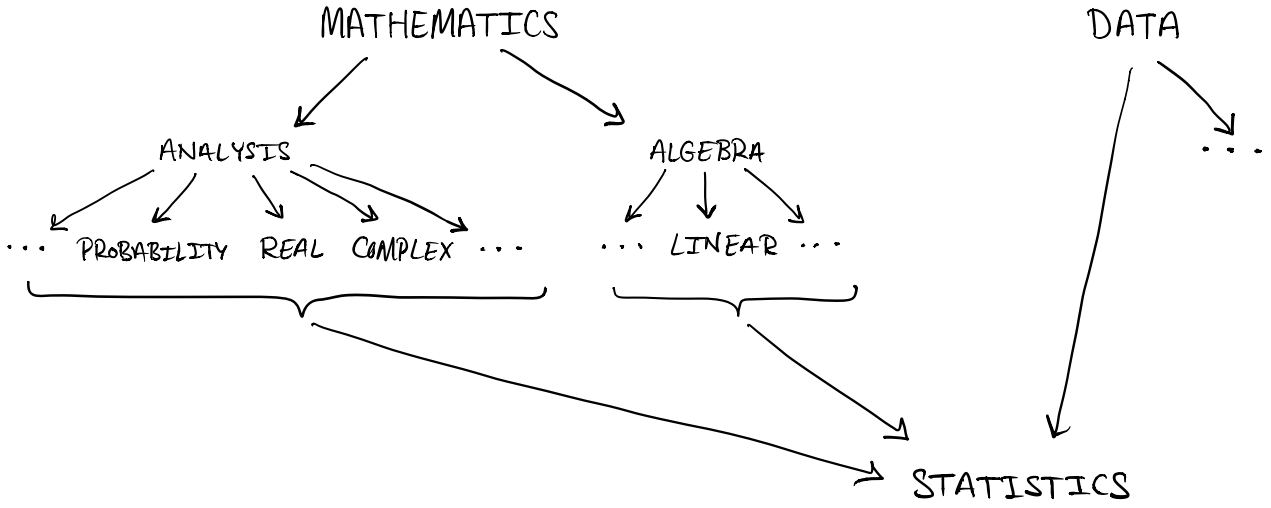
ST 705 LINEAR MODELS AND VARIANCE COMPONENTS

PRELIMINARIES

- Why is mathematics important for statistics?



Statisticians care about coherency (adjective; logical and consistent)

- Why are linear models important in statistics?

Linear models are the foundation of just about any data model. Even a neural net is built from linear model expressions,

"$b + Wx$".

Mathematicians and physicists like to call linear models a "first-order Taylor approximation".

APPENDIX A. REVIEW OF LINEAR ALGEBRA

We begin the course material with a review of linear algebra. Note, however, that although this should be mostly a review of prerequisite topics it is also consider content of ST 705. Accordingly, linear algebra topics are fair game for any assignment, midterm, final, or qualifying exam.

Most importantly, this course is designed to prepare students for a career in research. A PhD is a research, *not* a professional degree.

Linear algebra is the study of vector spaces, linear transformations, matrices, and inner product spaces.

DEFINITION. A vector space V over a field F is a set of elements $v \in V$, with the operations of addition and scalar multiplication, that satisfy the following axioms.

(1) $\forall x, y \in V, \quad x + y = y + x$

(2) $\forall x, y, z \in V, \quad (x+y) + z = x + (y+z)$

(3) There exists an element "0" $\in V$ s.t. $x + 0 = x, \ \forall x \in V$

(4) $\forall x \in V \ \exists y \in V$ s.t. $x + y = 0$

(5) There exists an element "1" $\in F$ s.t. $1 \cdot x = x, \ \forall x \in V$

(6) $\forall a, b \in F, \ \forall x \in V, \quad (ab)x = a(bx)$

(7) $\forall a \in F, \ \forall x, y \in V, \quad a(x+y) = ax + ay$

(8) $\forall a, b \in F, \ \forall x \in V, \quad (a+b)x = ax + bx$ #

In statistics contexts, typically $V = \mathbb{R}^n$ and $F = \mathbb{R}$.

The notion of a vector space is truly foundational for data science. A vector space gives us an entire framework and intuition for thinking about data objects (even as data are stored as objects in a computer).

DEFINITION. A collection of vectors $x_1, \ldots, x_p$ are said to be linearly dependent if and only if there exist scalars $c_1, \ldots, c_p$, not all zero, such that

$$\sum_{i=1}^{p} c_i x_i = 0.$$

A set of vectors that are *not* linearly dependent are said to be linearly independent. #

Often, to show that a set of vectors are linearly independent, a good strategy is to demonstrate that $c_1 x_1 + \cdots + c_p x_p = 0$ implies that $c_1 = \cdots = c_p = 0$.

Note that if $x_1, \ldots, x_p \in \mathbb{R}^n$, then there can be at most $\min\{n, p\}$ linearly independent of the p vectors.

DEFINITION. A basis, B, for a vector space, V, is a linearly independent subset of V with $\text{span}(B) = V$. #

EXAMPLE. For $V = \mathbb{R}^n$, the set $\{e_1, \ldots, e_n\}$, where $e_i \in \mathbb{R}^n$ has ith component with value 1 and 0 for all other components, is a basis. This is commonly called the standard basis, and e_i a standard basis vector. #

THEOREM. Let V be a vector space and $B := \{u_1, \ldots, u_n\} \subseteq V$. Then B is a basis for V if and only if every $v \in V$ can be expressed uniquely as

$$v = a_1 u_1 + \cdots + a_n u_n$$

for some scalars $a_1, \ldots, a_n$. #

Proof. Left as an exercise. #

THEOREM. If a vector space V is generated by some finite set S, then some subset of S is a basis for V. #

Proof. Left as an exercise. #

DEFINITION. Let V and W be vector spaces over field $\mathbb{R}$. We call a function $T : V \to W$ a linear transformation from V to W if $\forall x, y \in V$ and $\forall c \in \mathbb{R}$,

(1) $T(x+y) = T(x) + T(y)$

(2) $T(cx) = c\,T(x)$. #

EXAMPLE. Let $A \in \mathbb{R}^{p \times q}$ and define $T(x) := Ax \quad \forall x \in \mathbb{R}^q$. Then $T : \mathbb{R}^q \to \mathbb{R}^p$ is a linear transformation since

(1) $\forall x, y \in \mathbb{R}^q$, $T(x+y) = A(x+y) = Ax + Ay = T(x) + T(y)$, and

(2) $\forall x \in \mathbb{R}^q$, $\forall c \in \mathbb{R}$, $T(cx) = A \cdot (cx) = c\,Ax = c\,T(x)$. #

EXAMPLE. Let $V = C(\mathbb{R})$, the vector space of continuous real-valued functions defined on $\mathbb{R}$. Then $\forall a, b \in \mathbb{R}$ with $a < b$, the transformation $T : V \to \mathbb{R}$ defined by, #

$$T(f) := \int_a^b f(x)\,dx$$

is linear.

DEFINITION. Let V and W be vector spaces, and $T: V \to W$ a linear transformation. The set,

$$\text{null}(T) := \{x \in V : T(x) = 0\},$$

is called the null space of T. The range of T is defined as,

$$\text{range}(T) := \{y \in W : T(x) = y \text{ for some } x \in V\}.$$

Further, $\text{rank}(T) := \dim(\text{range}(T))$. #

THEOREM. (dimension theorem) Let V and W be vector spaces, and $T: V \to W$ be linear. If V is finite-dimensional, then

$$\dim(\text{null}(T)) + \text{rank}(T) = \dim(V). \quad \#$$

Proof. See any linear algebra textbook. #

Recall that if $T(x) := Ax$, then range(T) is called the column space of A, denoted col(A). Similarly for range(T'), called the row space of A, and denoted by row(A).

DEFINITION. Let $A \in \mathbb{R}^{p \times q}$ and $B \in \mathbb{R}^{q \times m}$. Then the matrix product, denoted by AB, has components defined as

$$(AB)_{ij} := \sum_{k=1}^{q} A_{ik} B_{kj}$$

for $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, m\}$. #

Further notions for matrices.

(1) The transpose of A, denoted A', satisfies the property that $(A')_{ij} = A_{ji}$ for all indices i, j.

(2) $A \in \mathbb{R}^{p \times q}$ is called square if $p = q$.

(3) A square matrix is called symmetric if $A' = A$.

(4) $A \in \mathbb{R}^{p \times p}$ is said to be invertible if $\exists B \in \mathbb{R}^{p \times p}$ s.t. $AB = I_p = BA$. In that case, $A^{-1} := B$.

(5) $A \in \mathbb{R}^{p \times p}$, the trace of A is defined as $tr(A) := \sum_{i=1}^{p} A_{ii}$.

Selected properties of matrices.

(1) $(AB)' = B'A'$

(2) $A \in \mathbb{R}^{p \times p}$ is invertible if and only if $rank(A) = p$ (equivalently $det(A) \neq 0$).

(3) If both $A, B \in \mathbb{R}^{p \times p}$ are invertible, then $(AB)^{-1} = B^{-1}A^{-1}$.

(4) For A and B of appropriate dimensions, $tr(AB) = tr(BA)$.

DEFINITION. Let $A \in \mathbb{R}^{p \times p}$. A nonzero vector $v \in \mathbb{R}^p$ is called an eigenvector of A if $Av = \lambda v$ for some scalar λ, called an eigenvalue. #

THEOREM. A scalar λ is an eigenvalue of a matrix $A \in \mathbb{R}^{p \times p}$ if and only if $det(A - \lambda I_p) = 0$. The polynomial $f(t) := det(A - tI_p)$ is called the characteristic polynomial of A. #

Proof. See any linear algebra textbook. #

Further matrix properties.

(1) If A is symmetric, then all of the eigenvalues of A are real-valued.

(2) The $rank(A)$ is equal to the number of nonzero eigenvalues of $A \in \mathbb{R}^{p \times p}$.

(3) Spectral theorem. For any symmetric matrix $A \in \mathbb{R}^{p \times p}$ there exists an orthogonal matrix Q (i.e., $Q'Q = QQ' = I_p$) such that,

$$A = QDQ',$$

where D is a diagonal matrix composed of the eigenvalues of A. #

Proof. See any linear algebra textbook. #

Recall that the trace, determinant, and spectrum of a matrix all have the property that they are "similarity invariant." That is, they remain the same

regardless of the basis used to express the coordinates of the matrix. Moreover, for any diagonalizable matrix A with eigenvalues $\lambda_1, \ldots, \lambda_p$,

$$tr(A) = \sum_{i=1}^{p} \lambda_i \qquad \text{and} \qquad det(A) = \prod_{i=1}^{p} \lambda_i .$$

THEOREM. Matrices $A, B \in \mathbb{R}^{p \times p}$ are called simultaneously diagonalizable if there exists an invertible matrix $Q \in \mathbb{R}^{p \times p}$ such that both QAQ^{-1} and QBQ^{-1} are diagonal matrices. #

Proof. See any linear algebra textbook. #

For example, every diagonalizable matrix A is simultaneously diagonalizable with the identity matrix. Consider QAQ^{-1} and $QQ^{-1} = I$.

DEFINITION. Let V be a vector space over F ($\mathbb{R}$ or $\mathbb{C}$). An inner product on V, $\langle \cdot, \cdot \rangle : V \times V \longrightarrow F$, satisfies the following axioms. $\forall x, y, z \in V$ and $\forall c \in F$,

(1) $\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

(2) $\langle cx, y \rangle = c \langle x, y \rangle$

(3) $\overline{\langle x, y \rangle} = \langle y, x \rangle$

(4) $\langle x, x \rangle > 0$ if $x \neq 0$. #

Note that conditions 1, 2, 3 imply that $\langle x, x \rangle$ is a real number.

EXAMPLE. Let $x, y \in \mathbb{R}^n$. Then the dot product $\langle x, y \rangle := x'y$ is an inner product. It is left as an exercise to verify this. #

EXAMPLE. The covariance between two random variables X and Y is an inner product. Recall that

$$Cov(X, Y) := E\left[(X - E(X)) \cdot (Y - E(Y))\right].$$

How about if $X, Y \in \mathbb{R}^n$? In that case,

$$Cov(X, Y) := E\left[(X - E(X)) \cdot (Y - E(Y))'\right] \in \mathbb{R}^{n \times n}.$$

In this case, $tr(Cov(X, Y))$ is an inner product. #

EXAMPLE. For $A, B \in \mathbb{R}^{n\times n}$, $\langle A, B\rangle := tr(B'A)$ is an inner product. Commonly it is referred to as the Frobenius inner product, and it induces the Frobenius norm,

$$\|A\|_F := \sqrt{\langle A, A\rangle} = \sqrt{tr(A'A)}. \qquad \#$$

THEOREM. Let V be an inner product space. Then for $x, y, z \in V$ and $c \in F$ ($F = \mathbb{R}$ or $\mathbb{C}$), the following properties hold.

(1) $\langle x, y+z\rangle = \langle x, y\rangle + \langle x, z\rangle$

(2) $\langle x, cy\rangle = \bar{c}\langle x, y\rangle$

(3) $\langle x, 0\rangle = \langle 0, x\rangle = 0$

(4) $\langle x, x\rangle = 0$ if and only if $x = 0$.

(5) If $\langle x, y\rangle = \langle x, z\rangle \quad \forall x$, then $y = z$. $\#$

Proof. Left as an exercise to verify each property. These mostly follow directly from the definition of an inner product. $\#$

DEFINITION. Let V be a vector space. For $x \in V$, the norm of x is defined as

$$\|x\| := \sqrt{\langle x, x\rangle}. \qquad \#$$

THEOREM. Let V be a vector space over $F = \mathbb{R}$ or $\mathbb{C}$. Then $\forall x, y \in V$ and every $c \in F$, the following statements are true.

(1) $\|cx\| = |c|\,\|x\|$

(2) $\|x\| = 0$ if and only if $x = 0$, and $\|x\| \geq 0 \quad \forall x \in V$.

(3) $|\langle x, y\rangle| \leq \|x\| \cdot \|y\|$ (Cauchy–Schwarz inequality)

(4) $\|x+y\| \leq \|x\| + \|y\|$ (triangle inequality) $\#$

Proof of (3). First suppose that $y = 0$. Then by the previous theorem, $\langle x, y\rangle = 0$, and $\|y\| = 0$ by (b). Thus, (c) holds. Next, consider $y \neq 0$ and $x \neq 0$. Then for any $c \in F$,

$$0 \leq \|x - cy\|^2$$

$$= \langle x - cy, x - cy\rangle$$

$$= \langle x,x\rangle - c\langle y,x\rangle - \bar{c}\langle x,y\rangle + c\bar{c}\langle y,y\rangle.$$

So choose $c = \dfrac{\langle x,y\rangle}{\langle y,y\rangle}$ which gives,

$$0 \le \|x\|^2 - \frac{\langle x,y\rangle}{\langle y,y\rangle}\cdot\langle y,x\rangle - \frac{\langle y,x\rangle}{\langle y,y\rangle}\cdot\langle x,y\rangle + \frac{\langle x,y\rangle\langle y,x\rangle}{\langle y,y\rangle}$$

Then

$$\frac{\langle x,y\rangle}{\langle y,y\rangle}\cdot\overline{\langle x,y\rangle} \le \|x\|^2$$

and so

$$|\langle x,y\rangle|^2 \le \|x\|^2\cdot\|y\|^2.$$

Proof of (4).

$$\begin{aligned}
\|x+y\|^2 &= \langle x+y, x+y\rangle \\
&= \|x\|^2 + \langle x,y\rangle + \overline{\langle x,y\rangle} + \|y\|^2 \\
&= \|x\|^2 + 2\cdot\text{real}(\langle x,y\rangle) + \|y\|^2 \\
&\le \|x\|^2 + 2|\langle x,y\rangle| + \|y\|^2 \\
&\le \|x\|^2 + 2\|x\|\cdot\|y\| + \|y\|^2 \qquad \text{by (3)} \\
&= (\|x\| + \|y\|)^2
\end{aligned}$$

#

DEFINITION. Let V be an inner product space. Then $x, y \in V$ are said to be orthogonal if $\langle x,y\rangle = 0$. #

CHAPTER 2. THE LINEAR LEAST SQUARES PROBLEM

The primary focus of ST 705 is the general linear model of the form

$$Y = X\beta + U,$$

where Y is $n\times 1$, X is $n\times p$, β is $p\times 1$, and U is $n\times 1$ with $E(U) = 0$.

Typically Y is some outcome of interest for which data has been observed, and describing variation in the observed or predicting new values of Y is of interest.

