Bios 762 Notes

1 Introduction

This is a set of abridged notes for BIOS 762 at the University of North Carolina compiled by Ethan Alt. These notes serve to help study for the PhD Qualifying Exam. The goal is to present key ideas and theorems from the course in note form rather than slide form. If there are any errors, please contact Ethan Alt at ethanalt@live.unc.edu.

2 Vector Spaces

Definition: A real vector space $\mathcal{M}$ is a set of elements (vectors) satisfying the addition and scalar multiplication axioms. For details, consult the notes or Wikipedia.

2.a Linear Independence

A set of vectors $D = \{x_1, \dots x_r\}$ is called linearly dependent if there is a set of scalars $\alpha_1, \dots \alpha_r$ not all zero such that

$$\sum_{i=1}^{r} \alpha_i x_i = 0$$

If $\sum_{i=1}^{r} \alpha_i x_i = 0 \implies \alpha_i = 0$ for every $i = 1, \dots, r$, then $D = \{x_1, \dots, x_r\}$ is linearly independent.

Note: if $0 \in D$, D is linearly dependent. If $\emptyset \in D$, D is linearly independent. If D is linearly independent, then $D_1 \subset D$ is linearly independent (proof by contradiction).

Below, we present an important theorem about a set of linearly dependent vectors. The proof is left as an exercise

Theorem: A set of vectors is linearly dependent if and only if some vector of the set can be written as a lienar combination of the others. That is, there exists a k such that

$$x_k = \sum_{i=1}^{r} -\frac{\alpha_i}{\alpha_k} x_i$$
$$\alpha_k \neq 0, i \neq k$$

2.b Basis

Definition: A basis in a vector space $\mathcal{M}$ is a set of linearly independent vectors such that every $x \in \mathcal{M}$ is a linear combination of vectors in the set.

Example: $\mathcal{M} = \mathbb{R}^n$. The set of the n unit vectors forms a basis.

Below, we list some useful relationships between basis and span.

1. The set of all possible linear combinations of elements of D is a vecotr space, called the span of D, written $\mathcal{S}(D)$
2. $\mathcal{S}(D) = \cap_i \mathcal{M}_i$ of all vector spaces containing D.
3. A basis for $\mathcal{M}$ is a linearly independent set of elements of $\mathcal{M}$ whose span is $\mathcal{M}$

This leads to an important theorem. **Theorem:** Every basis of a vector space $\mathcal{M}$ contains the same number of elements. This number is called the dimension of $\mathcal{M}$, written $\dim(M)$. This number is also called the rank of $\mathcal{M}$, written $r(\mathcal{M})$.

The next theorem is critical for linear models. **Theorem:** If $\{x_1, \ldots, x_r\}$ is a linearly independent set of vectors in $\mathcal{M}$ and $\dim(\mathcal{M}) = n$, $r < n$, then there exists some elements $x_{r+1}, \ldots, x_n$ such that $\{x_1, \ldots, x_n\}$ is a basis for $\mathcal{M}$.

This theorem says that any set of linearly independent set of vectors can be extended to a basis.

2.c Linear Subspaces

Definition: Let $\mathcal{M}$ be a vector space and let N be any set with $N \subset \mathcal{M}$. Then N is a subspace of $\mathcal{M}$ if and only if N is a vector space.

Theorem: Let $\mathcal{M}$ be a vector space and let N be a nonempty subset of $\mathcal{M}$. If N is closed under addition and scalar multiplication, then N is a subspace of $\mathcal{M}$.

Let H and K be two linear subspaces. Define the sum of H and K as

$$H + K = \{x + y : x \in H, y \in K\}$$

and define the intersection of H and K as

$$H \cap K = \{x : x \in H, x \in K\}$$

Theorem: Both $H+K$ and $H \cap K$ are subspaces. The proof is left as an exercise. Essentially, you only need to show that they are closed under addition and scalar multiplication.

2.c.1 Disjoint Subspaces

Definition: Two subspaces are disjoint if $H \cap K = 0$, where 0 is the null vector.

The concept of disjointness leads to a very important theorem in linear models.

Theorem: If $H \cap K = 0$ and $z \in H + K$, then the decomposition $z = x + y$ with $x \in H$ and $y \in K$ is unique. The proof is in the notes on page 42.

