-informative prior is

$$g(\theta) = g_1(\beta)g_2(\sigma^2) \propto \frac{1}{\sigma}$$

Thus,

$$f(\theta|y) \propto \frac{1}{\sigma}(\frac{1}{\sqrt{2\pi\sigma}})^n \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right)$$

If interest is in β along (exercises),

$$\begin{aligned}
f(\beta|y) &= \int_0^\infty \frac{1}{(\sqrt{2\pi})^n}\frac{1}{\sigma^{n+1}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right) d\sigma^2 \\
&= \frac{1}{2}\left(\frac{\|y - X\beta\|_2^2}{2}\right)^{-\frac{n}{2}} \Gamma(\frac{n}{2}) \\
&\propto \|y - X\beta\|_2^{-n} \\
&= \left((n-p)s^2 + \|X\widehat{\beta} - X\beta\|\right)^{-\frac{n}{2}} \\
&\propto \left(1 + \frac{(\widehat{\beta} - \beta)^T X^T X(\widehat{\beta} - \beta)}{(n-p)s^2}\right)^{-\frac{n}{2}}
\end{aligned}$$

In conclusion,

$$\beta|y \sim t_p((n-p), \widehat{\beta}, (X^TX/s^2)^{-1})$$

Definition 7.3 (Multivariate t-distribution)**.** We call y has a Multivariate t-distribution and write

$$y \sim t_m(\nu, \mu, \Sigma)$$

if the density of y is

$$f_Y(y) = \frac{\Gamma(\frac{\nu+m}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{\frac{m}{2}}} \frac{1}{|\Sigma|^{1/2}} \left(1 + \frac{(y-\mu)^T\Sigma^{-1}(y-\mu)}{\nu}\right)^{-\frac{\nu+m}{2}}$$

For details, see p475 of [1].

Example 7.2 (Conjugate prior)**.** We need formula 7.9 (exercise).

We put a conjugate prior on $\theta^T = (\beta^T, \sigma^2)$, i.e.,

$$g(\beta, \sigma^2) = g(\beta|\sigma^2)g(\sigma^2)$$

where $\beta|\sigma^2 \sim \mathcal{N}_p(m, \sigma^2 V)$ and σ^2 has an inverted gamma density (7.10) with hyper-parameters $(d, \frac{a}{2})$. Thus,

$$f(\sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{d+2}{2}}} \exp(-a/2\sigma^2)$$

The posterior of θ is proportional to

$$\begin{aligned} f(\theta|y) &\propto f(y|\theta)g(\theta) \\ &\propto (\sigma^2)^{-\frac{n+d+p+2}{2}} \exp\left(-\frac{1}{2\sigma^2}(Q+a)\right) \end{aligned}$$

where

$$Q = (y - X\beta)^T(y - X\beta) + (\beta - m)^T V^{-1}(\beta - m)$$

If interest is in β alone, then (exercise)

$$\begin{aligned} f(\beta|y) &\propto \int_0^\infty (\sigma^2)^{-\frac{n+d+p+2}{2}} \exp\left(-\frac{1}{2\sigma^2}(Q+a)\right) d\sigma^2 \\ &\propto (Q+a)^{-\frac{d+n+p}{2}} \\ &\propto \left(1 + \frac{1}{n+d}(\beta - m_*)^T W_*^{-1}(\beta - m_*)\right)^{-\frac{d+n+p}{2}} \end{aligned}$$

For definition of m_* and W_*, see exercise 7.9. In conclusion,

$$\beta|y \sim t_p(n+d, m_*, W_*)$$

7.5 Exercises

1. Prove inequality 7.3 for some κ.

2. Let X be a p-dimensional random vector with covariance matrix Σ. Let $u = (u_1, u_2, \cdots, u_p)^T$ be a vector of principal components of X. Then $u_i = a_i^T X$ for some vector a_i of length 1, $i = 1, 2, \cdots, p$. Show that

   (a) $Var(a^T X) \leq Var(u_1)$.

   (b) If $a^T X$ is uncorrelated with $u_1, u_2, \cdots, u_{i-1}$, then $Var(a^T X) \leq Var(U_1)$.