The matrix X is called the "design" matrix, and its columns are regarded as covariates or predictors that are relevant for describing variation in Y. Note that X is not regarded as data in the general linear model; the data is Y.

If the general linear model is a meaningful representation of the data, then the values in the vector β have inferential or predictive utility. The problem is that the β values typically are *not* known, or may not even exist.

→ This is why I have a job!

Observe that regarding X and β as fixed implies that U is inherently the random component of the model.

The linear model seems almost trivial now, but 100+ years ago it was the object of many open questions. Nonetheless, research relating to linear models continues to this day. There is an argument to be made that the general linear model formulation is a revolutionary achievement of humankind. The same is true for the sample mean. These ideas were not obvious until they were. We now have laws of large numbers and central limit theorems.

Remarks:

(1) Do *not* underestimate the importance of ideas.
→ Preceding any technology is an idea

(2) Do *not* underestimate the importance of *simple* ideas.
→ Law of parsimony; "The simplest explanation is usually the right one."
→ Look up Occam's razor.

Note that the general linear model is not to be confused with the *generalized* linear model.

Observe that in the context of linear algebra if $U = 0$, then

$$Y = X\beta$$

which implies that $Y \in \text{Col}(X)$. If the columns of X form a basis for $\mathbb{R}^n$, then β is simply the coordinate vector of Y w.r.t. the basis X. In this case,

$$\beta = X^{-1}Y.$$

However, in the context of repeated sampling $p = n$ is problematic. Why?

→ We regard the datum, $Y_i = X_i^t \cdot \beta$, as a single instance of the generative linear model. So if $\beta \in \mathbb{R}^n$, then β changes as new data are

observed, and in fact the model changes. This is one description of what is referred to as "overfitting".

The $p<n$ paradigm is considered the classical setting. The paradigms for $p \geq n$ or $p >> n$ are considered high-dimensional.

With $p<n$ strategies for inferring β involve finding the values that make $X\beta$ "closest" to Y. In the context of Euclidean space, this is most naturally defined as the β that minimizes the sum of squared deviations,

$$Q(\beta) := \|Y - X\beta\|^2.$$

The $\underset{\beta}{\operatorname{argmin}}\{Q(\beta)\}$ is called the least squares solution.

Desirable properties of $Q(\beta)$:

(1) Convex

(2) Differentiable

(3) $\underset{\beta}{\operatorname{argmin}}\{Q(\beta)\} = \{\beta : \nabla_\beta Q(\beta) = 0\}$

Recall that for a function $f : \mathbb{R}^p \to \mathbb{R}$,

$$\nabla_x f(x) := \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_p} \end{pmatrix} \in \mathbb{R}^p.$$

LEMMA. For $a, b \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p \times p}$, the following properties hold.

(1) $\nabla_b(a'b) = a$

(2) $\nabla_b(b'Ab) = (A + A')b$ #

Proof of (1).

$$\nabla_b(a'b) = \begin{pmatrix} \frac{\partial}{\partial b_1}\sum_1^p a_i b_i \\ \vdots \\ \frac{\partial}{\partial b_p}\sum_1^p a_i b_i \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = a.$$

Proof of (2). For all $k \in \{1, \dots, p\}$,

$$\begin{aligned}
\frac{\partial}{\partial b_k}(b'Ab) &= \frac{\partial}{\partial b_k}\left(\sum_{i=1}^{p} A_{ii} b_i^2 + \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j \neq i}}^{p} A_{ij} b_i b_j \right) \\
&= 2A_{kk} b_k + \frac{\partial}{\partial b_k} \sum_{j=2}^{p} A_{1j} b_1 b_j + \cdots + \frac{\partial}{\partial b_k} \sum_{j=1}^{p-1} A_{pj} b_p b_j \\
&= 2A_{kk} b_k + A_{1k} b_1 + \cdots + \sum_{\substack{j=1 \\ j \neq k}}^{p} A_{kj} b_j + \cdots + A_{pk} b_p \\
&= \sum_{j=1}^{p} A_{kj} b_j + \sum_{i=1}^{p} A_{ik} b_i \\
&= A_{k\cdot}' b + b' A_{\cdot k}
\end{aligned}$$

Thus,

$$\begin{aligned}
\nabla_b(b'Ab) &= \begin{pmatrix} A_{1\cdot}' b + b' A_{\cdot 1} \\ \vdots \\ A_{p\cdot}' b + b' A_{\cdot p} \end{pmatrix} \\
&= Ab + (b'A)' \\
&= (A + A')b
\end{aligned}$$

#

Observing the fact that

$$\begin{aligned}
Q(\beta) &= (y - X\beta)'(y - X\beta) \\
&= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta,
\end{aligned}$$

it follows by the lemma that

$$\begin{aligned}
\nabla_\beta Q(\beta) &= -2X'y + (X'X + X'X)\beta \\
&= -2X'(y - X\beta).
\end{aligned}$$

Setting $\nabla_\beta Q(\beta) = 0$ yields the "normal equations"

$$X'X\beta = X'y.$$

EXAMPLE. Recall a simple linear regression,

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad \text{for } i \in \{1, \dots, n\}.$$

In this case,

$$y = \underbrace{\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{=: X} \beta + u.$$

Then

$$X'X\beta = \begin{pmatrix} n & \sum_1^n x_i \\ \sum_1^n x_i & \sum_1^n x_i^2 \end{pmatrix} \beta \quad \text{and} \quad X'y = \begin{pmatrix} \sum_1^n y_i \\ \sum_1^n x_i y_i \end{pmatrix}.$$

Accordingly, the normal equations are

$$n\beta_0 + n\bar{x}_n \beta_1 = n\bar{y}_n$$

$$n\bar{x}_n \beta_0 + \sum_1^n x_i^2 \beta_1 = \sum_1^n x_i y_i$$

and so the least squares solution for β_0 and β_1 are

$$\hat{\beta}_0 = \bar{y}_n - \bar{x}_n \hat{\beta}_1$$

$$\hat{\beta}_1 = \frac{\sum_1^n y_i (x_i - \bar{x}_n)}{\sum_1^n (x_i - \bar{x}_n)^2}$$

Note that if $\sum_1^n (x_i - \bar{x}_n)^2 = 0$, then $\hat{\beta}_1$ is not defined. What does such a condition imply about the design matrix?

Another important identity to recall is $\sum_1^n (x_i - \bar{x}_n) = 0$. #

Next, we will consider the geometry of the least squares solution via the normal equations

$$X'X\beta = X'y.$$

Is there a solution β? Equivalently, is $X'y \in \text{Col}(X'X)$?

→ An approach to show that this is the case is to show $\text{Col}(X'X) = \text{Col}(X')$.

In this direction,

LEMMA. For $X \in \mathbb{R}^{n \times p}$, $\text{null}(X'X) = \text{null}(X)$. #

Proof. First observe that if $v \in \text{null}(X)$, then

$$X'Xv = X'(Xv) = X'0 = 0,$$

and so $v \in \text{null}(X'X)$. Hence, $\text{null}(X) \subseteq \text{null}(X'X)$.

For the other direction, assume $v \in \text{null}(X'X)$. Accordingly,

$$X'Xv = 0$$

$$v'X'Xv = 0$$

$$\|Xv\|^2 = 0.$$

Since a norm only maps zero to zero, $v \in \text{null}(X)$. Therefore, $\text{null}(X) \supseteq \text{null}(X'X)$. Both set inclusions have now been demonstrated for equivalence. #

LEMMA. For $X \in \mathbb{R}^{n \times p}$, $\text{rank}(X) = \text{rank}(X')$. #

Proof. Let $k := \text{rank}(X') = \dim(\text{col}(X')) = \dim(\text{row}(X))$, and let $\{u_1, \ldots, u_k\}$ be a basis for $\text{row}(X)$. Then $u_i \in \mathbb{R}^p \ \forall i \in \{1, \ldots, k\}$, and it can be shown that $\{Xu_1, \ldots, Xu_k\}$ is a linearly independent subset of $\text{col}(X)$, as follows. Suppose that

$$\sum_1^k a_i X u_i = X \cdot \sum_1^k a_i u_i = 0.$$

Then $\sum_1^k a_i u_i \perp \text{row}(X)$. However, since $\sum_1^k a_i u_i \in \text{row}(X)$ it must be the case that

$$\sum_1^k a_i u_i = 0$$

which implies that $a_1 = \cdots = a_k = 0$ because $u_1, \ldots, u_k$ are linearly independent. Thus, $\text{rank}(X) \geq k = \text{rank}(X')$.

Reversing the roles of X and X' demonstrates that $\text{rank}(X') \geq \text{rank}(X)$. Hence,

$$\text{rank}(X) \geq \text{rank}(X') \geq \text{rank}(X).$$ #

Observe that this result and argument follow from the fact that $\text{col}(X') \perp \text{null}(X)$.

THEOREM. For $X \in \mathbb{R}^{n \times p}$, $\text{col}(X'X) = \text{col}(X')$. #

Proof. If $v \in \text{col}(X'X)$, then for some $a \in \mathbb{R}^p$,

$$v = X'Xa = X'(Xa) \in \text{col}(X'),$$

and so $\text{col}(X'X) \subseteq \text{col}(X')$.

Next, recall from the dimension theorem that for $X'X : \mathbb{R}^p \to \mathbb{R}^p$,

$$\text{rank}(X'X) + \dim(\text{null}(X'X)) = p$$

$$\text{rank}(X) + \dim(\text{null}(X)) = p.$$

By the first lemma, $\text{null}(X) = \text{null}(X'X)$, so it follows that

$$\text{rank}(X'X) = \text{rank}(X) = \text{rank}(X'),$$

where the second equality was established in the second lemma. The result is now given from a direct application of a previous homework problem. That is, for any basis $\{u_1, \ldots, u_k\}$ for the $\text{col}(X'X)$, for $k := \text{rank}(X'X)$,

$$\{u_1, \ldots, u_k\} \subseteq \text{col}(X'X) \subseteq \text{col}(X').$$

Since $k = \text{rank}(X'X) = \text{rank}(X')$, $\{u_1, \ldots, u_k\}$ is also a basis for $\text{col}(X')$. #

Returning to the question of whether there exists a solution β to the normal equations $X'X\beta = X'y$, we now know that

$$X'y \in \text{col}(X') = \text{col}(X'X).$$

Therefore, there must exist a solution. That being so, is a solution to the normal equation necessarily a least squares solution?

Note that from the geometry perspective, we are avoiding calculus to solve the least squares problem.

THEOREM. $\hat{\beta} \in \{\beta : X'X\beta = X'y\}$ if and only if $\hat{\beta} \in \underset{\beta}{\text{argmin}}\{Q(\beta)\}$. #

Proof. First assume $\hat{\beta} \in \{\beta : X'X\beta = X'y\}$. Then

$$Q(\beta) = \|y - X\beta\|^2$$

$$= \|y - X\hat{\beta} + X\hat{\beta} - X\beta\|^2$$

$$= \|y - X\hat{\beta}\|^2 + 2(X\hat{\beta} - X\beta)'(y - X\hat{\beta}) + \|X\hat{\beta} - X\beta\|^2$$

$$= Q(\hat{\beta}) + 2(\hat{\beta} - \beta)'\underbrace{(X'y - X'X\hat{\beta})}_{=0 \text{ by assumption}} + \|X(\hat{\beta} - \beta)\|^2$$

$$= Q(\hat{\beta}) + \|X(\hat{\beta} - \beta)\|^2,$$

which is minimum when $\|X(\hat{\beta} - \beta)\|^2 = 0$. Hence, $\hat{\beta} \in \underset{\beta}{\operatorname{argmin}}\{Q(\beta)\}$.

Conversely, assume $\hat{\beta} \in \underset{\beta}{\operatorname{argmin}}\{Q(\beta)\}$, and recall that for any solution, $\tilde{\beta}$, to the normal equations,

$$Q(\beta) = Q(\tilde{\beta}) + \|X(\tilde{\beta} - \beta)\|^2 \geq Q(\tilde{\beta}).$$

Thus, $\|X(\tilde{\beta} - \hat{\beta})\|^2 = 0$ which implies that

$$X(\tilde{\beta} - \hat{\beta}) = 0$$

$$X'X(\tilde{\beta} - \hat{\beta}) = X'0 = 0$$

$$X'X\hat{\beta} = X'X\tilde{\beta} = X'y. \qquad \#$$

COROLLARY. The vector $X\hat{\beta}$ is invariant to the choice of solution, $\hat{\beta}$, to the normal equations. $\#$

Proof. As observed in the previous proof, for any solution, $\hat{\beta}$, to the normal equations,

$$Q(\beta) = Q(\hat{\beta}) + \|X(\hat{\beta} - \beta)\|^2,$$

which is minimized for any β satisfying $\|X(\hat{\beta} - \beta)\|^2 = 0$ so that $X\hat{\beta} = X\beta$. Thus, by the theorem, $\forall\, \tilde{\beta} \in \{\beta : X'X\beta = X'y\}$, $\tilde{\beta} \in \underset{\beta}{\operatorname{argmin}}\{Q(\beta)\}$, and

$$X\hat{\beta} = X\tilde{\beta}. \qquad \#$$

A geometric interpretation of this corollary is that $\hat{\beta} - \tilde{\beta} \in \text{null}(X'X) = \text{null}(X)$. This gives $X\hat{\beta} = X\tilde{\beta}$ for any solutions to the normal equation, meaning that any solution to the normal equations yields the same prediction,

$$\hat{y} := X\hat{\beta} \in \text{col}(X).$$

If we decompose $y = X\hat{\beta} + \underbrace{y - X\hat{\beta}}_{=:\hat{e}}$, it turns out that $\hat{e} \in \text{null}(X')$ since

$$X'\hat{e} = X'y - X'X\hat{\beta} = 0.$$

Moreover, note that this decomposition is unique because $\text{col}(X) \perp \text{null}(X')$, and

$$\begin{aligned}\|y\|^2 &= (X\hat{\beta}+\hat{e})'(X\hat{\beta}+\hat{e}) \\ &= \|X\hat{\beta}\|^2 + 2\hat{\beta}'X'\hat{e} + \|\hat{e}\|^2 \\ &= \|\hat{y}\|^2 + \|\hat{e}\|^2.\end{aligned}$$

The uniqueness of this decomposition / Pythagorean theorem implies that for the orthogonal projection matrix P onto $\text{col}(X)$,

$$y = Py + (I-P)y = X\hat{\beta} + \hat{e}$$

with $Py = X\hat{\beta}$ and $(I-P)y = \hat{e}$.

While the normal equations $X'X\beta = X'y$ characterize the least squares solution, and we have established that a solution β exists, we still do not have an explicit expression for the solution. In the case that X has full column rank,

$$\hat{\beta} = (X'X)^{-1}X'y.$$

But what if $\text{rank}(X) < p$?

$\longrightarrow$ In this case, it turns out that $\hat{\beta} = (X'X)^{g}X'y$, but we need to first construct the orthogonal projection onto $\text{col}(X)$ to understand why.

First a few preliminary results.

LEMMA. For $X \in \mathbb{R}^{n\times p}$ and $A, B \in \mathbb{R}^{p\times q}$, $X'XA = X'XB$ if and only if $XA = XB$. #

Proof. If $XA = XB$, then $X'XA = X'XB$.

For the reverse direction, suppose that $X'XA = X'XB$. Then

$$\begin{aligned}\|XA - XB\|_F^2 &= \text{tr}\big((XA-XB)'(XA-XB)\big) \\ &= \text{tr}\big((A-B)'(X'XA - X'XB)\big) \\ &= 0.\end{aligned}$$

Therefore, $XA = XB$. #

DEFINITION. For $A \in \mathbb{R}^{n\times p}$, the Moore-Penrose conditions for the pseudo-inverse of A, say G, are the following.

(1) $AGA = A$

(2) $GAG = G$

(3) $(AG)' = AG$

(4) $(GA)' = GA$

A pseudo-inverse satisfying (1) is called a generalized inverse, denoted A^g
A pseudo-inverse satisfying (1)-(4) is called a Moore-Penrose generalized inverse, denoted A^+. #

LEMMA. For $X \in \mathbb{R}^{n \times p}$, $(X'X)^g X'$ is a generalized inverse of X. #

Proof. By definition,

$$X'X \underbrace{(X'X)^g X'X}_{=:A} = X'X = X'X \underbrace{I_p}_{=:B}.$$

Then by the previous lemma, $(X'X)^g X'X = I_p$, and so

$$X \underbrace{(X'X)^g X'}_{= X^g} X = X.$$

#

Note that a generalized inverse always exists, and the Moore-Penrose generalized inverse is unique.