Theorem: If $H \cap K = 0$, then $r(H + K) = r(H) + r(K)$. In general, we have

$$r(H + K) = r(H) + r(K) - r(H \cap K)$$

Definition: If N and N^C are disjoint subspaces of $\mathcal{M}$ and $\mathcal{M} = N + N^C$, then N^C is called a complement of N.

Remark: The complement is not unique.

Definition: Suppose $\mathcal{M}$ is a vector space in R^n. Let x and y be two vectors in $\mathcal{M}$. Then x and y are said to be orthogonal, written $x \perp y$, if $x'y = 0$ where x' denotes the transpose of x.

This leads to an important definition in linear models.

Definition: Suppose N is a subspace of R^n. Then $\{x_1, \ldots, x_r\}$ is an orthogonal basis for N if for every $i \neq j$, $x_i'x_j = 0$. $\{x_1, \ldots, x_r\}$ is an orthonormal basis if, in addition, $x_i'x_i = 1$ for all i.

The concept of an orthonormal basis leads to one of the most important theorems in linear models, which will be used extensively throughout the course.

Theorem: (Gram-Schmidt)
Let N be a subspace of R^n with basis $\{x_1, \ldots, x_r\}$. Then there exists an orthonormal basis for N, $\{y_1, \ldots, y_r\}$ with $y_s \in \mathcal{S}(x_1, \ldots x_s), s = 1, \ldots r$.

Explicitly, the y_s's are given by:

$$y_1 = (x_1'x_1)^{-1/2}x_1$$

$$w_s = x_s - \sum_{i=1}^{s-1}(x_s'y_i)y_i, s = 2, \ldots, r$$

$$y_s = (w_s'w_s)^{-1/2}w_s, s = 2, \ldots, r$$

The proof of the theorem simply involves showing the constructed y_s's are orthonormal and by showing that they span R^n. Note that since the vectors are orthogonal, they are automatically linearly independent.

Another important definition and theorem are both crucial to the theory of linear models.

Definition: (Orthogonal Complement)
Let N be a subspace of a vector space $M \subset R^n$. Define

$$N^{\perp} := \{y \in \mathcal{M} : y \perp N\}$$

$N^{\perp}$ is called the orthogonal complement of N with respect to $\mathcal{M}$. If $\mathcal{M} = R^n$, then $N^{\perp}$ is referred to as the orthogonal complement of N.

Theorem: Let $\mathcal{M}$ be a vector space and let $N^{\perp}$ be the orthogonal complement of N with respect to $\mathcal{M}$. Then $N^{\perp}$ is a subspace of $\mathcal{M}$, and if $x \in \mathcal{M}$, x can be written uniquely as $x = x_0 + x_1$, with $x_0 \in N and x_1 \in N^{\perp}$. The ranks of these subspaces satisfy

$$r(\mathcal{M}) = r(N) + r(N^{\perp}$$

Also

$$\mathcal{M} = N + N^{\perp} = \{x : x = x_0 + x_1, x_0 \in N, x_1 \in N^{\perp}\}$$

This theorem will be used extensively throughout linear models, in particular for dealing with tests and independence of quadratic forms.

2.d Matrices

Suppose A is an $n \times p$ matrix. A' will denote the transpose of A. A matrix can be defined as a linear transformation on a vector space.

Definition: Suppose $\mathcal{M}$ is an arbitrary vector space. A linear transformation A on a vector space $\mathcal{M}$ is a function mapping $\mathcal{M} \to \mathcal{M}$ such that

$$A(\alpha x + \beta y) = \alpha A(x) + \beta A(y)$$

for all scalars α, β and all $x, y \in \mathcal{M}$.

Definition: (Column Space)
Let A be an $n \times p$ matrix and write A in terms of its columns

$$A = (x_1, \ldots, x_p)$$

where each $x_i \in R^n$, $i = 1, \ldots, p$. The space spanned by the columns of A is called the column space of A, written $C(A)$. That is, $\mathcal{S}(A) = C(A)$. Also, $r(A)$ will denote the rank of A.

Definition: (Trace)
Let A be an $n \times n$ square matrix with ij-th element a_{ij}. The trace of A is defined as

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

Properties of Trace:

1. tr(A + B) = tr(A) + tr(B)
2. tr(ABC) = tr(BCA) = tr(CAB)

Note for the second property, we can only pick off the end (cannot pick off the middle).

2.e Inverses of Matrices and Singularity

Definition (Matrix Inverse)
Suppose A is an $n \times n$ square matrix. Then A is said to be nonsingular if there exists a matrix A^{-1} such that $A^{-1}A = AA^{-1} = I$. If no such matrix exists, then A is said to be singular.