3. Let μ belong to $\mathcal{V}$, a p-dimensional vector space of $\mathbb{R}^n$, let y belong to $\mathbb{R}^n$ and let $P_{\mathcal{V}}$ be the orthogonal projection onto $\mathcal{V}$. If $\widehat{\mu} = P_{\mathcal{V}} y$ and $\widehat{\sigma}^2 = \|(I - P_{\mathcal{V}})y\|_2^2/(n-p)$, show that

   (a) $\mathbb{E}\|\widehat{\mu}\|_2^2$ always over-estimates $\|\mu\|_2^2$.

   (b) $\mathbb{E}\|\widehat{\mu} - \mu\|_2^2/\sigma^2 = p$ and identify the distribution of $\|\widehat{\mu} - \mu\|_2^2$.

   (c) $\mathbb{E}\widehat{\sigma}^4 = (n-p+2)\sigma^4/(n-p)$.

4. Recall that if X and Y are random variables with finite means and variances, then

$$\mathbb{E}\left(\mathbb{E}(X|Y)\right) = \mathbb{E}X$$

   Use this result to show that if the conditional distribution of $V|K \sim \chi^2_{p+2K}(0)$, then

$$\mathbb{E}\left(\frac{1}{V}|K\right) = \frac{1}{p-2+2K}$$

5. (Beta-Binomial) Let $Y \sim$ Bin(n, θ), that is, $\mathbb{P}(Y = y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$. Let $\theta \sim$ Beta(α, β). Find the posterior of θ.

6. (Normal sequence) Suppose the conditional density of $y|\mu$ is $\mathcal{N}(\mu, \sigma^2)$. If μ is univariate normal with mean μ_0 and variance σ_0^2, then

   (a) Find the conditional density of $\mu|y$.

   (b) Generalize the setting when we have a random sample $y_1, \cdots, y_n$.

7. Prove the following:

   (a) Recall $\Gamma(x) = \int_0^\infty \exp(-t)t^{x-1}dt$, then show

$$\int_0^\infty \exp(-\frac{k}{x})x^{-\nu-1} = k^{-\nu}\Gamma(\nu) \tag{7.6}$$

   (b) Let a and b be non-negative, then show

$$\int_0^\infty \exp -\frac{a}{x^2}x^{-b-1}dx = \frac{1}{2}a^{-b/2}\Gamma(\frac{b}{2}) \tag{7.7}$$

8. Show that

$$\|y - X\beta\|_2^2 = (n-p)s^2 + \|X\widehat{\beta} - X\beta\|_2^2 \tag{7.8}$$

where s^2 is usual estimate of σ^2 and $\widehat{\beta}$ is the LSE.

9. Use binomial inversion theorem to show the equality:

$$\begin{aligned}&(y - X\beta)^T(y - X\beta) + (\beta - m)^T V^{-1}(\beta - m) = \\ &(\beta - m^*)^T V^{*-1}(\beta - m^*) + (y - Xm)^T(I + XVX^T)^{-1}(y - Xm)\end{aligned} \tag{7.9}$$

where $V^* = (X^TX + V^{-1})^{-1}$ and $m^* = V^*(X^Ty + V^{-1}m)$.

10. (Inverted gamma) Suppose X has an inverted gamma with density given by

$$f(x|a,b) = b^{-a}x^{-a-1}\exp -\frac{b}{x}/\Gamma(a) \tag{7.10}$$

where $a > 0$ and $b > 0$ and Γ is the the Gamma function.

    (a) Describe how values of a and b affect $f(x|a,b)$ in terms of shape, skewness and symmetry.
    (b) Find its mean and describe how this density is typically used in estimating paramters in a linear model using a Bayesian paradigm.

11. Suppose Z has a density proportional to $z^{n/2-1}\exp -nz/2$ and conditional on Z, $X|Z = z$ is multivariate normal with mean 0 and covariance $\frac{1}{z}I_k$.

    (a) What is the density of X?
    (b) Determine the mean and covariance matrix of X.

12. (posterior mean of β, [1]) Using the non-informative prior for $\theta^T = (\beta^T, \log\sigma)$, show that the conditional posterior density $f(\beta|y,\sigma)$ is multivariate normal. Hence deduce that the posterior mean of β is $\widehat{\beta}$, the LSE.

13. Read the paper on **Bayes Estimation of Two-Phase Linear Regression Model** [9].

14. (Posterior mean of σ^2, [1]) Suppose that we use the non-informative prior for θ.