THEOREM. For $X \in \mathbb{R}^{n \times p}$, $P_X := X(X'X)^g X'$ is the symmetic projection matrix onto $col(X)$. That is, P_X is

(1) Idempotent
(2) Projection onto $col(X)$
(3) Invariant to the choice of generalized inverse
(4) Symmetric
(5) unique. #

Proof of (1).

$$P_X \cdot P_X = \underbrace{X(X'X)^g X'X}_{=X}(X'X)^g X' = X(X'X)^g X' = P_X.$$

Proof of (2). Let $u \in \mathbb{R}^n$. Then

$$P_X u = X(X'X)^g X'u = X \cdot [(X'X)^g X'u] \in col(X).$$

Further, if $u \in \text{col}(X)$, then for some $a \in \mathbb{R}^p$,

$$P_X u = X(X'X)^g X'u = X(X'X)^g X'Xa = Xa = u.$$

Proof of (3). Let G_1 and G_2 be any two generalized inverses of $X'X$. Then

$$X'XG_1X'X = X'X = X'XG_2X'X.$$

Taking $A = G_1X'X$ and $B = G_2X'X$, by a previous lemma,

$$XG_1X'X = XG_2X'X,$$

and so $X'XG_1'X' = X'XG_2'X'$. By the same logic, $XG_1'X' = XG_2'X'$. Since $X'X$ is symmetric,

$$XG_1X' = XG_2X' = P_X.$$

Proof of (4). Follows from the fact that $X'X$ is symmetric.

Proof of (5). Assume that Q is a symmetric projection matrix onto $\text{col}(X)$. Then for any $v \in \mathbb{R}^n$, $Qv \in \text{col}(X)$. Accordingly, $P_XQv = Qv$, and since this is true for every v, it follows that $P_XQ = Q$. Similary, $QP_X = P_X$. Hence,

$$\begin{aligned} \|P_X - Q\|_F^2 &= \text{tr}\left((P_X - Q)'(P_X - Q)\right) \\ &= \text{tr}(P_X - QP_X - P_XQ + Q) \\ &= \text{tr}(P_X - P_X - Q + Q) \\ &= 0. \end{aligned}$$

#

COROLLARY. The matrix $I - P_X$ is the unique, symmetric projection onto $\text{null}(X')$. #

Proof. Similar to the theorem. #

EXAMPLE. Let $X = 1_n$. Then $\text{col}(X)$ is the subspace of vectors that have all components equal to the same value. That being true, projecting onto $\text{col}(X)$ is a projection of a vector to a single value.

$$P_X = X(X'X)^g X' = \frac{1}{n} XX' = \frac{1}{n} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$$

Then for any $y \in \mathbb{R}^n$,

$$P_X y = \frac{1}{n}\begin{pmatrix} \sum_1^n y_i \\ \vdots \\ \sum_1^n y_i \end{pmatrix} = \bar{y}_n \cdot 1_n$$

#

Returning back to the least squares problem in the general linear model with the intuition of the decomposition,

$$y = P_X y + (I - P_X) y = X\hat{\beta} + \hat{e},$$

where $\hat{\beta} \in \{X'X\beta = X'y\}$ and $\hat{e} = y - X\hat{\beta}$, observe that by the uniqueness of the decomposition,

$$X\hat{\beta} = P_X y = X(X'X)^g X'y.$$

This suggests $\tilde{\beta} = (X'X)^g X'y$ as a natural candidate for $\hat{\beta}$. Moreover, since

$$X'X\tilde{\beta} = X'X(X'X)^g X'y = X'y,$$

$\tilde{\beta}$ is in fact a least squares solution. We also conclude that $P_X y$ is the least squares projection, which provides a lot of insight for how the least squares solution characterizes the association between y and X. Note also that the decomposition

$$y = P_X y + (I - P_X) y$$

is the unique decomposition from the direct sum $col(X) \oplus null(X')$.

THEOREM. $\{\beta : X'X\beta = X'y\} = \{\beta : X\beta = P_X y\}$ #

Proof. First suppose that $\beta \in \{X'X\beta = X'y\}$. Then

$$X'X\beta = X'y$$
$$= X'X(X'X)^g X'y,$$

and so $X\beta = X(X'X)^g X'y = P_X y$. For the reverse direction, if $\beta \in \{X\beta = P_X y\}$, then

$$X'X\beta = X'P_X y = X'X(X'X)^g X'y = X'y. \quad \#$$

We conclude this section with one final result that will be useful later on.

THEOREM. Let $X \in \mathbb{R}^{n \times p}$ and $W \in \mathbb{R}^{n \times q}$. If $col(W) \subseteq col(X)$, then $P_X - P_W$ is the

projection onto $\mathrm{col}((I_n - P_W)X)$. #

Proof. Need to show three things: (i) that $P_X - P_W$ is an orthogonal projection, (ii) that $\mathrm{col}(P_X - P_W) \subseteq \mathrm{col}((I_n - P_W)X)$, and (iii) that $P_X - P_W$ acts as the identity operator on $\mathrm{col}((I_n - P_W)X)$. Note that symmetry gives uniqueness.

To show (i), $\forall v \in \mathbb{R}^n$,

$$(P_X - P_W)(P_X - P_W)v = P_X v - P_W P_X v - P_X P_W v - P_W v.$$

By assumption, $P_W v \in \mathrm{col}(W) \subseteq \mathrm{col}(X)$, and so $P_X P_W v = P_W v$. Since this is true for any v, it follows that $P_X P_W = P_W$. Accordingly,

$$P_W P_X = (P_X' P_W')' = (P_X P_W)' = P_W' = P_W.$$

Hence, for any v,

$$\begin{aligned}(P_X - P_W)(P_X - P_W)v &= P_X v - P_W P_X v - P_X P_W v - P_W v \\ &= P_X v - P_W v - P_W v - P_W v \\ &= (P_X - P_W)v.\end{aligned}$$

To show (ii), $\forall v \in \mathbb{R}^n$,

$$\begin{aligned}(P_X - P_W)v &= (P_X - P_W P_X)v \\ &= (I_n - P_W)P_X v \\ &= (I_n - P_W)Xa \in \mathrm{col}((I_n - P_W)X),\end{aligned}$$

for some $a \in \mathbb{R}^p$.

To show (iii), let $u \in \mathrm{col}((I_n - P_W)X)$. Then for some $a \in \mathbb{R}^p$,

$$\begin{aligned}(P_X - P_W)u &= (P_X - P_W)(I_n - P_W)Xa \\ &= (P_X - P_X P_W - P_W + P_W)Xa \\ &= (P_X - P_W)Xa \\ &= (P_X - P_W P_X)Xa \\ &= (I_n - P_W)Xa \\ &= u.\end{aligned}$$

#

REPARAMETERIZATION

In the context of linear models, reparameterization refers to the equivalence of two or more models with different design matrices but the same least squares fit to the data. An advantage of reparameterizations is that they allow the interpretability of one model with the computational ease of another. A major disadvantage, however, is that the parameters are not all identifiable.

For example, if

$$Y = X\beta + U,$$

for some fixed coefficient vector β, then reparameterization is possible if $\text{rank}(X) < p$. In that case, β is not unique. Suppose $p = 2$, so that for some $\alpha \in \mathbb{R}$,

$$X\beta = X_1\beta_1 + X_2\beta_2 = X_1\beta_1 + \alpha X_1\beta_2 = X_1 \underbrace{(\beta_1 + \alpha\beta_2)}_{=:\gamma}.$$

DEFINITION. Let $X \in \mathbb{R}^{n\times p}$ and $W \in \mathbb{R}^{n\times q}$. Two linear models $Y = X\beta + U$ and $Y = W\gamma + U$ are said to be reparameterizations if $\text{col}(X) = \text{col}(W)$. #

Note that if $\text{col}(X) = \text{col}(W)$, then there exist matrices S and T such that $W = XT$ and $X = WS$, and we will show that,

(1) $P_X = P_W$

(2) $X\hat{\beta} = \hat{y} = W\hat{\gamma}$

(3) $(I - P_X)y = \hat{e} = (I - P_W)y$.

EXAMPLE. One-way ANOVA. Let

$$X\beta = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & 0 \\ 1_{n_2} & 0 & 1_{n_2} & 0 \\ 1_{n_3} & 0 & 0 & 1_{n_3} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix},$$

and observe the full column rank reparameterization,

$$X\beta = W\gamma = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 \\ 1_{n_2} & 0 & 1_{n_2} \\ 1_{n_3} & 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{\mu} \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{pmatrix},$$

where

$$\underbrace{\begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & 0 \\ 1_{n_2} & 0 & 1_{n_2} & 0 \\ 1_{n_3} & 0 & 0 & 1_{n_3} \end{pmatrix}}_{=X} = \underbrace{\begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 \\ 1_{n_2} & 0 & 1_{n_2} \\ 1_{n_3} & 0 & 0 \end{pmatrix}}_{=W} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}}_{=:S} \quad \#$$

THEOREM. Let $X \in \mathbb{R}^{n \times p}$ and $W \in \mathbb{R}^{n \times q}$. If $col(X) = col(W)$, then $P_X = P_W$. #

Proof. $\forall y \in \mathbb{R}^n$, $y = P_X y + (I - P_X)y$ and $y = P_W y + (I - P_W)y$. Since the expression of y from $col(X) \oplus null(X')$ is unique, it follows that $\forall y$, $P_X y = P_W y$. Thus, $P_X = P_W$. #

COROLLARY. Let $X \in \mathbb{R}^{n \times p}$ and $W \in \mathbb{R}^{n \times q}$. If $col(X) = col(W)$, then $\hat{y} := P_X y = P_W y$ and $\hat{e} := (I - P_X)y = (I - P_W)y$. #

Proof. Direct application of the theorem. #

THEOREM. Let $X \in \mathbb{R}^{n \times p}$, $W \in \mathbb{R}^{n \times q}$, and $col(X) = col(W)$. If $\hat{\gamma}$ solves the normal equations in W, then $\hat{\beta} := T\hat{\gamma}$ solves the normal equations in X, where $W = XT$. #

Proof. By assumption, $W'W\hat{\gamma} = W'y$. Then

$$X'X\hat{\beta} = X'XT\hat{\gamma} = X'W\hat{\gamma} = X'P_W y = X'P_X y = X'X(X'X)^g X'y = X'y \quad \#$$

GRAM-SCHMIDT ORTHONORMALIZATION

Here we study a procedure for how to construct a set of mutually orthogonal vectors from a set of linearly independent vectors. Suppose that $x_1, \ldots, x_p \in \mathbb{R}^n$ are linearly independent. The result of the Gram-Schmidt orthonormalization procedure is a set of orthogonal vectors $u_1, \ldots, u_p \in \mathbb{R}^n$ such that

$$span\{u_1, \ldots, u_p\} = span\{x_1, \ldots, x_p\}.$$

In matrix form with $U := (u_1, \ldots, u_p)$ and $X := (x_1, \ldots, x_p)$, that is

$$col(U) = col(X).$$

How to construct u_1?

$\longrightarrow$ Take $u_1 := x_1$

Next, for u_2, modify x_2 so that it is orthogonal to u_1. Such a vector is in the orthogonal complement of the span of u_1,

$\longrightarrow$ Take $u_2 := (I_n - P_{u_1}) x_2$

Verify that

$$u_2' u_1 = x_2' (I_n - P_{u_1}) u_1 = 0.$$

Continuing on, the third vector must be orthogonal to both u_1 and u_2. That is,

$\longrightarrow$ Take $u_3 := (I_n - P_{u_1} - P_{u_2}) x_3$

so that

$$\begin{aligned} u_3' u_2 &= x_3' (I_n - P_{u_1} - P_{u_2}) u_2 \\ &= -x_3' P_{u_1} u_2 \\ &= -x_3' u_1 (u_1' u_1)^{-1} u_1' u_2 \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} u_3' u_1 &= x_3' (I_n - P_{u_1} - P_{u_2}) u_1 \\ &= -x_3' P_{u_2} u_1 \\ &= -x_3' u_2 (u_2' u_2)^{-1} u_2' u_1 \\ &= 0. \end{aligned}$$

Accordingly,

$$\begin{aligned} u_1 &= x_1 \\ u_2 &= x_2 - \frac{u_1 u_1' x_2}{\|u_1\|^2} \\ u_3 &= x_3 - \frac{u_1 u_1' x_3}{\|u_1\|^2} - \frac{u_2 u_2' x_3}{\|u_2\|^2} \\ &\vdots \end{aligned}$$

More concisely, for $j \in \{1, \dots, n\}$,

$$u_j = x_j - \left(\sum_{k=1}^{j-1} \frac{u_k u_k'}{\|u_k\|^2} \right) x_j .$$

So now we have an orthogonal set $\{u_1, \dots, u_p\}$. How to show that,

$$\text{span}\{u_1, \dots, u_p\} = \text{span}\{x_1, \dots, x_p\} ?$$

Since this is equivalent to $\text{col}(U) = \text{col}(X)$, express

$$X = (u_1, \dots, u_p) \cdot \underbrace{\begin{pmatrix} 1 & \frac{u_1'x_2}{\|u_1\|^2} & \frac{u_1'x_3}{\|u_1\|^2} & \frac{u_1'x_4}{\|u_1\|^2} & \dots & \frac{u_1'x_p}{\|u_1\|^2} \\ 0 & 1 & \frac{u_2'x_3}{\|u_2\|^2} & \frac{u_2'x_4}{\|u_2\|^2} & \dots & \frac{u_2'x_p}{\|u_2\|^2} \\ 0 & 0 & 1 & \frac{u_3'x_4}{\|u_3\|^2} & \dots & \frac{u_3'x_p}{\|u_3\|^2} \\ 0 & 0 & 0 & 1 & \dots & \frac{u_4'x_p}{\|u_4\|^2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}}_{=: S}$$

$$= \underbrace{\left(\frac{u_1}{\|u_1\|}, \dots, \frac{u_p}{\|u_p\|} \right)}_{=: Q} \cdot \underbrace{\begin{pmatrix} \|u_1\| & & \\ & \ddots & \\ & & \|u_p\| \end{pmatrix} \cdot S}_{=: R}$$

Note that $X = QR$ is commonly called the "QR decomposition" for an orthogonal matrix Q, and an upper triangular matrix R. This expression also yields the Cholesky decomposition of $X'X$,

$$X'X = R'Q'QR = R'R,$$

where R is upper triangular. Think about the uniqueness of these decompositions.

CHAPTER 3. ESTIMABILITY AND LEAST SQUARES ESTIMATORS

In the previous chapter we essentially considered $X\beta$ as a mathematical approximation to a data vector $y \in \mathbb{R}^n$. In this chapter we will consider

$$Y = X\beta + U$$

as a statistical model with the assumption that $E(U) = 0$. Namely, if the general linear model describes how Y is actually generated, then what can be said about the feasibility of estimating features relating to β? To contrast, in the previous chapter the columns of X were rather arbitrary vectors in $\mathbb{R}^n$.

Goals:

(i) Determine whether certain functions of parameters are "estimable".

(ii) Construct "unbiased estimators" for the "estimable" functions.

DEFINITION. An estimator $t(y)$ is an unbiased estimator for the scalar $\lambda'\beta$ if and only if $E(t(Y)) = \lambda'\beta$, $\forall \beta$. #

DEFINITION. An estimator $t(y)$ is a linear estimator in y if and only if $t(y) = c + a'y$, for some constants a and c. #

DEFINITION. A function $\lambda'\beta$ is linearly estimable if and only if there exists a linear unbiased estimator of it. If no such estimator exists, then the function is called nonestimable. #

THEOREM. $\lambda'\beta$ is linearly estimable if and only if there exists a vector a such that $E(a'Y) = \lambda'\beta$, $\forall \beta$. #

Proof. First suppose that $\exists a \in \mathbb{R}^n$ such that $E(a'Y) = \lambda'\beta$, $\forall \beta \in \mathbb{R}^p$. Then $t(y) = a'y$ is an unbiased linear estimator of $\lambda'\beta$, $\forall \beta$, with $c = 0$.

Conversely, if $\lambda'\beta$ is linearly estimable, then there exist constants a and c such that $t(y) = c + a'y$, and $\forall \beta$,

$$\lambda'\beta = E(t(Y)) = c + a'E(Y) = c + a'X\beta.$$

Choosing $\beta = 0$ demonstrates that $c = 0$, so that $\forall \beta \in \mathbb{R}^p$, $t(y) = a'y$ is a linear unbiased estimator of $\lambda'\beta$. #

In words, $\lambda'\beta$ is linearly estimable if and only if it is equal to the expected value of a linear combination of the data. Observe the condition that $E(a'Y) = \lambda'\beta$, $\forall \beta$ implies that $\lambda = X'a$. So essentially, linear estimability of a function $\lambda'\beta$ is equivalent to the condition that $\lambda \in \text{col}(X')$.