If B is nonsingular, then $tr(A) = tr(BAB^{-1}) = tr(B^{-1}AB)$.

Theorem
An $n \times n$ matrix A is nonsingular if an donly if $r(A) = n$, i.e., $C(A)$ is a basis for R^n. Thus, A is nonsingular if and only if all its columns are linearly independent.

NB: If A is singular, then there exists a nonzero vector x such that $Ax = 0, x \in R^n$.

Definition (Null Space)
The set of all vectors x such that $Ax = 0$ is a vector space, and it is called the null space of A, written $\mathcal{N}(A)$.

Theorem:
Suppose A is $n \times n$. If $r(A) = r$, then $r(\mathcal{N}(A)) = n - r$.

2.f Eigenvalues and Eigenvectors

Definition: (Eigenvalue and Eigenvector) Suppose A is an $n \times n$ square matrix. An eigenvector of A is any nonzero vector x satisfying

$$Ax = \lambda x, \quad \lambda \in R^1$$

λ is called an eigenvalue of A.

Theorem
If A is a symmetric matrix, then th eigenvalues of A are real.

Theorem
λ is an eigenvalue of A if and only if $A - \lambda I$ is singular.

Below, we list some important facts about eigenvectors and eigenvalues.

1. If λ is an eigenvalue of A and $\lambda \neq 0$, then the eigenvectors corresponding to λ form a subspace of $C(A)$.
2. If A is symmetric and λ and γ are distinct eigenvalues, then the eigenvectors corresponding to λ and γ are orthogonal. These eigenvectors form a basis for a subspace of $C(A)$.
3. A is nonsingular if and only if all of its eigenvalues are nonzero.
4. Theorem:
   If A is a symmetric matrix, then there exists a basis for $C(A)$ consisting of eigenvectors of nonzero eigenvalues.
5. if A is $n \times n$ and symmetric, and some if its eigenvalues are 0, then the eigenvectors corresponding to the nonzero eigenvalues are a basis for $C(A) \subset R^n$. Thus, $r(A) =$ number of nonzero eigenvalues of A.
6. If A is $n \times n$ and symmetric, then the eigenvectors corresponding to the 0 eigenvalues (if any) are a basis for $\mathcal{N}(A)$.
7. **Important:** If A is symmetric, then $\mathcal{N}(A) = C(A)^{\perp}$. That is, the null space of A corresponds to the orthogonal complement of A.
8. **Theorem:**
   Suppose A is $n \times n$ and symmetric. Then there exists eigenvectors of A that are an orthogonal basis for $C(A)$. If A is nonsingular, then they are an orthogonal basis for R^n. If we normalize these eigenvectors, they are an orthonormal basis.

Theorem:
Suppose A is $n \times n$ symmetric of rank $r \leq n$. Then

1. $\mathcal{N}(A) = C(A)^{\perp}$
2. $C(A) \cap \mathcal{N}(A) = 0$
3. $C(A) + \mathcal{N}(A) = R^n$
4. $r(A) = r$, $r(\mathcal{N}(A)) = n - r$

The eigenvalues of a matrix A are found by finding the zeroes of the equation $\det(A - \lambda I) = 0$.

Suppose A is $n \times n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then

1. $\det(A) = \prod_{i=1}^{n} \lambda_i$
2. If A is singular, then $\det(A) = 0$.
3. If A is nonsingular, then A^{-1} exists and the eigenvalues are given by $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$.
4. The eigenvalues of A' are the same as those of A.

5. $tr(A) = \sum_{i=1}^n \lambda_i$ and $tr(A^{-1}) = \sum_{i=1}^n \lambda_i^{-1}$.
6. If A is symmetric, then $tr(A^r) = \sum_{i=1}^n \lambda_i^r$ for any integer r.

2.g Orthogonal Matrices

Definition: (Orthogonal Matrices)
A square matrix P is said to be orthogonal if $P' = P^{-1}$. If P is an orthogonal matrix, then so is P'. Thus, a square matrix is orthogonal if $PP' = P'P = I$.

Theorem:
The product of two orthogonal matrices is orthogonal. The proof is on page 56 in the notes, but is very easy.

Theorem:
An $n \times n$ matrix P is orthogonal if and only if the columns of P form an orthonormal basis for R^n.