    (a) Obtain an expression proportional to $f(\beta, \sigma^2|y)$.
    (b) Integrate out β to obtain
    $$f(\sigma^2|y) \propto (\sigma^2)^{-\frac{\nu}{2}-1}\exp\left(-\frac{a}{v}\right)$$
    where $\nu = n - p$ and $a = \|y - X\widehat{\beta}\|_2^2/2$.
    (c) Find the posterior mean of σ^2.

8 ANOVA Mixed Models

Random effect, fixed effect and mixed effect ANOVA models are included.

8.1 One-way ANOVA with random effects

Recall the one-way ANOVA with fixed effects:

$$\begin{aligned}
y_{ij} &= \mu + \tau_i + \epsilon_{ij} \\
\epsilon_{ij} &\sim_{\text{iid}} \mathcal{N}(0, \sigma^2) \\
\tau_i &= \text{ effect of treatment } i
\end{aligned}$$

We want to test

$$\begin{aligned}
&H_0 : \tau_1 = \tau_2 = \cdots = \tau_k \\
&H_1 : \text{They are not the same.}
\end{aligned}$$

When interest is in more than k treatments, i.e., interest is in a population of treatment, then we sample k of them, suggesting that

$$\begin{aligned}
\tau_i &\sim \mathcal{N}(0, \sigma_\tau^2) \\
\epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \\
\tau_i &\perp\!\!\!\perp \epsilon_{ij} \ \forall i \& j
\end{aligned}$$

Here τ_i's are called random effects. Hypothesis of interest is

$$H_0 : \sigma_\tau^2 = 0 \quad H_1 : \sigma_\tau^2 > 0$$

and we have data $\{y_{ij} : i = 1, \cdots, k; j = 1, \cdots, n\}$.

$$\begin{aligned}
\sum_i \sum_j (y_{ij} - \bar{y})^2 &= \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}) \\
&= \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_i \sum_j (\bar{y}_{i\cdot} - \bar{y})^2 \\
&= \text{SSE} + \text{SSTr}
\end{aligned}$$

Since $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, then

$$\text{SSTr} = \sum_i^k \sum_j^n (\tau_i - \bar{\tau} + \bar{\epsilon}_{i\cdot} - \bar{\epsilon})^2$$
$$\text{SSE} = \sum_i^k \sum_j^n (\epsilon_{ij} - \bar{\epsilon}_{i\cdot})^2$$

and

$$\mathbb{E}\,(\text{SSTr}) = (k-1)(n\sigma_\tau^2 + \sigma^2)$$
$$\mathbb{E}\,(\text{SSE}) = n(k-1)\sigma^2$$

Example 8.1 (Point estimation and confidence interval)**.** The point estimation for $\widehat{\sigma}_\tau^2$ is

$$\widehat{\sigma}_\tau^2 = \frac{\text{MSTr} - \text{MSE}}{n}$$

where $\text{MSTr} = \frac{\text{SSTr}}{k-1}$ and $\text{MSE} = \frac{\text{SSE}}{n(k-1)}$. Also,

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{k(n-1)}$$

then a $(1-\alpha)$ confidence interval for σ^2 is

$$\frac{\text{SSE}}{\chi^2_{k(n-1),1-\frac{\alpha}{2}}} < \sigma^2 < \frac{\text{SSE}}{\chi^2_{k(n-1),\frac{\alpha}{2}}} \tag{8.1}$$

and

$$\frac{(k-1)\text{MSTr}}{n\sigma_\tau^2 + \sigma^2} \sim \chi^2_{k-1}$$

indicates a $(1-\alpha)$ confidence interval for σ_τ^2

$$\frac{\text{SSTr}}{\chi^2_{k-1,1-\frac{\alpha}{2}}} < \sigma_\tau^2 < \frac{\text{SSTr}}{\chi^2_{k-1,\frac{\alpha}{2}}} \tag{8.2}$$

Further, MSE $\perp\!\!\!\perp$ MSTr implies

$$\frac{\text{MSTr}}{\text{MSE}} \frac{\sigma^2}{n\sigma_\tau^2 + \sigma^2} \sim F_{k-1,(n-1)k}$$