EXAMPLE. Suppose $Y = 1_n\beta + U$. Then for $\lambda = 1$, $\lambda'\beta = \beta$. Choosing $a = e_1 \in \mathbb{R}^n$,

$$E(a'Y) = a'1_n\beta = \beta,$$

$\forall \beta \in \mathbb{R}$. That being true, $t(y) = y_1$ is a linear unbiased estimator of β. Probably y_1 is not a great estimator of β due to high variability, and the discard of the information $y_2, \ldots, y_n$, but the theorem only gives the necessary and sufficient conditions for the existence of a linear estimator. A better choice would be $a = \frac{1}{n} \cdot 1_n$. Then $t(y) = a'y = \bar{y}_n$, and

$$E(\bar{Y}_n) = E(a'Y) = \frac{1}{n} 1_n' 1_n \beta = \beta. \quad \#$$

EXAMPLE. Suppose

$$Y = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \underbrace{\begin{pmatrix} \mu \\ \alpha \end{pmatrix}}_{=:\beta} + U$$

To estimate α, choose $\lambda = (0, 1)'$ since $\lambda'\beta = \alpha$. For α to be estimable, it must be the case that $\lambda \in \text{col}(X')$, so determine if there exists a solution to

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

One solution is $a = (1, 0, -1, 0)'$, and so $t(y) = y_1 - y_3$ is a linear unbiased estimator of α. #

EXAMPLE. Now let's consider an example where $\lambda'\beta$ is not linearly estimable.

$$Y = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}}_{=:\beta} + U$$

In this case, $\lambda = (0,1,0)'$ to estimate α_1, but

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

does not have a solution. Thus, there exists no linear unbiased estimator for α_1 because $\lambda \neq X'a$ for any $a \in \mathbb{R}^n$. #

Notice that the set of vectors λ for which $\lambda'\beta$ is linearly estimable actually forms an entire subspace,

$$\{\lambda \in \mathbb{R}^p : \lambda = X'a \text{ for some } a \in \mathbb{R}^n\} = \text{col}(X').$$

Moreover, if $\text{rank}(X') = p$, then $\text{col}(X') = \mathbb{R}^p$ and so $\lambda'\beta$ is linearly estimable $\forall \lambda \in \mathbb{R}^p$. An important corollary of this discussion is that each component of β is linearly estimable (i.e., identifiable) if X has full column rank.

With the notion of linear estimability now established, what are some effective strategies for determining if a function $\lambda'\beta$ is linearly estimable?

METHOD 1. If $\lambda \in \text{col}(X')$, then $\lambda'\beta$ is estimable. Accordingly, for any basis $\{v_1, \ldots, v_k\}$ for $\text{col}(X')$ determine if the system

$$v_1 a_1 + \cdots + v_k a_k = \lambda$$

has a solution. Equivalently, since $P_{X'} = X'(XX')^g X$ and $X'X(X'X)^g$ are projections onto $\text{col}(X')$, it suffices to show that either $P_{X'}\lambda = \lambda$ or $X'X(X'X)^g\lambda = \lambda$.

METHOD 2. Since $\text{col}(X') \oplus \text{null}(X) = \mathbb{R}^n$, $\lambda \in \text{col}(X')$ if and only if $\lambda \perp \text{null}(X)$. That being so, construct a basis for $\text{null}(X)$, $\{w_1, \ldots, w_{p-k}\}$, and show that $\lambda' w_j = 0 \quad \forall j \in \{1, \ldots, p-k\}$.

METHOD 3. If $\lambda'\beta$ can be expressed as a linear combination of $E(Y_1), \ldots, E(Y_n)$, then $\lambda'\beta$ is estimable.

THEOREM. If $\lambda'\beta$ is estimable, then the least squares estimator $\lambda'\hat{\beta}$ is the same for all solutions to the normal equations. #

Proof. Recall that any $\hat{\beta} \in \{X'X\beta = X'y\}$ can be expressed as

$$\hat{\beta} = X^g y + (I_p - X^g X)z$$

for some $z \in \mathbb{R}^p$. Assuming $\lambda'\beta$ is estimable, $\lambda = X'a$ for some $a \in \mathbb{R}^n$, and so

$$\begin{aligned}\lambda'\hat{\beta} &= a'X\left[X^g y + (I_p - X^g X)z\right] \\ &= a'X(X'X)^g X'y + a'(X - XX^g X)z \\ &= a'P_X y\end{aligned}$$

which does not depend on z nor the choice of generalized inverse $(X'X)^g$.

The converse direction is left as a homework problem. #

Alternative proof. Let $\hat{\beta}_1, \hat{\beta}_2 \in \{X'X\beta = X'y\}$. Then

$$\begin{aligned}X'X\hat{\beta}_1 &= X'X\hat{\beta}_2 \\ X'X(\hat{\beta}_1 - \hat{\beta}_2) &= 0\end{aligned}$$

and so $\hat{\beta}_1 - \hat{\beta}_2 \in \text{null}(X'X) = \text{null}(X)$. For estimable $\lambda'\beta$, $\lambda \perp \text{null}(X)$ so it follows that $\lambda'(\hat{\beta}_1 - \hat{\beta}_2) = 0$. #

THEOREM. The least squares estimator $\lambda'\hat{\beta}$ of an estimable function $\lambda'\beta$ is a linear unbiased estimator of $\lambda'\beta$. #

Proof. First observe that for some $a \in \mathbb{R}^n$ and $z \in \mathbb{R}^p$,

$$\begin{aligned}\lambda'\hat{\beta} &= a'X\left[X^g y + (I_p - X^g X)z\right] \\ &= a'X(X'X)^g X'y + a'(X - XX^g X)z \\ &= a'P_X y\end{aligned}$$

Hence, $\lambda'\hat{\beta} = v'y$ for some $v \in \mathbb{R}^n$. Next,

$$E(a'\hat{\beta}) = a'P_X E(y) = a'P_X X\beta = a'X\beta = a'\beta. \qquad \#$$

REPARAMETERIZATION REVISITED

Recall the scenario for reparameterization:

$$\text{Model 1}: \; Y = X\beta + U \qquad\qquad \text{Model 2}: \; Y = W\gamma + U$$

$$X \in \mathbb{R}^{n\times p} \qquad\qquad W \in \mathbb{R}^{n\times q}$$

$\longrightarrow$ Same U for both models with $E(U) = 0$, and $\text{col}(X) = \text{col}(W)$. Moreover, there exists matrices S and T such that

$$X = WS \qquad \text{and} \qquad W = XT$$

In this section we investigate the effect of reparameterization on estimability.

THEOREM. If $a'\beta$ is estimable in the model with design X, and $\hat{\gamma}$ solves the normal equations in design W, then $a'T\hat{\gamma}$ is the least squares estimator of $a'\beta$. $\#$

Proof. Recall from a result in our initial investigation of reparameterizations,

$$X'XT\hat{\gamma} = X'W\hat{\gamma} = X'P_W y = X'P_X y = X'X(X'X)^{g}X'y = X'y,$$

since $\{W'W\gamma = W'y\} = \{W\gamma = P_W y\}$. Therefore, $T\hat{\gamma} \in \underset{\beta}{\text{argmin}}\{\|y - X\beta\|^2\}$. This completes the argument because we proved a previous theorem demonstrating the invariance of the least squares estimator $a'\hat{\beta}$ over $\{X'X\beta = X'y\}$. $\#$

EXAMPLE. One-way ANOVA. Let

$$X\beta = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & 0 \\ 1_{n_2} & 0 & 1_{n_2} & 0 \\ 1_{n_3} & 0 & 0 & 1_{n_3} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix},$$

and recall the full column rank reparameterization,

$$X\beta = W\gamma = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 \\ 1_{n_2} & 0 & 1_{n_2} \\ 1_{n_3} & 0 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}.$$

For matrices,

$$S = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

we have $X = WS$ and $W = XT$. Then the unique solution to the normal equations in the full column rank design W has the form,

$$\begin{aligned} \hat{\gamma} &= (W'W)^{-1}W'y \\ &= \begin{pmatrix} n_1+n_2+n_3 & n_1 & n_2 \\ n_1 & n_1 & 0 \\ n_2 & 0 & n_2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} n\bar{y}_{n} \\ n_1\bar{y}_{1\cdot} \\ n_2\bar{y}_{2\cdot} \end{pmatrix} \\ &= \begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot}-\bar{y}_{3\cdot} \\ \bar{y}_{2\cdot}-\bar{y}_{3\cdot} \end{pmatrix}. \end{aligned}$$

Suppose we want to estimate $\beta_2 - \beta_3$ in design X. Then $\lambda = (0,1,-1,0)'$, and the least squares estimator can be constructed as,

$$\begin{aligned} \lambda'T\hat{\gamma} &= (0,1,-1,0)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot}-\bar{y}_{3\cdot} \\ \bar{y}_{2\cdot}-\bar{y}_{3\cdot} \end{pmatrix} \\ &= (0,1,-1,0)\begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot}-\bar{y}_{3\cdot} \\ \bar{y}_{2\cdot}-\bar{y}_{3\cdot} \\ 0 \end{pmatrix} \\ &= \bar{y}_{1\cdot} - \bar{y}_{2\cdot} \end{aligned}$$

Note that a least squares solution in design X is given by,

$$\hat{\beta} = T\hat{\gamma} = \begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot}-\bar{y}_{3\cdot} \\ \bar{y}_{2\cdot}-\bar{y}_{3\cdot} \\ 0 \end{pmatrix},$$

but in fact any value for the last component gives the same estimator $\lambda'\hat{\beta} = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$. #

THEOREM. If $\delta'\gamma$ is estimable in the reparameterized model (i.e., $\delta \in \text{col}(W')$), then $\delta'S\beta$ is estimable in the original model, and its least squares estimator is $\delta'\hat{\gamma}$, where $\hat{\gamma} \in \{W'W\gamma = W'y\}$. #

Proof. Since $\delta \in \text{col}(W')$, $\exists a \in \mathbb{R}^n$ such that $\delta = W'a$. Then

$$S'\delta = S'W'a = (WS)'a = X'a \in \text{col}(X'),$$

so $\delta'S\beta = (S'\delta)'\beta$ is estimable in the original model. Moreover, we have proven a previous theorem that shows

$$\hat{\beta} := T\hat{\gamma} \in \{X'X\beta = X'y\}, \qquad \forall\, \hat{\gamma} \in \{W'W\gamma = W'y\}.$$

Therefore, by the invariance of the least squares estimator to the least squares solution (see a previous theorem),

$$\delta'\hat{\gamma} = a'W\hat{\gamma} = a'XT\hat{\gamma} = a'WST\hat{\gamma} = \delta'S\hat{\beta}$$

is the least squares estimator of $\delta'S\beta$. #

EXAMPLE (continued). Since every function $\delta'\gamma$ is estimable in the full column rank reparameterization,

$$W\gamma = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 \\ 1_{n_2} & 0 & 1_{n_2} \\ 1_{n_3} & 0 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix},$$

$(S'\delta)'\beta = \delta'S\beta$ is estimable with design X, $\forall\, \delta \in \mathbb{R}^q$, where

$$X = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & 0 \\ 1_{n_2} & 0 & 1_{n_2} & 0 \\ 1_{n_3} & 0 & 0 & 1_{n_3} \end{pmatrix}, \qquad S\beta = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1+\beta_4 \\ \beta_2-\beta_4 \\ \beta_3-\beta_4 \end{pmatrix},$$

and so

$$\delta'S\hat{\beta} = \delta'\hat{\gamma} = \delta' \begin{pmatrix} \bar{y}_{3\cdot} \\ \bar{y}_{1\cdot} - \bar{y}_{3\cdot} \\ \bar{y}_{2\cdot} - \bar{y}_{3\cdot} \end{pmatrix}.$$

Take $\delta \in \{e_1, e_2, e_3\} \subseteq \mathbb{R}^3$ to estimate $\beta_1+\beta_4$, $\beta_2-\beta_4$, and $\beta_3-\beta_4$.

Notice that in this example, there are vectors λ such that $\lambda'T\gamma$ is estimable in the reparameterized model (i.e., $T'\lambda \in \text{col}(W')$), but λ is not estimable in the original model (i.e., $\lambda \notin \text{col}(X')$). For instance, take $\lambda = e_1 \in \mathbb{R}^4$. Then

$$\lambda' T \gamma = (1,0,0,0)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \gamma_1,$$

but $\beta_1 = \lambda'\beta$ and $\lambda \neq \text{col}(X')$. #

In summary, $E(Y) = X\beta = WS\beta = W\gamma = XT\gamma$, and the goal is to find a full column rank reparameterization of the original model so that the solution to the normal equations is unique.

EXAMPLE. Sometimes there may be reasons for a full rank reparameterization of a full rank model. Consider the transformation of centering the covariate in a simple linear regression;

$$y_i = \beta_0 + x_i\beta_1 + u_i$$

reparameterized as,

$$y_i = \gamma_0 + (x_i - \bar{x}_n)\gamma_1 + u_i.$$

Under what conditions do both of these models have a full column rank design?

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad W = \begin{pmatrix} 1 & x_1 - \bar{x}_n \\ 1 & x_2 - \bar{x}_n \\ \vdots & \vdots \\ 1 & x_n - \bar{x}_n \end{pmatrix},$$

so $\text{rank}(X) = \text{rank}(W) = 2$ if $x_j \neq x_i$ for some $i, j \in \{1, \ldots, n\}$. Moreover, observe that $W = XT$ and $X = WS$ for

$$T = \begin{pmatrix} 1 & -\bar{x}_n \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 1 & \bar{x}_n \\ 0 & 1 \end{pmatrix}.$$

An advantage of using design W is that the least squares estimates of the intercept, γ_0, and slope, γ_1, are uncorrelated (assuming $E(YY')$ exists).

$$\hat{\gamma} = (W'W)^{-1}W'y = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_i (x_i - \bar{x}_n)^2} \end{pmatrix}\begin{pmatrix} n\bar{y}_n \\ \sum_i (x_i - \bar{x}_n) y_i \end{pmatrix} = \begin{pmatrix} \bar{y}_n \\ \frac{\sum_i (x_i - \bar{x}_n) y_i}{\sum_i (x_i - \bar{x}_n)^2} \end{pmatrix}$$

whereas,

$$\hat{\beta} = T\hat{\gamma} = \begin{pmatrix} 1 & -\bar{x}_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_n \\ \frac{\sum_i (x_i - \bar{x}_n) y_i}{\sum_i (x_i - \bar{x}_n)^2} \end{pmatrix} = \begin{pmatrix} \bar{y}_n - \bar{x}_n \frac{\sum_i (x_i - \bar{x}_n) y_i}{\sum_i (x_i - \bar{x}_n)^2} \\ \frac{\sum_i (x_i - \bar{x}_n) y_i}{\sum_i (x_i - \bar{x}_n)^2} \end{pmatrix}. \quad \#$$

IMPOSING CONDITIONS FOR A UNIQUE SOLUTION TO THE NORMAL EQUATIONS

In the general linear model setup, $Y = X\beta + U$, with $X \in \mathbb{R}^{n \times p}$ and $r := \text{rank}(X)$, if $r = p$, then the unique solution to the normal equations is

$$\hat{\beta} := (X'X)^{-1}X'y.$$

However, if $r < p$, then

$$\{X'X\beta = X'y\} = \{X^g y + (I_p - X^g X)z : z \in \mathbb{R}^p\}.$$

We can always find a unique solution if we reparameterize to a full column rank design, but how about constraining in the original parameterization to obtain a unique solution?

In the one-way ANOVA model, $y_{ij} = \mu + \alpha_i + u_{ij}$, adding conditions to the solution to the normal equations yields a unique solution. Common choices of conditions are,

(1) $\alpha_i = 0$, for some $i \in \{1, \ldots, k\}$, where k is the number of groups.