Definition:
Suppose A is an $n \times p$ matrix. Then

$$C(A) = \{z : Ax = z, x \in R^p\}$$

Below is a very important theorem.

Theorem:
Suppose A is an $n \times p$ matrix. Then

$$\mathcal{N}(A) = C(A')^{\perp}$$

Proof is fundamental, and is on page 57 in the notes.

Miscellaneous Results

1. Suppose A is an $n \times p$ matrix of rank r. Then $r \leq \min(n, p)$.
2. Suppose A is an $n \times p$ matrix of rank r and B is a $p \times s$ matrix with $r(B) \geq r$. Then $C(A) = C(AB)$, and thus $r(AB) = r(A) = r$. If $r(B) < r$, then $C(AB) \subset C(A)$, and therefore, $r(AB) \leq r(A)$.
3. If B is a square nonsingular matrix, then $C(A) = C(AB)$.
4. In general, $C(AB) \subset C(A)$ and $\mathcal{N}(B) \subset \mathcal{N}(AB)$.

2.h Matrix Decompositions

2.h.1 Spectral Theorem

Theorem (Spectral Theorem)
Suppose A is an $n \times n$ symmetric matrix. Then there exists an orthogonal matrix P such

that

$$A = P\Lambda P'$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n$ is an $n \times n$ matrix of the eigenvalues of A with $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Here, P is an orthogonal matrix of eigenvectors corresponding to the eigenvalues of A.

Theorem:
The rank of a symmetric matrix is the number of nonzero eigenvalues.

Theorem:
The eigenvalues of a positive definite matrix are all positive, and the eigenvalues of a positive semidefinite matrix are all nonnegative.

Proof:
Let λ be an eigenvalue of A and let x be the corresponding eigenvector. Then $Ax = \lambda x$ and so since A is positive definite, $0 < x'Ax = \lambda x'x$ Note that $x'x > 0$, so we must have $\lambda > 0$. Thus, the eigenvalues of A are positive.

Theorem:
A is positive semidefinite if and only if there exists a matrix Q such that $r(A) = r(Q)$ and $A = QQ'$.

Theorem:
A is positive definite if and only if there exists a nonsingular matrix Q such that $A = QQ'$

Proof:
Suppose A is positive semidefinite. By the spectral theorem, we can write $A = P\Lambda P'$, where Λ is a diagonal matrix of eigenvectors and P is orthogonal. Let $Q = P\Lambda^{1/2}$. Then $A = QQ'$. Clearly, Q is nonsingular since $\Lambda^{1/2}$ is nonsingular because all eigenvalues are nonzero and P is nonsingular by definition or orthogonal matrix, and the product of two nonsingular matrices is nonsingular.

2.h.2 Singular Value Decomposition

Theorem: (Singular Value Decomposition) Suppose A is an $n \times p$ matrix of rank r, ($r \leq \min(n, p)$). There exists orthogonal matrices $U_{p \times p}$ and $V_{n \times n}$ such that

$$V'AU = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix}$$

where $\Delta = \text{diag}(\delta_1, \ldots, \delta_r)$ is an $r \times r$ diagonal matrix with $\delta_1 \geq \ldots \geq \delta_r$. The δ_i's are called the singular values of A.

Below are several important properties of the SVD.

1. $$A = VDU, \text{ where } D = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix}$$

2. Split $V = (V_1, V_2)$ and $U = (U_1, U_2)$, where V_1 is $n \times r$ and V_2 is $n \times (n-r)$, U_1 is $p \times r$ and U_2 is $p \times (p-r)$. Then

$$\begin{aligned} A = VDU' &= (V_1, V_2) \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_1' \\ U_2' \end{pmatrix} \\ &= V_1 \Delta U_1' + V_2 0 U_2' \\ &= V_1 \Delta U_1' \end{aligned}$$

   This implies that

$$\begin{aligned} C(A) &= \{z : z = Ax = V_1 \Delta U_1' x, x \in R^p\} \\ &= \{z : z = V_1 x^*, x^* \in R^r\} \\ &= C(V_1) \end{aligned}$$

   Note that since Δ and U_1' have rank r, we can set $x^* = \Delta U_1' x$ and get the equality above. Thus, the columns of V_1 span the same space as the columns of A. Since V is an orthonormal basis for R^n, the columns of V_1 are an orthonormal basis for $C(A)$.