Thus, if we are interested in the ratio:

$$\theta = \frac{\sigma_\tau^2}{\sigma^2}$$

then a $100(1-\alpha)\%$ confidence interval is

$$\left(\frac{1}{F_{k-1,(n-1)k,1-\frac{\alpha}{2}}}\frac{\text{MSTr}}{\text{MSE}}-1\right)\frac{1}{n} \le \theta \le \left(\frac{1}{F_{k-1,(n-1)k,\frac{\alpha}{2}}}\frac{\text{MSTr}}{\text{MSE}}-1\right)\frac{1}{n} \tag{8.3}$$

The intraclass correlation coefficient for effect τ is defined as

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2+\sigma^2}$$

So a $100(1-\alpha)\%$ confidence interval for ρ is

$$\frac{l}{1+l} \le \rho \le \frac{u}{1+u}$$

where l and u are confidence limits for θ and noting that $\rho = \frac{\theta}{\theta+1}$.

8.2 Two factor ANOVA

Suppose we have two factors A and B, each with a and b levels. Then the ANOVA model is

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$$

For fixed effects, we have, for example, constraint on τ:

$$\sum_{i=1}^{a} \tau_i = 0$$

For random effects, we assume, say,

$$\tau_i \sim \mathcal{N}(0, \sigma_\tau^2),\ i = 1, \cdots, a$$

so τ_i's are independent but not identically distributed.

Example 8.2 (A fixed, B fixed)**.** see figure 2.

Example 8.3 (A random, B random)**.** see figure 3.

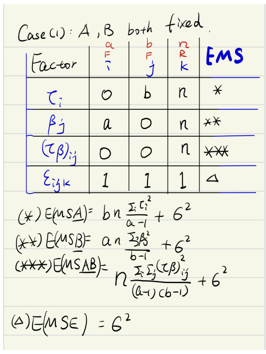


Figure 2: Fixed effect.

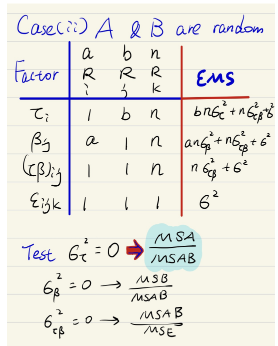


Figure 3: Random effect.

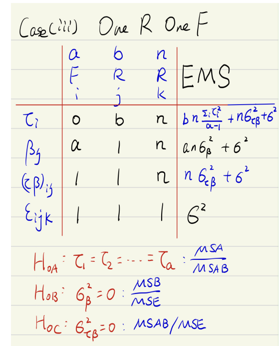


Figure 4: Mixed effect.

Example 8.4 (A fixed, B random)**.** see figure 4. The assumptions for $(\tau\beta)$ are:

$$
(\tau\beta)_{ij} \sim \mathcal{N}(0, \frac{a-1}{a}\sigma^2_{\tau\beta})
$$

$$
\sum_{i=1}^{a} (\tau\beta)_{ij} = 0
$$

8.3 Satterwaite approximation

The Satterwaite approximation (a.k.a. Welch-Satterwaite equation) is a powerful tool to derive approximate F-test in random effect models. It is also used to solve the famous Berhens-Fisher problem (exercise).

Theorem 8.1 (Satterwaite approximation)**.** Let $U_1, U_2, \cdots, U_k$ be independent χ^2 random variables with $r_1, r_2, \cdots, r_k$ degrees of freedom respectively. Define

$$
U = \sum_{i=1}^{k} a_i U_i
$$

where $a_1, \cdots, a_k$ is a sequence of positive numbers. Then we have

$$
\frac{bU}{\mathbb{E}(U)} \sim \chi^2_b
$$

where

$$
b = \frac{(\mathbb{E}U)^2}{\sum_{i=1}^{k} \frac{(a_i \mathbb{E}U_i)^2}{r_i}} = \frac{\left(\sum_{i=1}^{k} a_i r_i\right)^2}{\left(\sum_{i=1}^{k} a_i^2 r_i\right)}
$$

Therefore, by method of moments, we can approximate b by

$$
\widehat{b} = \frac{(\sum_{i=1}^{k} a_i U_i)^2}{\sum_{i=1}^{k} (a_i^2 U_i^2 / r_i)} \tag{8.4}
$$

Proof. Suppose we want to find constant v s.t.