(2) $\sum_{i=1}^{k} \alpha_i = 0$

(3) $\sum_{i=1}^{k} n_i \alpha_i = 0$

Recall that the normal equations, here, are of the form,

$$\begin{pmatrix} n & n_1 & n_2 & \cdots & n_k \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & \cdots & n_k \end{pmatrix} \beta = X'X\beta = X'y = \begin{pmatrix} n\bar{y}_{n\cdot} \\ n_1\bar{y}_{1\cdot} \\ n_2\bar{y}_{2\cdot} \\ \vdots \\ n_k\bar{y}_{k\cdot} \end{pmatrix}$$

which gives,

$$
\begin{aligned}
n\mu + \sum_{i=1}^{k} n_i \alpha_i &= n\,\bar{y}_{n} \\
n_1\mu + n_1\alpha_1 &= n_1\,\bar{y}_{1\cdot} \\
n_2\mu + n_2\alpha_2 &= n_2\,\bar{y}_{2\cdot} \\
&\vdots \\
n_k\mu + n_k\alpha_k &= n_k\,\bar{y}_{k\cdot}
\end{aligned}
$$

Imposing the third constraint yields the unique solution,

$$
\begin{aligned}
\hat{\mu} &= \bar{y}_{n} \\
\hat{\alpha}_1 &= \bar{y}_{1\cdot} - \bar{y}_{n} \\
\hat{\alpha}_2 &= \bar{y}_{2\cdot} - \bar{y}_{n} \\
&\vdots \\
\hat{\alpha}_k &= \bar{y}_{k\cdot} - \bar{y}_{n}\,.
\end{aligned}
$$

The question becomes whether there always exist conditions that afford a unique solution. Accordingly, consider constraint equations of the form,

$$
C\beta = 0\,,
$$

where $C \in \mathbb{R}^{(p-r)\times p}$ and $\text{rank}(C) = p - r$. Then augment the normal equations as,

$$
\begin{pmatrix} X'X \\ C \end{pmatrix} \beta = \begin{pmatrix} X'y \\ 0 \end{pmatrix},
$$

or equivalently, since $\{X'X\beta = X'y\} = \{X\beta = P_X y\}$,

$$
\begin{pmatrix} X \\ C \end{pmatrix} \beta = \begin{pmatrix} P_X y \\ 0 \end{pmatrix}.
$$

If $\text{rank}\begin{pmatrix} X \\ C \end{pmatrix} = p$, then we are done. This is equivalent to $\text{rank}(X', C') = p$ and

$$
\text{col}(X', C') = \mathbb{R}^p.
$$

That being so, the columns of C' (i.e., the rows of C) must not be orthogonal to the $\text{null}(X)$ because they are independent of the columns of X', and $\text{col}(X') \oplus \text{null}(X) = \mathbb{R}^p$. Hence, $C'\beta$ is a collection of non-estimable functions. More concisely,

$$
\text{col}(X') \cap \text{col}(C') = \{0\}.
$$

EXAMPLE. Consider again, the one-way ANOVA model, with three groups and design,

$$X = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & 0 \\ 1_{n_2} & 0 & 1_{n_2} & 0 \\ 1_{n_3} & 0 & 0 & 1_{n_3} \end{pmatrix}.$$

A basis for $col(X')$ is

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

Adding the constraint vector $C = (0, 1, 1, 1)$ is such that

$$col(X', C') = col \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} = \mathbb{R}^4.$$

Note that $0 = C\beta = \sum_{i=1}^{k} \alpha_i$ which is listed as constraint equation (2). #

LEMMA. Let $C \in \mathbb{R}^{(p-r) \times p}$ with $rank(C) = p - r$ and $col(X') \cap col(C') = \{0\}$. Then the following systems of equations are equivalent.

(1) $\begin{pmatrix} X'X \\ C'C \end{pmatrix} \beta = \begin{pmatrix} X'y \\ 0 \end{pmatrix}$

(2) $\begin{pmatrix} X'X \\ C \end{pmatrix} \beta = \begin{pmatrix} X'y \\ 0 \end{pmatrix}$

(3) $(X'X + C'C)\beta = X'y$ #

Proof. Using the fact that for any matrix C, $null(C) = null(C'C)$, it follows that

$$\beta \in \{C'C\beta = 0\} \quad \text{if and only if} \quad \beta \in \{C\beta = 0\}.$$

Thus, (1) and (2) are equivalent.

System (1) implies $X'X\beta = X'y$ and $C'C\beta = 0$, which gives $(X'X + C'C)\beta = X'y$, as in (3).

From (3) it follows that $C'C\beta = X'(y - X\beta) \in col(X') \cap col(C') = \{0\}$. That is, $C'C\beta = 0$ and $X'X\beta = X'y$, which is precisely the system (1). #

THEOREM. Let $C \in \mathbb{R}^{(p-r)\times p}$ with $\text{rank}(C) = p - r$ and $\text{col}(X') \cap \text{col}(C') = \{0\}$. Then

(1) The matrix $X'X + C'C$ is nonsingular.

(2) $(X'X + C'C)^{-1}X'y$ uniquely solves $X'X\beta = X'y$ and $C'C\beta = 0$.

(3) $(X'X + C'C)^{-1}$ is a generalized inverse of $X'X$.

(4) $C(X'X + C'C)^{-1}X' = 0$.

(5) $C(X'X + C'C)^{-1}C' = I$. #

Proof.

(1) By a homework problem, $p = \text{rank}(X', C') = \text{rank}\left[(X', C')\begin{pmatrix} X \\ C \end{pmatrix}\right] = \text{rank}(X'X + C'C)$. To complete the argument, note that,

$$X'X + C'C \in \mathbb{R}^{p\times p}.$$

(2) By (1), $(X'X + C'C)^{-1}$ exists, so the solution to $(X'X + C'C)\beta = X'y$ is uniquely given by $(X'X + C'C)^{-1}X'y$. Recall the equivalence of (1) and (3) from the Lemma.

(3) By (2), $X'X(X'X + C'C)^{-1}X'y = X'y$, $\forall y \in \mathbb{R}^n$. Thus,

$$X'X(X'X + C'C)^{-1}X' = X',$$

and so $X'X(X'X + C'C)^{-1}X'X = X'X$.

(4) Similarly, by (2), $C'C(X'X + C'C)^{-1}X'y = 0$, $\forall y \in \mathbb{R}^n$. Thus,

$$C'C(X'X + C'C)^{-1}X' = 0.$$

(5) Homework problem. #

CONSTRAINED PARAMETER SPACE

Consider the general linear model, assuming $E(U) = 0$,

$$Y = X\beta + U,$$

restricted to $\beta \in \{P'\beta = \delta\}$, for some full column rank matrix P. This is called the restricted general linear model.

DEFINITION. The function $\lambda'\beta$ is estimable in the restricted model if and only if there exists a scalar c and a vector a such that

$$E(c+a'Y) = \lambda'\beta, \quad \forall \beta \in \{P'\beta = \delta\}. \qquad \#$$

Observe that if $\lambda'\beta$ is estimable in the unrestricted model (i.e., $\forall \beta \in \mathbb{R}^p$), then it is estimable in the restricted model (i.e., $\forall \beta$ in some subset of $\mathbb{R}^p$).

THEOREM. In the restricted model, $c+a'y$ is unbiased for $\lambda'\beta$ if and only if $\exists d$ such that $\lambda = X'a + Pd$ and $c = d'\delta$. $\#$

Proof. First suppose that $\exists d$ such that $\lambda = X'a + Pd$ and $c := d'\delta$. Then for any $\beta \in \{P'\beta = \delta\}$,

$$\begin{aligned} E(c+a'Y) &= c + a'X\beta \\ &= d'\delta + (X'a)'\beta \\ &= d'P'\beta + (X'a)'\beta \\ &= (Pd + X'a)'\beta \\ &= \lambda'\beta. \end{aligned}$$

Conversely, assume that $c+a'y$ is unbiased for $\lambda'\beta$, $\forall \beta \in \{P'\beta = \delta\}$. Let W be a matrix with columns comprising a basis for null(P'). Then since $\forall z$,

$$P'(\beta + Wz) = P'\beta,$$

$\beta \in \{P'\beta = \delta\}$ implies $\beta + Wz \in \{P'\beta = \delta\}$. Accordingly, $\forall \beta \in \{P'\beta = \delta\}$, $\forall z$,

$$\begin{aligned} \lambda'(\beta + Wz) &= c + a'X(\beta + Wz) \\ \lambda'\beta + \lambda'Wz &= \underbrace{c + a'X\beta}_{= \lambda'\beta \text{ since } \beta \in \{P'\beta = \delta\}} + a'XWz. \end{aligned}$$

Hence, $\lambda'Wz = a'XWz$, $\forall z$, which gives $(\lambda' - a'X)W = 0$. Equivalently stated, $\lambda - X'a \in$ null$(W') \perp$ col(W), and so $\lambda - X'a \in$ col(P). That is,

$$\lambda - X'a = Pd$$

for some vector d. Further, $\forall \beta \in \{P'\beta = \delta\}$, $c + a'X\beta = \lambda'\beta = (Pd + X'a)'\beta$, so that $c = d'P'\beta = d'\delta$. $\#$

The restricted normal equations (RNEs) are defined as

$$\begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix}$$

THEOREM. If $\delta \in \text{col}(P')$, then there exists a solution to the restricted normal equations. #

Proof. The right side of the restricted normal equations is of the form

$$\begin{pmatrix} X'y \\ P'\beta \end{pmatrix} = \begin{pmatrix} X' & 0 \\ 0 & P' \end{pmatrix}\begin{pmatrix} y \\ \beta \end{pmatrix} \in \text{col}\begin{pmatrix} X' & 0 \\ 0 & P' \end{pmatrix},$$

so if it can be shown that,

$$\text{col}\begin{pmatrix} X' & 0 \\ 0 & P' \end{pmatrix} \subseteq \text{col}\begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix},$$

then the result will have been established. This is equivalent to showing

$$\text{null}\begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix} \subseteq \text{null}\begin{pmatrix} X & 0 \\ 0 & P \end{pmatrix}.$$

Accordingly, let $v := (v_1', v_2')' \in \text{null}\begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}$, where

$$0 = \begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix} v = \begin{pmatrix} X'Xv_1 + Pv_2 \\ P'v_1 \end{pmatrix}.$$

Then $P'v_1 = 0$, and $0 = v_1'X'Xv_1 + v_2'P'v_1 = \|Xv_1\|^2$ so that $v_1 \in \text{null}(X)$. That being true, $X'Xv_1 + Pv_2 = 0$ gives $v_2 \in \text{null}(P)$. Therefore,

$$\begin{pmatrix} X & 0 \\ 0 & P \end{pmatrix} v = \begin{pmatrix} Xv_1 \\ Pv_2 \end{pmatrix} = 0,$$

and so $\forall \beta \in \{P'\beta = \delta\}$,

$$\begin{pmatrix} X'y \\ \delta \end{pmatrix} \subseteq \text{col}\begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}.$$

#

THEOREM. If $\hat{\beta}_H$ denotes the first component of a solution,

$$\begin{pmatrix} \hat{\beta}_H \\ \hat{\theta}_H \end{pmatrix} \in \left\{ \begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix} \right\},$$

then $\hat{\beta}_H$ minimizes $Q(\beta) = \|y - X\beta\|_2^2$ over the constrained space $\{P'\beta = \delta\}$. #

Proof. Let $\beta \in \{P'\beta = \delta\}$. Then

$$\begin{aligned} Q(\beta) &= \|y - X\hat{\beta}_H + X\hat{\beta}_H - X\beta\|_2^2 \\ &= \|y - X\hat{\beta}_H\|_2^2 + 2(X\hat{\beta}_H - X\beta)'(y - X\hat{\beta}_H) + \|X\hat{\beta}_H - X\beta\|_2^2 \\ &= Q(\hat{\beta}_H) + 2(\hat{\beta}_H - \beta)'X'(y - X\hat{\beta}_H) + \|X(\hat{\beta}_H - \beta)\|_2^2 \\ &= Q(\hat{\beta}_H) - 2(\hat{\beta}_H - \beta)'P\hat{\theta}_H + \|X(\hat{\beta}_H - \beta)\|_2^2 \\ &= Q(\hat{\beta}_H) - 2\underbrace{(P'\hat{\beta}_H - P'\beta)}_{=0}\hat{\theta}_H + \|X(\hat{\beta}_H - \beta)\|_2^2 . \end{aligned}$$

Hence, $Q(\beta) \geq Q(\hat{\beta}_H)$, $\forall \beta \in \{P'\beta = \delta\}$, and $Q(\beta) = Q(\hat{\beta}_H)$ if $\beta = \hat{\beta}_H$. #

Note that the difference between the systems of equations,

$$\begin{pmatrix} X'X \\ C \end{pmatrix}\beta = \begin{pmatrix} X'y \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix},$$

is the first only includes nonestimable constraints while the second includes both nonestimable and estimable constraints. Recall that every row of C is linearly independent and not contained in $\text{col}(X')$ (i.e. $\text{col}(X') \cap \text{col}(C') = \{0\}$). This need not be the case with the matrix P. We may encounter estimable constraints based on contextual knowledge or otherwise.

THEOREM. Let $\hat{\beta}_H$ denote the first component of a solution,

$$\begin{pmatrix} \hat{\beta}_H \\ \hat{\theta}_H \end{pmatrix} \in \left\{ \begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix} \right\}.$$

If $\beta \in \{P'\beta = \delta\}$, then $Q(\beta) = Q(\hat{\beta}_H)$ if and only if β also solves the RNEs. #

Proof. First suppose that $\beta \in \{P'\beta = \delta\}$ and $Q(\beta) = Q(\hat{\beta}_H)$. Then by the previous theorem, $\|X(\hat{\beta}_H - \beta)\|_2^2 = 0$, and so,

$$X'y = X'X\hat{\beta}_H + P\hat{\theta}_H = X'X\beta + P\hat{\theta}_H .$$

For the reverse direction, assume that $(\beta', \theta')'$ solves the RNEs. Then

$$X'X\beta + P\theta = X'y = X'X\hat{\beta}_H + P\hat{\theta}_H ,$$

which gives $X'X(\hat{\beta}_H - \beta) = P(\theta - \hat{\theta}_H)$, and so

$$
\begin{aligned}
Q(\beta) - Q(\hat{\beta}_H) &= \|X(\hat{\beta}_H - \beta)\|_2^2 \\
&= (\hat{\beta}_H - \beta)' X'X (\hat{\beta}_H - \beta) \\
&= (\hat{\beta}_H - \beta)' P (\theta - \hat{\theta}_H) \\
&= \underbrace{(P'\hat{\beta}_H - P'\beta)'}_{=0} (\theta - \hat{\theta}_H)
\end{aligned}
$$

#

CHAPTER 4. GAUSS-MARKOV MODEL

The motivation for this chapter is to understand additional inferences that ensue from making various assumptions on the variance of the errors in the general linear model. We begin with the Gauss-Markov assumptions/model,

$$Y = X\beta + U,$$

with $E(U) = 0 \in \mathbb{R}^n$, $\text{Var}(U) = \sigma^2 I_n$, and $\sigma \in \mathbb{R}_+$. In particular, $E(U_i) = 0$ for $i \in \{1, \ldots, n\}$, and the errors are homoskedastic; that is,

$$\text{Cov}(U_i, U_j) = \begin{cases} \sigma^2 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}.$$

EXAMPLE. Suppose that $\lambda'\beta$ is an estimable function. Then for $\hat{\beta} \in \{X'X\beta = X'y\}$,

$$
\begin{aligned}
\text{Var}(\lambda'\hat{\beta}) &= \lambda' \text{Var}\left[X^g Y + (I_p - X^g X) z\right] \lambda \\
&= \lambda' X^g \text{Var}(Y) (X^g)' \lambda \\
&= \sigma^2 \lambda' (X'X)^g X' X (X'X)^g \lambda \\
&= \sigma^2 a' \underbrace{X (X'X)^g X' X}_{=X} (X'X)^g X' a, \quad \text{for some } a, \text{ since } \lambda \in \text{col}(X') \\
&= \sigma^2 a' X (X'X)^g X' a \\
&= \sigma^2 \lambda' (X'X)^g \lambda.
\end{aligned}
$$

#

It turns out that the least squares estimate plays an important role under the Gauss-Markov assumptions, as described by the following result.