3. Similar arguments show that

$$C(A') = C(U_1)$$

4. Claim: $\mathcal{N}(A) = C(U_2) = C(A')^{\perp}$. See proof on page 68 in notes.

5. We have the following summary:

Matrix	Column Space	Null Space
A	$C(V_1)$	$C(U_2)$
A'	$C(U_1)$	$C(V_2)$

6. The columns of V_1 are the eigenvectors corresponding to the nonzero eigenvalues of AA' and the columns of U_1 are the eigenvectors corresponding to the nonzero eigenvalues of $A'A$.

7. If $r(A) = r$, then $\delta_1^2, \ldots, \delta_r^2$ are the eigenvalues of $A'A$.

2.h.3 Q-R Factorization

Theorem: (Q-R Factorization) Suppose A is $n \times p$ with linearly independent columns. Then A can be written uniquely in the form

$$A = QR$$

where $Q_{n \times p}$ has orthonormal columns and $R_{p \times p}$ is an upper triangular matrix with positive diagonal elements. The proof is given on page 69 in the notes.

2.i Projections

Definition (Projection)
Suppose $\mathcal{M}$ is a vector space and N_1 and N_2 are two subspaces in $\mathcal{M}$ where $N_1 + N_2 = \mathcal{M}$ and $N_1 \cap N_2 = 0$. Consider the unique decomposition $z = x + y$ where $x \in N_1$ and $y \in N_2$. The linear transformation

$$P_{N_1|N_2} z = x$$

is called the projection of z onto the subspace N_1 along the subspace N_2.

Theorem: The projection operator onto N_2 along N_1 is given by

$$P_{N_2|N_1} = I - P_{N_1|N_2}$$

Proof:
We want to show that $P_{N_2|N_1} z = y$ We have

$$\begin{aligned} P_{N_2|N_1} z &= (I - P_{N_1|N_2}) z \\ &= z - P_{N_1|N_2} z \\ &= (x + y) - x \\ &= y \end{aligned}$$

By definition of projection, $I - P_{N_1|N_2}$ is the projection operator onto N_2 along N_1.

Remarks:

1. If the projection is at a ”right angle”, then it is called an **orthogonal** projection and such projections are unique.
2. If the projection is not at a right angle, then it is called a projection and it is not unique.

This leads us to an important definition **Definition:** (Projection Operator)
Let A be an $n \times n$ matrix. A is said to be a projection operator onto $C(A)$ along $\mathcal{N}(A)$ if for any $v \in C(A)$,

$$Av = v$$

Definition: (Idempotent Matrix)
If $A^2 = A$, then A is said to be an idempotent matrix.

Theorem:
$A^2 = A$ if and only if A is a projection matrix.

Proof: ($\implies$)
Suppose A is idempotent. Let $v \in C(A)$. Then $v = Ab$ for some $b \in R^n$. Thus,

$$Av = A(Ab) = (AA)b = A^2 b = Ab = v$$

By definition of a projection, we have that A is a projection matrix.

$(\Longleftarrow)$
Let $x \in R^n$. We can express x as $x = x_0 + x_1$, where $x_0 \in C(A)$ and $x_1 \in \mathcal{N}(A)$. We have

$$A^2x = A[A(x_0 + x_1)] = Ax_0 = Ax$$

The first equality follows by the decomposition of x and the associative property of matrix multiplication. The second equality follows because A is a projection by assumption. The last equality follows because $Ax = A(x_0 + x_1) = Ax_0$ by definition of projection. Since $A^2x = Ax$ for every vector x, we have $A = A^2$. Hence, A is idempotent.

The next definition and theorems are critical for the theory of linear models, more than anything else discussed thus far. **Definition:** (Othogonal Projection Operator)
M is an orthogonal projection operator onto $C(X)$ if and only if:

1. $v \in C(X) \implies Mv = v$ (projection)
2. $w \in C(X)^{\perp} \implies Mw = 0$ (orthogonal)

Theorem:
If M is an orthogonal projection operator onto C(X), then C(M) = C(X). *Proof:*
We must show $C(M) \subset C(X)$ and $C(M) \supset C(X)$.
$(\supset)$
Let $v \in C(X)$. Note that $Mv \in C(M)$. Since M is a projection onto $C(X)$, we have $Mv = v$. Thus, $v \in C(M)$. Since $v \in C(X)$ was arbitrary, we have $C(X) \subset C(M)$.