$$\frac{vU}{\mathbb{E}(U)} \sim \chi_b^2$$

Take expectation on both sides:

$$\mathbb{E}\left(\frac{vU}{\mathbb{E}U}\right) = b \Rightarrow v = b$$

Take variance on both sides:

$$\begin{aligned}
2b &= Var(\chi_b^2) \\
&= Var\left(\frac{vU}{\mathbb{E}(U)}\right) \\
&= \frac{v^2}{(\mathbb{E}U)^2}\sum_{i=1}^{k} a_i^2 Var(U_i) \\
&= \frac{v^2}{(\mathbb{E}(U))^2}\sum_{i=1}^{k}(2r_i a_i^2)
\end{aligned}$$

Replace v with b gives the desired result. □

8.4 Exercises

1. (Behrens-Fisher problem) Show that the statistic for testing equality of means from two normal populations with different variances using sample size n_1 and n_2 has a t-distribution with degrees of freedom **approximately** equal to

$$\frac{(g_1+g_2)^2}{g_1^2/(n_1-1)+g_2^2/(n_2-1)}$$

   where $g_i = s_i^2/n_i, i = 1, 2$ and s_1^2 and s_2^2 are the sample variances.

2. (Final exam) Suppose we have an **unbalanced** one-way ANOVA setting with the usual assumptions, except that we do not assume the variances of the responses from each group are equal. Derive a statistical procedure to test whether a given linear combination of the group means is zero.

3. (Three factor mixed ANOVA model) Suppose we conduct an experiment with 3 factors A, B and C. The two factors A, B are fixed and C is a random factor. The number of levels for each of the factors are a, b and c, respectively, with n observations per cell.

   (a) Find the expected mean square errors (EMS) for all the effects, see figure 2, 3 and 4.

   (b) Provide the test statistics for testing all effects in the model and state the rejection rules. This is known as approximate F-test.

9 Linear Mixed Models

Consider a linear mixed model (LMM):

$$Y = X\beta + Z\gamma + \epsilon$$
$$\mathbb{E}\epsilon = 0,\ Cov(\epsilon) = R$$

and

- β: an unknown vector of fixed effects.
- γ: a vector of random effects with mean 0 covariance D.
- X, Z are known and $Cov(\gamma, \epsilon)$=0.

We have

$$\mathbb{E}Y = X\beta;\ Cov(Y) = ZDZ^T + R := V$$

In this section, we discuss the Henderson's equation for solving MLE in LMMs. The Best Linear Unbiased Predictor (BLUP) theorem is also proved.

9.1 Henderson's equation

To find MLEs, we assume

$$\gamma \sim \mathcal{N}(0, D);\ \epsilon \sim \mathcal{N}(0, R)$$

Write down the likelihood function for Y and γ in two steps:

$$\begin{aligned} f_{y,\gamma}(y, \gamma) &= f_{y|\gamma}(y) f_\gamma(\gamma) \\ &\propto \frac{1}{|R|^{1/2}} \exp\left(-\frac{1}{2}(y - X\beta - Z\gamma)^T R^{-1}(y - X\beta - Z\gamma)\right) \times \\ &\quad \frac{1}{|D|^{1/2}} \exp\left(-\frac{1}{2}\gamma^T D^{-1}\gamma\right) \end{aligned} \tag{9.1}$$

Differentiate $\ln f_{y,\gamma}(y, \gamma)$ w.r.t. β and γ results

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & D^{-1} + Z^T R^{-1} Z \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{pmatrix} \tag{9.2}$$

This is the set of Henderson's equations. It can be shown (exercise) the solution is

$$\begin{aligned} \widehat{\beta}_w &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \\ \widehat{\gamma} &= D Z^T V^{-1}(y - X\widehat{\beta}_w) \end{aligned} \tag{9.3}$$

A hint is

$$V^{-1} = (ZDZ^T + R)^{-1} = R^{-1}Z(D^{-1} + Z^T R^{-1} Z)^{-1} Z^T R^{-1}$$

and note that $\widehat{\beta}_w$ is unbiased for β and $\mathbb{E}\widehat{\gamma} = 0$.