THEOREM. (Gauss-Markov theorem) Under the Gauss-Markov assumptions, if $\lambda'\beta$ is estimable, then $\lambda'\hat{\beta}$ is the best (minimum variance) linear unbiased estimator (BLUE) of $\lambda'\beta$, where $\hat{\beta} \in \{X'X\beta = X'y\}$. #

Proof. Assume that $c + d'y$ is another unbiased estimator of $\lambda'\beta$. Then, $\forall \beta \in \mathbb{R}^p$,

$$\lambda'\beta = E(c + d'Y) = c + d'X\beta.$$

Choosing $\beta = 0$ demonstrates that $c = 0$. Accordingly, $\lambda = X'd$, and

$$\begin{aligned} \mathrm{Var}(c + d'Y) &= \mathrm{Var}(d'Y) \\ &= \mathrm{Var}(\lambda'\hat{\beta} + d'Y - \lambda'\hat{\beta}) \\ &= \mathrm{Var}(\lambda'\hat{\beta}) + 2\,\mathrm{cov}(\lambda'\hat{\beta}, d'Y - \lambda'\hat{\beta}) + \mathrm{Var}(d'Y - \lambda'\hat{\beta}). \end{aligned}$$

Next,

$$\begin{aligned} \mathrm{cov}(\lambda'\hat{\beta}, d'Y - \lambda'\hat{\beta}) &= \mathrm{Cov}(\lambda'(X'X)^{g}X'Y, d'Y - \lambda'(X'X)^{g}X'Y) \\ &= \lambda'(X'X)^{g}X' \cdot \mathrm{Var}(Y) \cdot (d - X(X'X)^{g}\lambda) \\ &= \sigma^2 \lambda'(X'X)^{g}X'(d - X(X'X)^{g}\lambda) \\ &= \sigma^2 (\lambda'(X'X)^{g}\lambda - \lambda'(X'X)^{g}\lambda) \\ &= 0. \end{aligned}$$

Hence,

$$\mathrm{Var}(c + d'Y) = \mathrm{Var}(\lambda'\hat{\beta}) + \mathrm{Var}(d'Y - \lambda'\hat{\beta}) \geq \mathrm{Var}(\lambda'\hat{\beta})$$

with equality if and only if

$$\begin{aligned} 0 &= \mathrm{Var}(d'Y - \lambda'\hat{\beta}) \\ &= (d' - \lambda'(X'X)^{g}X')\,\mathrm{Var}(Y)\,(d' - \lambda'(X'X)^{g}X')' \\ &= \sigma^2 \| d - X(X'X)^{g}\lambda \|_2^2. \end{aligned}$$

That is, $\mathrm{Var}(c + d'Y) = \mathrm{Var}(\lambda'\hat{\beta})$ if and only if $d = X(X'X)^{g}\lambda$, in which case

$$d'y = \lambda'(X'X)^{g}X'y = \lambda'\hat{\beta},$$

so that $\lambda'\hat{\beta}$ is the unique BLUE of $\lambda'\beta$. #

VARIANCE ESTIMATION

Recall the unique decomposition,

$$\begin{aligned} Y &= P_X Y + (I - P_X) Y \\ &= X\hat{\beta} + (I - P_X)(X\beta + U) \\ &= X\hat{\beta} + (I - P_X) U \end{aligned}$$

for any $\hat{\beta} \in \{X'X\beta = X'Y\} = \{X\beta = P_X Y\}$. Since $P_X Y$ is related to the estimation of β, in the decomposition, the intuition is that $(I - P_X)Y$ is related to the estimation of σ^2.

THEOREM. Under the Gauss-Markov assumptions, $\hat{\sigma}^2 := Y'(I - P_X)Y/(n-r)$ is an unbiased estimator of σ^2, where $r := \text{rank}(X) \leq p$. #

Proof.

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{1}{n-r} E\left[(X\beta + U)'(I - P_X)(X\beta + U)\right] \\ &= \frac{1}{n-r} E\left[U'(I - P_X)U\right] \\ &= \frac{1}{n-r} E\left[\text{tr}(U'(I - P_X)U)\right] \\ &= \frac{1}{n-r} \text{tr}\left[(I - P_X) E(UU')\right] \\ &= \frac{1}{n-r} \text{tr}(I - P_X) \cdot \sigma^2 \\ &= \sigma^2. \end{aligned}$$

#

Note that we have not shown that $\sqrt{\hat{\sigma}^2}$ is an unbiased estimator for σ.

IMPLICATIONS ON MODEL SELECTION

Here we make a distinction between the true data generating model, and the model being used. In the context of a linear model, two discrepencies are

(1) Underfitting (missing covariates)

(2) Overfitting (redundant covariates)

MISPECIFICATION FROM UNDERFITTING

$$Y = \underbrace{X\beta}_{\text{model used by the practitioner}} + \eta + u$$

Using the practitioner model, $Y = X\beta + \tilde{u}$, under the Gauss-Markov assumptions results in an omitted variable bias. Consider the effect on the least squares estimator. For $\lambda \in \text{col}(X')$,

$$\begin{aligned} E(\lambda'\hat{\beta}) &= \lambda'(X'X)^{g}X'(X\beta+\eta) \\ &= a'X(X'X)^{g}X'(X\beta+\eta) \\ &= \lambda'\beta + \underbrace{a'P_x\eta}_{\text{mispecification bias}} \end{aligned}$$

Observe that if $\eta \perp \text{col}(X)$, then $\lambda'\hat{\beta}$ is unbiased. How about estimation of σ^2?

$$\begin{aligned} E(Y'(I-P_x)Y) &= (X\beta+\eta)'(I-P_x)(X\beta+\eta) + \sigma^2 \cdot tr(I-P_x) \\ &= \eta'(I-P_x)\eta + \sigma^2 \cdot (n-r) \end{aligned}$$

That is, if $\eta \in \text{col}(X)$ then $\hat{\sigma}^2$ is unbiased. In the case that $\eta \in \text{col}(X)$, $\exists \gamma$ such that

$$Y = X\beta + \eta + u = X(\beta+\gamma) + u.$$

OVERFITTING AND MULTICOLLINEARITY

Practitioner model:

$$Y = X\beta + W\gamma + u,$$

where $\gamma = 0$ in the true data generating model. Assume that the concatenated matrix (X,W) has full column rank. Then under the Gauss-Markov assumptions,

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} X'X & X'W \\ W'X & W'W \end{pmatrix}^{-1} \begin{pmatrix} X'Y \\ W'Y \end{pmatrix}$$

with

$$E\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} \beta \\ 0 \end{pmatrix} \qquad \text{and} \qquad \text{Var}\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \sigma^2 \begin{pmatrix} X'X & X'W \\ W'X & W'W \end{pmatrix}^{-1}.$$

It can be worked out that,

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1} + \sigma^2(X'X)^{-1}X'W(W'(I-P_X)W)^{-1}W'X(X'X)^{-1},$$

where the second term can be understood as the increase in variability of the least squares estimates that results from including redundant covariates.

However, estimation of σ^2 remains unbiased;

$$E\left[Y'(I-P_{X,W})Y\right] = \sigma^2\left[n - \text{rank}(X,W)\right].$$

Notice that if the columns of W are not only independent but are also orthogonal to the columns of X, then the variance penalty is zero. In this case, there is no cost to overfitting. More generally, though, if W is not orthogonal to X, then multicollinearity ensues.

A measure to assess the influence of multicollinearity relates to the SVD of (X,W).

DEFINITION. The mean squared error (MSE) of an estimator $\hat{\theta}$ for some parameter θ is

$$\text{MSE}(\hat{\theta}) := E\left[\|\hat{\theta}-\theta\|_2^2\right]. \qquad \#$$

For some unbiased estimator $\tilde{\beta}$ of β,

$$\begin{aligned}\text{MSE}(\tilde{\beta}) &= E\left[\|\tilde{\beta}-\beta\|_2^2\right] \\ &= \text{tr}\left(E\left[(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)'\right]\right) \\ &= \text{tr}\left(\text{Var}(\tilde{\beta})\right).\end{aligned}$$

Further, denoting the SVD of the design matrix as $X = U\begin{pmatrix}\Lambda \\ 0\end{pmatrix}V'$,

$$\begin{aligned}\text{MSE}(\hat{\beta}) &= \text{tr}\left(\text{Var}(\hat{\beta})\right) \\ &= \sigma^2 \cdot \text{tr}\left[(X'X)^{-1}\right] \\ &= \sigma^2 \cdot \text{tr}\left(V\Lambda^{-2}V'\right) \\ &= \sigma^2 \sum_1^p \lambda_i^{-2},\end{aligned}$$

where λ_i is the i-th eigenvalue of $X'X$. In the case that multicollinearity is severe, $X'X$ is nearly singular. As such, $\min\{\lambda_i\}$ will be close to zero, and so $\text{MSE}(\hat{\beta})$ will be excessively large. Another measure of multicollinearity is the condition number,

$$\frac{\max\{\lambda_i\}}{\min\{\lambda_i\}}.$$

Alternatively, there is a residual sum of squares approach to assessing the severity of multicollinearity, called variance inflation factors.

THE AITKEN MODEL AND GENERALIZED LEAST SQUARES

The Aitken model is a relaxation of certain assumptions in the Gauss-Markov model. Namely, the Aitken assumptions are,

$$Y = X\beta + U,$$

with $E(U) = 0$, $\text{Var}(U) = \sigma^2 \cdot V$, $\sigma \in \mathbb{R}_+$, and V some known symmetric positive-definite matrix. Under these conditions, the least squares estimator may not be the BLUE for an estimable function $\lambda'\beta$. Accordingly, we will construct a generalized least squares (GLS) estimator for $\lambda'\beta$ that is the BLUE.

The idea is to decompose $V = LL'$ for some positive-definite matrix L. Two options are:

(1) Cholesky factorization, where L is lower triangular.

(2) Spectral decomposition, $V = Q\Lambda Q' = \underbrace{Q\Lambda^{1/2}Q'}_{=L} \cdot \underbrace{Q\Lambda^{1/2}Q'}_{=L'}$

Then the transformed model

$$L^{-1}Y = L^{-1}X\beta + L^{-1}U$$

satisfies $E(L^{-1}U) = L^{-1}E(U) = 0$ and

$$\begin{aligned}\text{Var}(L^{-1}U) &= L^{-1}\text{Var}(U)(L^{-1})' \\ &= \sigma^2 L^{-1} L L' (L')^{-1} \\ &= \sigma^2 \cdot I_n.\end{aligned}$$

Recall this is precisely the Gauss-Markov model. That being so, for any estimable function $\lambda'\beta$, the GLS estimator $\lambda'\hat{\beta}_{GLS}$ is the BLUE, where

$$\hat{\beta}_{GLS} \in \{X'V^{-1}X\beta = X'V^{-1}Y\}.$$

The system, $X'V^{-1}X\beta = X'V^{-1}Y$ is called the Aitken equations. Solutions to the

Aitken equations are called GLS or weighted least squares solutions. Observe the condition for estimability, $\lambda \in \text{col}(X'(L^{-1})') = \text{col}(X')$ since L^{-1} is nonsingular.

THEOREM. (Aitken's theorem) Under the Aitken assumptions, if $\lambda'\beta$ is estimable, then $\lambda'\hat{\beta}_{GLS}$ is the BLUE for $\lambda'\beta$, where $\hat{\beta}_{GLS} \in \{X'V^{-1}X\beta = X'V^{-1}Y\}$. #

Proof. Homework problem. #

For estimation of σ^2, the estimator used in the Gauss-Markov model remains unbiased in the transformed Aitken model. That is,

$$\hat{\sigma}^2_{GLS} := \frac{1}{n-r} \cdot (Y - X\hat{\beta}_{GLS})' V^{-1} (Y - X\hat{\beta}_{GLS}).$$

EXAMPLE. Heteroskedasticity. Let $y_i = x_i\beta + u_i$ with $E(u_i) = 0$, $\text{Var}(u_i) = \sigma^2 x_i$, and u_i uncorrelated with x_i. Then

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \beta + \sigma^2 \underbrace{\begin{pmatrix} 1/x_1 & & \\ & \ddots & \\ & & 1/x_n \end{pmatrix}}_{= L^{-1}} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

satisfies the Gauss-Markov assumptions. #

THEOREM. The estimator $t'Y$ is the BLUE for $E(t'Y)$ if and only if $t'Y$ is uncorrelated with all unbiased estimators of zero. #

Proof. First assume that $t'Y$ is uncorrelated with all unbiased estimators of zero, and let $a'Y$ be an unbiased estimator of $E(t'Y)$. Then

$$\begin{aligned} \text{Var}(a'Y) &= \text{Var}(t'Y + a'Y - t'Y) \\ &= \text{Var}(t'Y) + \text{Var}(a'Y - t'Y) + 2\,\text{Cov}(t'Y, a'Y - t'Y). \end{aligned}$$

Since $E(a'Y - t'Y) = E(a'Y) - E(t'Y) = 0$, by assumption, $\text{Cov}(t'Y, a'Y - t'Y) = 0$. Hence, $\text{Var}(a'Y) \geq \text{Var}(t'Y)$.

For the other direction, assume that $t'Y$ is the BLUE for $E(t'Y)$, and let $a'Y$ be an unbiased estimator of zero, for $a \neq 0$. Next, define

$$s'Y = t'Y - \frac{\text{Cov}(t'Y, a'Y)}{\text{Var}(a'Y)} \cdot a'Y.$$

By assumption, $E(s'Y) = E(t'Y)$, and

$$\mathrm{Var}(s'Y) = \mathrm{Var}(t'Y) + \frac{\mathrm{Cov}^2(t'Y, a'Y)}{\mathrm{Var}(a'Y)} - 2\frac{\mathrm{Cov}(t'Y, a'Y)}{\mathrm{Var}(a'Y)} \cdot \mathrm{Cov}(t'Y, a'Y)$$

$$= \mathrm{Var}(t'Y) - \underbrace{\frac{\mathrm{Cov}^2(t'Y, a'Y)}{\mathrm{Var}(a'Y)}}_{\geq 0}.$$

So $\mathrm{Var}(s'Y) \leq \mathrm{Var}(t'Y) \leq \mathrm{Var}(s'Y)$, where the second inequality follows by assumption that $t'Y$ is the BLUE for $E(t'Y)$. Therefore,

$$\mathrm{Var}(t'Y) = \mathrm{Var}(t'Y) - \frac{\mathrm{Cov}^2(t'Y, a'Y)}{\mathrm{Var}(a'Y)},$$

and so $\mathrm{Cov}(t'Y, a'Y) = 0$. #

COROLLARY. Under the Aitken assumptions, $t'Y$ is the BLUE for $E(t'Y)$ if and only if $Vt \in \mathrm{col}(X)$. #

Proof. From the previous theorem, $t'Y$ is the BLUE for $E(t'Y)$ if and only if $\mathrm{Cov}(t'Y, a'Y) = 0$ for every a such that $a'X\beta = E(a'Y) = 0$, $\forall \beta \in \mathbb{R}^p$. That is, $t'Y$ is the BLUE for $E(t'Y)$ if and only if $\mathrm{Cov}(t'Y, a'Y) = 0$, $\forall a \in \mathrm{null}(X')$. Moreover, $\mathrm{Cov}(t'Y, a'Y) = \sigma^2 \cdot t'Va$.

Accordingly, $t'Y$ is the BLUE for $E(t'Y)$ if and only if $t'Va = 0$, $\forall a \in \mathrm{null}(X')$, if and only if $Vt \perp \mathrm{null}(X')$, if and only if $Vt \in \mathrm{col}(X)$. #

The corollary gives a tool for determining whether $t'Y$ is the BLUE for $E(t'Y)$. The next result gives a condition for when $\lambda'\hat{\beta}_{OLS}$ remains the BLUE for an estimable $\lambda'\beta$ even under the Aitken assumptions.

THEOREM. Under the Aitken assumptions, $\lambda'\hat{\beta}_{OLS}$ is the BLUE for any estimable $\lambda'\beta$ if and only if $\exists Q$ such that $VX = XQ$. #

Proof. Suppose that $VX = XQ$ for some Q. Then

$$\lambda'\hat{\beta}_{OLS} = \lambda'(X'X)^{g}X'y = t'y,$$

for $t := X(X'X)^{g}\lambda$. Further,

$$Vt = VX(X'X)^{g}\lambda = XQ(X'X)^{g}\lambda \in \mathrm{col}(X),$$

and so $\lambda'\hat{\beta}_{OLS}$ is the BLUE for any estimable $\lambda'\beta$, by the corollary.