9.2 Best linear unbiased predictor (BLUP)

A recall of generalized linear model:

$$y = X\beta + \epsilon, \ \text{rank}(X) = p$$

where $\mathbb{E}(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 V$. Then we have the weighted LSE:

$$\widehat{\beta}_w = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

Definition 9.1 (Linear unbiased predictor (LUP))**.** Suppose u is a random variable with mean 0 and finite variance. A linear predictor $d + a^T y$ of $c^T\beta + u$ is **unbiased** iff

$$\mathbb{E}(d + a^T y) = \mathbb{E}(c^T\beta + u) = c^T\beta$$

or $c^T\beta + u$ is **predictable** iff $\exists$ an LUP of $c^T\beta + u$.

Lemma 10. *$c^T\beta + u$ is predictable iff*

$$\exists \ a \ \textit{s.t.} \ c = X^T a$$

This implies $d + a^T y$ is an LUP of $c^T\beta + u$ iff

$$d = 0, \ c = X^T a$$

Definition 9.2 (BLUP)**.** Let $\widehat{u}$ be a predictor of u. If $\widehat{u}$ satisfies the following three requirements, we call it a <u>best unbiased linear predictor</u> (BLUP).

- $\widehat{u}$ is a linear function of y.
- $\widehat{u}$ is unbiased for u.
$$\mathbb{E}(\widehat{u} - u) = 0$$
- If v is any LUP, then
$$Var(\widehat{u} - u) \leq Var(v - u)$$

Recall the MSE of $\widehat{\theta}$ of θ is

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta} - \theta)^2 = Var(\widehat{\theta}) + (\text{Bias}(\widehat{\theta}))^2$$

Theorem 9.1 (BLUP)**.** Consider a generalized linear model:

$$y = X\beta + \epsilon$$

A linear mixed model can be written as a generalized linear model where ϵ is replaced by

$$Z\gamma + \epsilon$$

and $V = ZDZ^T + R$. Assume u is a variable with mean 0 and finite variance and

$$\begin{aligned}\mathbb{E}(\epsilon) &= 0 \\ Cov(\epsilon) &= \sigma^2 V \\ Cov(\epsilon, u) &= \sigma^2 K\end{aligned}$$

where K is a known vector. Note we have no assumption of normality. Then $c^T\widehat{\beta}_w + \widehat{u}$ has the smallest MSE among all LUP of $c^T\beta + u$, where

$$\begin{aligned}\widehat{u} &= K^T V^{-1}\left(y - X\widehat{\beta}_w\right) \\ \widehat{\beta}_w &= (X^T V^{-1} X)^{-1} X^T V^{-1} y\end{aligned}$$

That is, $c^T\widehat{\beta}_w + \widehat{u}$ is BLUP of $c^T\beta + u$.

Proof. Note that

$$c^T\widehat{\beta}_w + \widehat{u} = c^T\widehat{\beta}_w + K^T V^{-1}(y - X\widehat{\beta}_w)$$

has expectation

$$c^T\beta + 0 = \mathbb{E}(c^T\beta + u)$$

Further,

$$\begin{aligned}c^T\widehat{\beta}_w + \widehat{u} &= (c^T + K^T V^{-1} X)\widehat{\beta} + K^T V^{-1} y \\ &= \left[(c^T - K^T V^{-1} X)(X^T V^{-1} X)^{-1} X^T V^{-1} + K^T V^{-1}\right] y \\ &= b^T y\end{aligned}$$

where $b^T = (c^T - K^T V^{-1} X)(X^T V^{-1} X)^{-1} X^T V^{-1} + K^T V^{-1}$. Thus, $c^T\widehat{\beta}_w + \widehat{u}$ is **linear** in y. It is also **unbiased** for $c^T\beta + u$, we have

$$c = X^T b$$

If $a^T y$ is any other LUP of $c^T\beta + u$, we have

$$c = X^T a$$

To derive MSE$(a^T y)$, note that

$$Cov((a-b)^T y, b^T y - u) = \sigma^2 (a-b)^T (Vb - K)$$

But

$$Vb - K = X(X^T V^{-1} X)^{-1}(c - X^T V^{-1} K)$$

Thus,

$$\begin{aligned}
Cov((a-b)^T y, b^T y - u) &= \sigma^2 (a-b)^T (Vb - K) \\
&= \sigma^2 \left[(a-b)^T X\right] \left[(X^T V^{-1} X)^{-1}(c - X^T V^{-1} K)\right] \\
&= 0
\end{aligned}$$