Conversely, assume that $\lambda'\hat{\beta}_{OLS}$ is the BLUE for any estimable $\lambda'\beta$. Next, define $\lambda_j := X'X_j \ \forall j \in \{1, \ldots, p\}$ so that for $t_j := X(X'X)^g\lambda_j$,

$$\lambda_j'\hat{\beta}_{OLS} = X_j'X(X'X)^gX'Y = t_j'Y.$$

That being true, $t_j'Y$ is the BLUE for $\lambda_j'\beta$, and by the corollary,

$$\begin{aligned} VX &= VX(X'X)^gX'(X_1, \ldots, X_p) \\ &= (Vt_1, \ldots, Vt_p) \\ &= (XQ_1, \ldots, XQ_p), \text{ for some } Q_1, \ldots, Q_p \in \mathbb{R}^p \\ &= XQ \end{aligned}$$

#

CHAPTER 5. DISTRIBUTIONAL THEORY

To this point we have considered the following assumptions on the general linear model, $Y = X\beta + U$:

(1) $E(U) = 0$ in the context of the least squares problem

(2) $E(U) = 0$ in the context of estimation

(3) $E(U) = 0$ and $Var(U) = \sigma^2 I_n$

(4) $E(U) = 0$ and $Var(U) = \sigma^2 V$, for some known positive-definite V.

In this chapter we consider the implications of a distributional assumption on U.

DEFINITION. A random vector $Y \in \mathbb{R}^p$ is said to follow the multivariate normal distribution with mean μ and covariance matrix Σ if $\forall v \in \mathbb{R}^p$ such that $v'\Sigma v \neq 0$

$$v'Y \sim N(v'\mu, v'\Sigma v).$$

Denote $Y \sim N_p(\mu, \Sigma)$. #

THEOREM. A random vector $Y \sim N_p(\mu, \Sigma)$ for some nonsingular matrix Σ if and only if Y has density,

$$f(y) = \det(2\pi\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}.$$

#

Proof. Exercise. #

Note that if $Z \sim N_p(0, I_p)$, then $Y := \mu + \Sigma^{1/2} Z \sim N_p(\mu, \Sigma)$.

Other than the density function, another quantity that uniquely defines the distribution of a random variable is the moment generating function.

DEFINITION. The moment generating function of a random vector $X \in \mathbb{R}^p$ is

$$m_X(t) := E(e^{t'X}), \quad t \in \mathbb{R}^p$$

provided that the expectation exists in a neighborhood of $t = 0$. #

Recall two important properties of moment generating functions:

(1) If the moment generating functions for two random variables X and Y exist, then they share identical CDFs if and only if $m_X(t) = m_Y(t)$ for every t in some neighborhood of zero.

(2) Let

$$X := \begin{pmatrix} X_1 \\ \vdots \\ X_K \end{pmatrix}$$

for some K. Then $X_1, \ldots, X_K$ are mutually independent if and only if

$$m_X(t) = m_{X_1}(t_1) \cdots m_{X_K}(t_K)$$

for every t in some neighborhood of zero.

Using (2), if $Z_1, \ldots, Z_p \overset{iid}{\sim} N(0, 1)$, then for $Z := (Z_1, \ldots, Z_p)'$,

$$\begin{aligned} m_Z(t) &= m_{Z_1}(t_1) \cdots m_{Z_p}(t_p) \\ &= e^{\frac{1}{2}t_1^2} \cdots e^{\frac{1}{2}t_p^2} \\ &= e^{\frac{1}{2}t't}. \end{aligned}$$

Then if $Y := \mu + \Sigma^{1/2} Z \sim N_p(\mu, \Sigma)$ for $Z \sim N_p(0, I_p)$,

$$m_Y(t) = E(e^{t'Y}) = e^{t'\mu} E(e^{t'\Sigma^{1/2} Z}) = e^{t'\mu} m_Z(\Sigma^{1/2} t) = e^{t'\mu + \frac{1}{2} t'\Sigma t}.$$

More generally, linear transformations of multivariate normal random variables can be summarized as in the following result.

THEOREM. If $X \sim N_p(\mu, \Sigma)$ and $Y = a + BX$ for some $a \in \mathbb{R}^q$ and $B \in \mathbb{R}^{q \times p}$, then

$$Y \sim N_q(a + B\mu, B\Sigma B')$$ #

Proof. Generalize the above argument. #

See the textbook for a variety of other properties of the multivariate normal distribution. Next, we will study quadratic forms of multivariate normal random variables.

DEFINITION. Let $Z \sim N_p(0, I_p)$. Then $U := Z'Z = \sum_1^p Z_i^2$ is said to follow the χ_p^2 distribution. The parameter p denotes the degrees of freedom. #

Some properties of the χ_p^2 distribution:

$$m_U(t) = (1-2t)^{-\frac{p}{2}} \quad \text{for } t < \tfrac{1}{2}$$

$$E(U) = p$$

$$\text{Var}(U) = 2p$$

Observe that the variance grows on the same order as the mean.

DEFINITION. Let $J \sim \text{Poisson}(\phi)$ and $U|J=j \sim \chi^2_{p+2j}$. Then (unconditionally) U is said to follow a noncentral chi-squared distribution with noncentrality parameter ϕ, denoted $U \sim \chi_p^2(\phi)$. #

We will see shortly that if $Z_1, \ldots, Z_p \overset{iid}{\sim} N(\mu_i, 1)$, then $U := \sum_1^p Z_i^2 \sim \chi_p^2(\phi)$, where

$$\phi = \tfrac{1}{2} \sum_1^p \mu_i^2.$$

The density for U can be derived as

$$f_U(u) = \sum_{j \geq 0} f_{U|J}(u|j) \cdot f_J(j) = \sum_{j \geq 0} \frac{u^{\frac{p+2j-2}{2}} e^{-\frac{u}{2}}}{\Gamma\left(\frac{p+2j}{2}\right) 2^{j+\frac{p}{2}}} \cdot \frac{e^{-\phi} \phi^j}{j!},$$

and the moment generating function is

$$m_U(t) = (1-2t)^{-\frac{p}{2}} \cdot e^{\frac{2\phi t}{1-2t}} \quad \text{for } t < \tfrac{1}{2}$$

$$E(U) = p + 2\phi$$

$$\text{Var}(U) = 2p + 8\phi$$

THEOREM. If $U_1, \ldots, U_m$ are mutually independent and $U_i \sim \chi^2_{p_i}(\phi_i)$, then

$$U := \sum_1^m U_i \sim \chi^2_p(\phi),$$

where $p := \sum_1^m p_i$ and $\phi := \sum_1^m \phi_i$. #

Proof.

$$\begin{aligned} m_U(t) &= E(e^{tU}) \\ &= E\left(e^{t\sum_1^m U_i}\right) \\ &= \prod_1^m E(e^{tU_i}) \\ &= (1-2t)^{-\frac{1}{2}\sum_1^m p_i} e^{\frac{2t\sum_1^m \phi_i}{1-2t}}, \quad t < \tfrac{1}{2} \end{aligned}$$ #

THEOREM. If $X \sim N(\mu, 1)$, then $X^2 \sim \chi^2_1(\frac{1}{2}\mu^2)$. #

Proof.

$$\begin{aligned} m_{X^2}(t) &= E(e^{tX^2}) \\ &= \int_{\mathbb{R}} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} dx \end{aligned}$$

Then "complete the square" to show $m_{X^2}(t) = (1-2t)^{-\frac{1}{2}} e^{\frac{2t(\frac{\mu^2}{2})}{1-2t}}$. #

THEOREM. If $X \sim N_p(\mu, I_p)$, then $X'X \sim \chi^2_p(\frac{1}{2}\mu'\mu)$. #

Proof. Since $\operatorname{cov}(X_i, X_j) = 0$ for every $i \neq j$, joint normality gives independence. Thus, by the previous theorem,

$$X_i^2 \sim \chi^2_1(\tfrac{1}{2}\mu^2), \quad \forall i \in \{1, \ldots, p\},$$

and so by two theorems previous, since $X_1^2, \ldots, X_p^2$ are independent,

$$X'X = \sum_1^p X_i^2 \sim \chi^2_p(\tfrac{1}{2}\mu'\mu).$$ #

COROLLARY. If $X \sim N_p(\mu, V)$ for some nonsingular V, then $X'V^{-1}X \sim \chi^2_p(\frac{1}{2}\mu'V^{-1}\mu)$. #

Proof. Rescale X and apply the previous theorem. #

DEFINITION. Let U_1 and U_2 be independent random variables with $U_1 \sim \chi^2_{p_1}$ and $U_2 \sim \chi^2_{p_2}$. Then the F distribution is defined as

$$\frac{U_1/p_1}{U_2/p_2} \sim F_{p_1,p_2}$$

#

DEFINITION. Let U_1 and U_2 be independent random variables with $U_1 \sim \chi^2_{p_1}(\phi)$ and $U_2 \sim \chi^2_{p_2}$. Then the noncentral F distribution is defined as

$$\frac{U_1/p_1}{U_2/p_2} \sim F_{p_1,p_2}(\phi)$$

#

DEFINITION. Let $X \sim N(\mu, 1)$, $U \sim \chi^2_p$, and X and U are independent, then the T distribution is defined as

$$\frac{X}{\sqrt{U/p}} \sim T_p(\mu).$$

#

Now we are ready to consider the Gaussian linear model,

$$Y = X\beta + U,$$

where $U \sim N_n(0, \sigma^2 I_n)$. The goal for the remainder of this chapter is to study the distributions of the components of,

$$Y'Y = Y'P_X Y + Y'(I - P_X)Y.$$

These are all quadratic forms of multivariate normal random vectors with projection matrices.

LEMMA. A $p \times p$ matrix A is symmetric and idempotent with rank s if and only if there exists a $p \times s$ matrix G with orthonormal columns such that $A = GG'$.

Proof. Lab problem. #

THEOREM. Let $X \sim N_p(\mu, I_p)$. If A is a symmetric and idempotent matrix with $\text{rank}(A) = s$, then

$$X'AX \sim \chi^2_s(\tfrac{1}{2}\mu'A\mu).$$

#

Proof. By the previous lemma, there exists a $p \times s$ matrix G with $G'G = I_s$

such that $A = GG'$. Then $G'X \sim N_s(G'\mu, I_s)$ so that by a previous theorem,

$$X'AX = (G'X)'G'X \sim \chi^2_p(\tfrac{1}{2}\mu'\underbrace{GG'}_{=A}\mu).$$

#

THEOREM. Let $X \sim N_p(\mu, V)$ with V positive definite. If A is a symmetric matrix and AV is idempotent with $\text{rank}(AV) = s$, then

$$X'AX \sim \chi^2_s(\tfrac{1}{2}\mu'A\mu).$$

#

Proof. Since V is symmetric, positive definite, $V = LL'$ for some $L > 0$. Then

$$Y := L^{-1}X \sim N_p(L^{-1}\mu, I_p),$$

so that $X'AX = (L^{-1}X)'L'AL\,L^{-1}X = Y'\underbrace{L'AL}_{=:B}Y$. Accordingly, since $B' = (L'AL)' = B$,

$$\begin{aligned} B \cdot B &= L'AL \cdot L'AL \\ &= L'AVAL \cdot L'(L')^{-1} \\ &= L'AVAV(L')^{-1} \\ &= L'AV(L')^{-1}, \quad \text{by assumption} \\ &= L'AL \\ &= B, \end{aligned}$$

and $\text{rank}(B) = \text{tr}(B) = \text{tr}(L'AL) = \text{tr}(AV) = \text{rank}(AV) = s$, by the previous theorem,

$$X'AX = Y'BY \sim \chi^2_s(\tfrac{1}{2} \cdot \mu'\underbrace{(L')^{-1}BL^{-1}}_{=(L')^{-1}L'ALL^{-1}=A}\mu).$$

#

Note that for the general linear model $Y \sim N_n(X\beta, \sigma^2 I_n)$,

$$\sigma^{-2} \cdot Y'(I - P_X)Y \sim \chi^2_{n-r},$$

since $\sigma^2 I_n > 0$, $\sigma^{-2} \cdot Y'(I - P_X)Y$ is symmetric, and

$$\sigma^{-2}(I - P_X) \cdot \sigma^2 I = I - P_X$$

is idempotent with rank $n - r$. The noncentrality parameter is zero because

$$\tfrac{1}{2}\mu'A\mu = \tfrac{1}{2\sigma^2}\beta'X'(I - P_X)X\beta = 0.$$

Similarly, $\sigma^{-2} \cdot Y'P_XY \sim \chi^2_r(\frac{1}{2\sigma^2}\beta'X'X\beta)$. Moreover,

$$\begin{pmatrix}\hat{Y}\\ \hat{u}\end{pmatrix} = \begin{pmatrix}P_x & 0\\ 0 & I-P_x\end{pmatrix} Y \sim N_{2n}\left[\begin{pmatrix}X\beta\\ 0\end{pmatrix}, \sigma^2\begin{pmatrix}P_x & 0\\ 0 & I-P_x\end{pmatrix}\right].$$

Since $\hat{Y}$ and $\hat{u}$ are jointly normal and have zero covariance, they are independent, and so

$$\frac{\|\hat{Y}\|^2/r}{\|\hat{u}\|^2/(n-r)} \sim F_{r,n-r}\left(\tfrac{1}{2\sigma^2}\|X\beta\|^2\right).$$

THEOREM. Let $X \sim N_p(\mu, V)$ and let A be a symmetric matrix with rank s. If $BVA = 0$ for a given matrix B, then BX and $X'AX$ are independent. #

Proof. Homework problem. #

COROLLARY. Let $X \sim N_p(\mu, V)$, let A be a symmetric matrix with rank r, and let B be a symmetric matrix with rank s. If $BVA = 0$, then $X'BX$ and $X'AX$ are independent. #

Proof. Homework problem. #

THEOREM. (Cochran's theorem) Let $Y \sim N_n(\mu, \sigma^2 I_n)$ and let A_i be a symmetric idempotent matrix with rank s_i for $i \in \{1, \dots, k\}$. If $\sum_1^k A_i = I_n$, then

$$\sigma^{-2} Y'A_iY \sim \chi^2_{s_i}\left(\tfrac{1}{2\sigma^2}\mu'A_i\mu\right),$$

$\sum_1^k s_i = n$, and $\sigma^{-2}Y'A_1Y, \dots, \sigma^{-2}Y'A_kY$ are independent. #

Proof. By the previous lemma, since the A_i are symmetric and idempotent there exists $n \times s_i$ matrices Q_i with $Q_i'Q_i = I_{s_i}$ such that $A_i = Q_iQ_i'$ $\forall i \in \{1, \dots, k\}$. Construct $Q := (Q_1, \dots, Q_k)$ with shape $n \times \sum_1^k s_i$, and observe that

$$QQ' = (Q_1, \dots, Q_k)\begin{pmatrix}Q_1'\\ \vdots\\ Q_k'\end{pmatrix} = Q_1Q_1' + \cdots + Q_kQ_k' = A_1 + \cdots + A_k = I_n.$$

Further, $n = \operatorname{tr}(I_n) = \operatorname{tr}(A_1) + \cdots + \operatorname{tr}(A_k) = \sum_1^k s_i$, and since Q is $n \times n$ and invertible

$$\begin{pmatrix}Q_1'Q_1 & \cdots & Q_1'Q_k\\ \vdots & \ddots & \vdots\\ Q_k'Q_1 & \cdots & Q_k'Q_k\end{pmatrix} = Q'Q = I_n,$$

so that $Q_i'Q_j = 0$ for every $i \neq j$. Note that $n \times n$ and $QQ' = I_n$ implies that

$\text{rank}(Q) = \text{rank}(QQ') = n$. That being true means Q is full rank, Q^{-1} exists, and

$$Q^{-1} \cdot QQ'Q = Q^{-1} \cdot Q.$$

Accordingly, $Q'Y \sim N_n(Q'\mu, \sigma^2 I_n)$, and so $Q_1'Y, \ldots, Q_k'Y$ are independent. Moreover, $\sigma^{-2}\|Q_1'Y\|^2, \ldots, \sigma^{-2}\|Q_k'Y\|^2$ are independent and $\sigma^{-1} Q_i'Y \sim N_{s_i}(\sigma^{-1}Q_i'\mu, I_{s_i})$. Therefore,

$$\sigma^{-2} Y'A_iY \sim \chi^2_{s_i}\left(\tfrac{1}{2\sigma^2}\mu'A_i\mu\right). \qquad \#$$

CHAPTER 6. STATISTICAL INFERENCE

Consider again the Gaussian linear model, $Y \sim N_n(X\beta, \sigma^2 I_n)$. Up to this point we have shown that the BLUE for an estimable $\lambda'\beta$ is $\lambda'\hat{\beta}$, for $\hat{\beta} \in \{X'X\beta = X'y\}$, and that

$$\lambda'\hat{\beta} \sim N\left(\lambda'\beta, \sigma^2 \lambda'(X'X)^g\lambda\right)$$

independent of

$$\hat{\sigma}^2 = \frac{1}{n-r} Y'(I - P_X)Y, \quad \text{where } \frac{(n-r)}{\sigma^2} \cdot \hat{\sigma}^2 \sim \chi^2_{n-r}.$$

The textbook presents a variety of other important properties of these estimators, one of which is the following theorem.