Finally,

$$\begin{aligned}
\text{MSE}(a^T y) &= \mathbb{E}\left(a^T y - (c^T\beta + u)\right)^2 \\
&= Var(a^T y - u) \\
&= Var(a^T y - b^T y + b^T y - u) \\
&= Var(a^T y - u) + 2Cov((a-b)^T y, b^T y - u) + Var(b^T y - u) \\
&= Var(a^T y - u) + Var(b^T y - u) \\
&= Var(a^T y - u) + Var(b^T y - (c^T\beta + u)) \\
&= Var((a-b)^T y) + \text{MSE}(b^T y) \\
&\geq \text{MSE}(b^T y) \qquad (9.4)
\end{aligned}$$

□

Recall the solution to Henderson's equation 9.3, we have

$$\widehat{\gamma} = DZ^T V^{-1}(y - X\widehat{\beta})$$

and note

$$\begin{pmatrix} \gamma \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} D & ? \\ ? & V \end{pmatrix}\right)$$

where

$$? = Cov(\gamma, y) = DZ^T$$

Thus (exercise),

$$\gamma | y \sim \mathcal{N}_q\left(DZ^T V^{-1}(y - X\beta), D - DZ^T V^{-1} ZD\right)$$

The **BLUP** of $\widehat{\gamma}$ (elementwise) is indeed

$$\widehat{\mathbb{E}(\gamma|y)} = DZ^T V^{-1}(y - X\widehat{\beta}_w) \tag{9.5}$$

Example 9.1 (BLUP for one-way ANOVA)**.** Assume for $i = 1, 2, 3$ and $j = 1, 2$,

$$\begin{aligned}
y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\
y &= \mathbf{1}_6\mu + Z\alpha + \epsilon \\
R &= Cov(\epsilon) = \sigma^2 I_6 \\
D &= Cov(\alpha) = \sigma_\alpha^2 I_3 \\
V &= ZDZ^T + R
\end{aligned}$$

where $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$, $\mathbf{1}_6$ is a 6-dimensional vector of all ones and

$$Z = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

In this case, the BLUP for α is

$$\widehat{\alpha} = DZ^T V^{-1}(y - X\widehat{\mu}_w)$$

where

$$\widehat{\mu}_w = \frac{\mathbf{1}^T V^{-1} y}{\mathbf{1}^T V^{-1}\mathbf{1}}$$

9.3 Restricted maximum likelihood (REML)

Let the linear model be

$$\begin{aligned}
Y &= X\beta + \epsilon \\
\mathbb{E}\epsilon &= 0,\ Cov(\epsilon) = V(\theta)
\end{aligned}$$

and

- β: an unknown vector of fixed effects.
- $V(\theta)$: the variance component depending on a parameter θ.

- θ: an unknown vector of parameters.

Suppose we are only interested in estimating $V(\theta)$. Thus, β is the **nuisance parameter** and will cause extra variance when estimating $V := V(\theta)$. In 1974, David Harville proposed the so-called restrcited maximum likelihood (REML) method [10]. In the original paper, David omitted details obtaining the REML. However, Lynn LaMotte provided a detailed derivation in 2007 [11]. In this subsection, I use LaMotte's derivation to get REML.

Let A be a matrix such that $A^TX = 0$, a particular choice of A would be

$$A = Q_X = I - X(X^TX)^{-1}X^T$$

Then we have

$$\mathbb{E}(AY) = 0,\ Var(AY) = AVA^T$$

and

$$L(\theta|y) = (2\pi)^{-\frac{n-p}{2}}|A^TVA|^{-\frac{1}{2}}\exp\{-\frac{1}{2}y^TA(A^TVA)^{-1}A^Ty\} \tag{9.6}$$

This is the restricted maximum likelihood function.

Theorem 9.2 (Re-parametrization of REML, [10], [11])**.** The REML of 9.6 can be re-written as

$$L(\theta|y) = (\text{Const.})|V|^{-\frac{1}{2}}|X^TV^{-1}X|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(y - X\widehat{\beta}_w)^TV^{-1}(y - X\widehat{\beta}_w)\right\} \tag{9.7}$$

with $\widehat{\beta}_w = (X^TV^{-1}X)^{-1}X^TV^{-1}y$ being the weighted LSE. In the special case $A^TA = I$,

$$L(\theta|y) = (2\pi)^{-\frac{n-p}{2}}|X^TX|^{\frac{1}{2}}|V|^{-\frac{1}{2}}|X^TV^{-1}X|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(y - X\widehat{\beta}_w)^TV^{-1}(y - X\widehat{\beta}_w)\right\}$$

LaMotte used the following two lemmas to show 9.7. Here I provide an easier proof of the first lemma and copy the original proof of the second lemma, which is first shown in Searle, 1979 [11].

Lemma 11. *If V is an $n \times n$ positive definite matrix and X and $A = Q_X$, then*

$$V^{-1} = V^{-1}X\left(X^TV^{-1}X\right)^{-1}X^TV^{-1} + A\left(A^TVA\right)^{-1}A^T$$

Proof. Left-multiplying $V^{\frac{1}{2}}$ and then right-multiplying $V^{\frac{1}{2}}$ on both sides, it is enough to show

$$I = P_{V^{-\frac{1}{2}}X} + P_{V^{\frac{1}{2}}A}$$

But this follows directly from the fact $V = V(\theta)$ is non-singular and

$$A^TX = 0,\ \mathcal{C}(X) \cup \mathcal{C}(A) = \mathbb{R}^n$$

□

Lemma 12. *If V is an $n \times n$ positive definite matrix, and (X, A) is an $n \times n$ matrix with full column rank, and $A^T X = 0$, then*

$$|V| = \frac{|A^T V A||X^T X|}{|A^T A||X^T V^{-1} X|}$$

Proof. Since (X, A) is a square matrix,

$$\begin{aligned}
|A^T A||X^T X||V| &= \left|\begin{pmatrix} X^T X & 0 \\ 0 & A^T A \end{pmatrix}\right| |V| \\
&= |(X, A)^T V (X, A)| \\
&= \left|\begin{pmatrix} X^T V X & X^T V A \\ A^T V X & A^T V A \end{pmatrix}\right| \\
&=_{(*)} |A^T V A||X^T V X - X^T V A (A^T V A)^{-1} A^T V X| \\
&=_{(@)} |A^T V A||X^T \left(X (X^T V X)^{-1} X^T \right) X| \\
&= |A^T V A| \frac{|X^T X|^2}{|X^T V^{-1} X|}
\end{aligned}$$

where $(*)$ follows from 1.4 and $(@)$ is due to 11. $\square$

Combine these two lemmas we can get theorem 9.2 easily (exercise).

9.4 Exercises

1. Find $\mathbb{E}(\gamma|y)$ under normality assumption.
2. Derive Henderson's equation 9.2.
3. Use binomial inversion theorem to derive 9.3.
4. Prove 9.7.

References

[1] George AF Seber and Alan J Lee. Linear regression analysis, volume 329. John Wiley & Sons, 2012.

[2] Calyampudi Radhakrishna Rao. Linear statistical inference and its applications, volume 2. Wiley New York, 1973.

[3] Franklin A Graybill and George Marsaglia. Idempotent matrices and quadratic forms in the general linear hypothesis. The Annals of Mathematical Statistics, 28(3):678–686, 1957.

[4] KS Banerjee. A note on idempotent matrices. The Annals of Mathematical Statistics, 35(2): 880–882, 1964.

[5] R Dennis Cook and Sanford Weisberg. Diagnostics for heteroscedasticity in regression. Biometrika, 70(1):1–10, 1983.

[6] Michael R Stoline and Hans K Ury. Tables of the studentized maximum modulus distribution and an application to multiple comparisons among means. Technometrics, 21(1):87–93, 1979.

[7] Yosef Hochberg. Some generalizations of the t-method in simultaneous inference. Journal of multivariate analysis, 4(2):224–234, 1974.

[8] Steven F Arnold. The theory of linear models and multivariate analysis. Wiley New York, 1981.

[9] Mayuri Pandya, Krishnam Bhatt, and Paresh Andharia. Bayes estimation of two-phase linear regression model. Journal of Quality and Reliability Engineering, 2011.

[10] David A Harville. Bayesian inference for variance components using only error contrasts. Biometrika, 61(2):383–385, 1974.

[11] Lynn Roy LaMotte. A direct derivation of the reml likelihood function. Statistical Papers, 48 (2):321–327, 2007.