THEOREM. In the model $Y \sim N_n(X\beta, \sigma^2 I_n)$ with unknown parameters β and σ^2, the maximum likelihood estimators are $\hat{\beta} = (X'X)^g X'y$ and $\hat{\sigma}^2 = \frac{1}{n} y'(I - P_X)y$, respectively. $\#$

Proof. The log-likelihood of the data is given by

$$\ell(\beta, \sigma^2; y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} Q(\beta),$$

where $Q(\beta) = \|y - X\beta\|^2$. Maximizing with respect to β is equivalent to minimizing $Q(\beta)$ with respect to β. Accordingly, the MLE of β is

$$\hat{\beta} \in \text{argmin}\{Q(\beta)\} = \{X'X\beta = X'y\}.$$

Next,

$$0 = \frac{\partial}{\partial \sigma^2}\ell(\beta, \sigma^2; y) = -\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} Q(\hat{\beta})$$

gives $\hat{\sigma}^2 = \frac{1}{n}\|y - X\hat{\beta}\|^2$. Left as an exercise to verify the second-order conditions. $\#$

TESTING THE GENERAL LINEAR HYPOTHESIS

With the distributional assumption $Y \sim N_n(X\beta, \sigma^2 I_n)$ we can construct and test a variety of hypotheses. For example,

H_1: $\beta_j = 0$

H_2: $\beta_1 = \beta_2 = \beta_3 = 0$

H_3: $\beta_1 + \beta_3 = 1$, $\beta_2 = 3$

H_4: $\beta_2 = \beta_3 = \beta_4$

H_5: $\beta \in \text{col}(B)$

In particular, hypotheses we will consider are those of the form,

$$H: K'\beta = m \qquad \text{versus} \qquad H^c: K'\beta \neq m,$$

where K is a $p \times s$ full column rank matrix and $K \in \text{col}(X')$ so that $K'\beta$ is estimable.

DEFINITION. The general linear hypothesis $H: K'\beta = m$ is said to be testable if and only if $K'\beta$ is estimable and K has full column rank. #

EXAMPLE. Consider testing $H_5: \beta \in \text{col}(B)$. To do so, construct a basis $c_1, \ldots, c_s$ for the null(B'), set $K = (c_1, \ldots, c_s)$ and $m = 0$. Then the hypothesis is equivalent to the form $H: K'\beta = m$. #

THEOREM. If $K'\beta$ is estimable, then the $s \times s$ matrix $H = K'(X'X)^g K$ is nonsingular. #

Proof. First recall that K is $p \times s$ and X is $n \times p$. Since $K'\beta$ is estimable, every column of K is in $\text{col}(X')$, and so

$$H = K'(X'X)^g K = K'(X'X)^g X' \underbrace{X(X'X)^g K}_{=: W}.$$

Then

$$\text{rank}(W) \leq \min\{n, s\} \leq s = \text{rank}(K) = \text{rank}(X'X(X'X)^g K) = \text{rank}(X'W) \leq \text{rank}(W).$$

Therefore, $s = \text{rank}(W) = \text{rank}(W'W) = \text{rank}(H)$. #

So assuming $Y \sim N_n(X\beta, \sigma^2 I_n)$ and K is estimable, it follows that

$$\begin{aligned}
\text{Var}(K'\hat{\beta}) &= K'(X'X)^{g}X'\,\text{Var}(Y)\,X(X'X)^{g}K \\
&= \sigma^2 \cdot A'X(X'X)^{g}X'X(X'X)^{g}K, \quad \text{for some matrix } A \text{ since } K \in \text{Col}(X') \\
&= \sigma^2 \cdot K'(X'X)^{g}K \\
&= \sigma^2 H.
\end{aligned}$$

Hence, $\text{Var}(K'\hat{\beta})$ is nonsingular and assuming $H_0: K'\beta = m$,

$$K'\hat{\beta} - m \sim N_s(0, \sigma^2 H).$$

Then by result 5.10, assuming $H_0: K'\beta = m$,

$$(K'\hat{\beta} - m)'(\sigma^2 H)^{-1}(K'\hat{\beta} - m) \sim \chi^2_s.$$

Observe that since σ^2 is unknown we are unable to use this as a test statistic. However, by result 5.16 with $H = LL'$,

$$B = L^{-1}K'(X'X)^{g}X'\frac{1}{\sigma},$$

$V = \sigma^2 I_n$, and $A = I_n - P_X$, because $BVA = 0$, BY is independent of $Y'AY$. Thus,

$$F := \frac{(K'\hat{\beta} - m)'(\sigma^2 H)^{-1}(K'\hat{\beta} - m)/s}{\sigma^{-2}Y'(I - P_X)Y/(n-r)} \sim F_{s,n-r}$$

under $H_0: K'\beta = m$.

The F-test is to reject H_0 if the p-value, $P(F > f)$, is sufficiently small, where f is an observed value of the test statistic.

EXAMPLE. Consider the one-way ANOVA model,

$$Y_{ij} = \mu + \alpha_i + U_{ij},$$

where $U_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ for $j \in \{1, \dots, n_i\}$, $i \in \{1, 2, 3\}$. Then

$$X = \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & 0 \\ 1_{n_2} & 0 & 1_{n_2} & 0 \\ 1_{n_3} & 0 & 0 & 1_{n_3} \end{pmatrix},$$

and denote $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3)'$. A common hypothesis to test is whether the group effects are the same;

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 .$$

Accordingly, $H_0 : K'\beta = m$, where

$$K = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad m = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The least squares estimator $\hat{\beta} = (0, \bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \bar{y}_{3\cdot})'$, and

$$H = K'(X'X)^{g}K = \begin{pmatrix} \frac{1}{n_1} + \frac{1}{n_2} & \frac{1}{n_1} \\ \frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_3} \end{pmatrix}.$$

Then

$$F = \frac{(K'\hat{\beta} - m)'(\sigma^2 H)^{-1}(K'\hat{\beta} - m)/s}{\sigma^{-2} Y'(I - P_X)Y/(n-r)}$$

$$= \frac{\frac{1}{2}\sum_{1}^{3} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2}{\frac{1}{n-3}\sum_{i=1}^{3}\sum_{j=1}^{n_i}(\bar{Y}_{ij} - \bar{Y}_{i\cdot})^2} \sim F_{2,n-3}$$

#

Next consider some useful properties of the constructed F-test. Suppose that two hypotheses are logically equivalent in the sense that

$$\{\beta : K'\beta = m\} = \{\beta : \tilde{K}'\beta = \tilde{m}\}.$$

Is it necessarily the case that the F-test is invariant to the choice of expression?

$\longrightarrow$ To see that this is the case, we will first show that $col(K) = col(\tilde{K})$.

Proof. Recall that

$$\{\beta : K'\beta = m\} = \{K(K'K)^{-1}m + (I - K(K'K)^{-1}K')z : z \in \mathbb{R}^p\}.$$

Since $\{\beta : K'\beta = m\} = \{\beta : \tilde{K}'\beta = \tilde{m}\}$,

$$\tilde{K}'K(K'K)^{-1}m + \tilde{K}'(I - K(K'K)^{-1}K')z = \tilde{m}, \quad \forall z.$$

So

$$\tilde{K}'(I - K(K'K)^{-1}K')z = \tilde{m} - \tilde{K}'K(K'K)^{-1}m, \quad \forall z,$$

and this implies that $P_K\tilde{K} = \tilde{K}$. Hence, $col(\tilde{K}) \subseteq col(K)$. The reverse direction is implied by the symmetry of the argument, and so $col(K) = col(\tilde{K})$. #

Accordingly, there exists a nonsingular matrix Q such that $KQ = \tilde{K}$. Then $Q = (K'K)^{-1}K'\tilde{K}$. From the argument above,

$$\tilde{m} = \tilde{K}'K(K'K)^{-1}m = Q'K'K(K'K)^{-1}m = Q'm.$$

Therefore,

$$(\tilde{K}'\hat{\beta} - \tilde{m})'(\sigma^2\tilde{K}'(X'X)^{g}\tilde{K})^{-1}(\tilde{K}'\hat{\beta} - \tilde{m}) = (Q'K'\hat{\beta} - Q'm)'(\sigma^2Q'K'(X'X)^{g}KQ)^{-1}(Q'K'\hat{\beta} - Q'm)$$

$$= (K'\hat{\beta} - m)'(\sigma^2H)^{-1}(K'\hat{\beta} - m).$$

Since the denominator of the F-statistic does not depend on K nor m, the relation establishes that the F-test is invariant to the choice of expression.

A simple example is to take $\tilde{K} = c \cdot K$ and $\tilde{m} = c \cdot m$ for some constant c.

Page 135 shows the construction of a T-test, the special case when $s = 1$.

LIKELIHOOD RATIO TEST

The LRT is developed in an alternative to the F-test, for testing. Again, begin by assuming $Y \sim N_n(X\beta, \sigma^2 I_n)$ and that the form of the hypothesis to test is $H_0 : K'\beta = m$ for an estimable $K'\beta$. Under this hypothesis, the parameter space is given by

$$\Omega_0 := \{(\beta, \sigma^2) : K'\beta = m, \sigma > 0\},$$

whereas the union of H_0 and H_0^c is given by

$$\Omega := \{(\beta, \sigma^2) : \beta \in \mathbb{R}^p, \sigma > 0\}.$$

Then a likelihood ratio is defined by

$$LR := \frac{\max\limits_{\Omega_0}\{L(\beta, \sigma^2)\}}{\max\limits_{\Omega}\{L(\beta, \sigma^2)\}},$$

where small values of LR suggest evidence against H_0. In particular, we can control the probability of a type 1 error by rejecting H_0 if $LR < C$, where C is a constant such that

$$P(LR < c \mid H_0) = \alpha.$$

Observe that

$$\max_{\Omega}\{L(\beta,\sigma^2)\} = L\left(\hat\beta, \tfrac{1}{n}\|y-X\hat\beta\|^2\right) = \left(\tfrac{2\pi}{n}\|y-X\hat\beta\|^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}},$$

and

$$\max_{\Omega_0}\{L(\beta,\sigma^2)\} = L\left(\hat\beta_H, \tfrac{1}{n}\|y-X\hat\beta_H\|^2\right) = \left(\tfrac{2\pi}{n}\|y-X\hat\beta_H\|^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}},$$

where $\hat\beta_H$ is part of a solution to the RNEs with constraint $K'\beta = m$. Thus,

$$LR = \left(\frac{\|y-X\hat\beta\|^2}{\|y-X\hat\beta_H\|^2}\right)^{\frac{n}{2}} = \left(\frac{Q(\hat\beta)}{Q(\hat\beta_H)}\right)^{\frac{n}{2}},$$

and reject H_0 if,

$$\left(\frac{Q(\hat\beta)}{Q(\hat\beta_H)}\right)^{\frac{n}{2}} < c \qquad \text{or} \qquad \frac{[Q(\hat\beta_H)-Q(\hat\beta)]/s}{Q(\hat\beta)/(n-r)} > \frac{n-r}{s}\left(c^{-\frac{2}{n}}-1\right).$$

We will see shortly that this is actually an F-test.

THEOREM. If $K'\beta$ is estimable and $\hat\beta_H$ is part of a solution to the RNEs with constraint $K'\beta = m$, then

$$\begin{aligned} Q(\hat\beta_H) - Q(\hat\beta) &= (\hat\beta_H-\hat\beta)'X'X(\hat\beta_H-\hat\beta) \\ &= (K'\hat\beta - m)'(K'(X'X)^{g}K)^{-1}(K'\hat\beta - m). \end{aligned}$$ #

Proof. Recall that

$$\begin{aligned} Q(\hat\beta_H) &= \|y - X\hat\beta + X\hat\beta - X\hat\beta_H\|^2 \\ &= \|y-X\hat\beta\|^2 + 2\underbrace{(y-X\hat\beta)'(X\hat\beta - X\hat\beta_H)}_{= (X'y - X'X\hat\beta)'(\hat\beta-\hat\beta_H) = 0} + \|X\hat\beta - X\hat\beta_H\|^2 \end{aligned}$$

So $Q(\hat\beta_H) - Q(\hat\beta) = \|X\hat\beta - X\hat\beta_H\|^2$. Next, observe the RNEs

$$\begin{pmatrix} X'X & K \\ K' & 0 \end{pmatrix}\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ m \end{pmatrix},$$

and let $\hat\theta_H$ be a solution such that $X'X\hat\beta_H + K\hat\theta_H = X'y$. Then

$$X'X(\hat{\beta}-\hat{\beta}_H) = X'y - X'y + K\hat{\theta}_H = K\hat{\theta}_H$$

$$K'(X'X)^{\vartheta}X'X(\hat{\beta}-\hat{\beta}_H) = K'(X'X)^{\vartheta}K\hat{\theta}_H$$

$$(K'(X'X)^{\vartheta}K)^{-1}\underbrace{K'(X'X)^{\vartheta}X'X}_{=K'}(\hat{\beta}-\hat{\beta}_H) = \hat{\theta}_H$$

Finally,

$$\begin{aligned} Q(\hat{\beta}_H) - Q(\hat{\beta}) &= \|X\hat{\beta} - X\hat{\beta}_H\|^2 \\ &= (\hat{\beta}-\hat{\beta}_H)'X'X(\hat{\beta}-\hat{\beta}_H) \\ &= (\hat{\beta}-\hat{\beta}_H)'K\hat{\theta}_H \\ &= (\hat{\beta}-\hat{\beta}_H)'K(K'(X'X)^{\vartheta}K)^{-1}K'(\hat{\beta}-\hat{\beta}_H) \\ &= (K'\hat{\beta}-m)'(K'(X'X)^{\vartheta}K)^{-1}(K'\hat{\beta}-m). \end{aligned}$$ #

COROLLARY. If $K'\beta$ is estimable and $\hat{\beta}$ solves the normal equations, then the $\hat{\beta}_H$ component of a solution to the RNE solves,

$$X'X\beta = X'y - K(K'(X'X)^{\vartheta}K)^{-1}(K'\hat{\beta}-m).$$ #

Proof. From the RNE, $X'X\hat{\beta}_H + K\hat{\theta}_H = X'y$, and from the proof of the theorem,

$$\hat{\theta}_H = (K'(X'X)^{\vartheta}K)^{-1}(K'\hat{\beta}-m).$$

Hence,

$$X'X\beta + K(K'(X'X)^{\vartheta}K)^{-1}(K'\hat{\beta}-m) = X'y.$$ #

THEOREM. If $P'\beta$ is a system of linearly independent, non-estimable functions and $\hat{\beta}_H$ is part of a solution to the RNEs with constraint $P'\beta = \delta$, then $Q(\hat{\beta}_H) = Q(\hat{\beta})$ and $\hat{\theta}_H = 0$. #

Proof. From the RNE, $X'X\hat{\beta}_H + P\hat{\theta}_H = X'y$, and so $P\hat{\theta}_H = X'(y - X\hat{\beta}_H) \in col(X')$. Since $P'\beta$ is non-estimable, $col(P) \cap col(X') = \{0\}$ which implies that $P\hat{\theta}_H = 0$. P has full column rank, so it must be the case that $\hat{\theta}_H = 0$. Therefore,

$$\hat{\beta}_H \in \{X'X\beta = X'y\}$$

so that $Q(\hat{\beta}_H) = Q(\hat{\beta})$. #

CONFIDENCE INTERVALS AND MULTIPLE COMPARISONS

Continue with the assumption that $Y \sim N_n(X\beta, \sigma^2 I_n)$, and recall that if $\lambda'\beta$ is estimable then,

$$\lambda'\hat{\beta} \sim N\left(\lambda'\beta, \sigma^2 \lambda'(X'X)^{g}\lambda\right),$$

for $\hat{\beta} \in \{X'X\beta = X'y\}$. If we estimate σ^2 with $\hat{\sigma}^2 = \frac{1}{n-r} y'(I-P_X)y$, then

$$\frac{1}{\hat{\sigma}^2}\left(\frac{\lambda'\hat{\beta} - \lambda'\beta}{\lambda'(X'X)^{g}\lambda}\right) \sim T_{n-r}.$$

Accordingly,

THEOREM. Let $\{E_j\}$ be a collection of measurable events. Then

(i) $P\left(\bigcup_j E_j\right) \leq \sum_j P(E_j)$

(ii) $P\left(\bigcap_j E_j\right) \geq 1 - \sum_j P(E_j^c)$ #

Proof.