$(\subset)$
Let $v \in C(M)$. Then $v = Mb$ for some $b \in R^n$. We can express $b = b_0 + b_1$, where $b_0 \in C(X)$ and $b_1 \in C(X)^{\perp}$. We have

$$v = Mb = M(b_0 + b_1) = Mb_0 + Mb_1$$

Now, $Mb_1 = 0$ and $Mb_0 = b_0$ by definition of orthogonal projection, so we have $v \in C(M)$ and $v = Mb_0 = b_0 \in C(X)$. Thus, $C(M) \subset C(X)$.

Theorem:
M is an orthogonal projection operator onto C(M) if and only if M is symmetric and idempotent.

Proof:
$(\implies)$
Suppose M is an orthogonal projection operator. Let $v \in R^n$ and write $v = v_0 + v_1$, where $v_0 \in C(M)$ and $v_1 \in C(M)^{\perp}$. We have

$$M^2v = M(Mv) = M[M(v_0+v_1)] = M(Mv_0+Mv_1) = M(v_0) = Mv_0+Mv_1 = M(v_0+v_1) = Mv$$

Since $M^2v = Mv$ for every $v \in R^n$, we have $M^2 = M$. Hence, M is idempotent.

It remains to be shown that $M = M'$. Let $w = w_1+w_2$, where $w_1 \in C(M)$ and $w_2 \in C(M)^\perp$. Also, write $v = v_1+v_2$ wehre $v_1 \in C(M)$ and $v_2 \in C(M)^\perp$. Note that $(I-M$ is the orthogonal projection operator onto $C(M)^\perp$. Thus,

$$(I - M)v = v_2$$

Thus, for every v and w

$$w'M'(I - M)v = w_1'M'(I - M)v_2 = w_1'v_2 = 0.$$

Hence, $M'(I-M) = 0 \implies M' = M'M$. Since $M'M$ is symmetric, we have M' is symmetric and hence $M = M'$.

Hence, orthogonal projection operators are idempotent and symmetric.

$(\Longleftarrow)$
Suppose M is idempotent and symmetric. We must show $Mv = v$ for every $v \in C(M)$ and $Mw = 0$ for every $w \in C(M)^\perp$.

Let $v \in C(M)$. Then $v = Mb$ for some $b \in R^n$. We have

$$Mv = M(Mb) = (MM)v = M^2v = Mv$$

Let $w \in C(M)^\perp$. Recall that the null space of M is the orthogonal complement to column space of M'. Using this with the fact that M is symmetric, we have that

$$\mathcal{N}(M) = [C(M')]^\perp = [C(M)]^\perp$$

And hence $w \in \mathcal{N}(M) = \{x \in R^n : Mx = 0\}$. By definition, $Mw = 0$. Thus, M is an orthogonal projection operator onto $C(M)$.

Theorem: (Uniqueness of Orthogonal Projection Operators)
Let M be an <u>orthogonal</u> projection operator. Then M is unique.

Proof:
Suppose M and P are two orthogonal projection operators onto $C(M)$. Let $x \in R^n$. We can express $x = x_0 + x_1$, where $x_0 \in C(M)$ and $x_1 \in C(M)^\perp$. We have

$$\begin{aligned} Mx &= M(x_0 + x_1) = Mx_0 + Mx_1 = x_0 + 0 = x_0 \\ Px &= P(x_0 + x_1) = Px_0 + Px_1 = x_0 = 0 = x_0 \end{aligned}$$

Hence, for every $x \in R^n$, we have $0 = Mx - Px = (M - P)x$. Thus, $M = P$.

Theorem: (Properties of Orthogonal Projections)
Suppose M is an orthogonal projection operator. Then

1. The eigenvalues of M are 0 or 1
2. $r(M) = tr(M) = r$
3. M is positive semidefinite

Proof
(i). Let λ be an eigenvalue of M. By definition, there exists $v \neq 0$ such that $Mv = \lambda v$. We have

$$\lambda v = Mv = M^2 v = M(Mv) = M(\lambda v) = \lambda(Mv) = \lambda(\lambda v) = \lambda^2 v$$

Hence,

$$\begin{aligned} &\lambda^2 v = \lambda v \\ \iff &(\lambda^2 - \lambda)v = 0 \\ \iff &\lambda(\lambda - 1)v = 0 \end{aligned}$$

Since v is an eigenvector, $v \neq 0$ and so $\lambda = 0$ or $\lambda = 1$.

(ii.) Note that M is symmetric. By the spectral theorem, we can write

$$M = PDP'$$

where D is a diagonal matrix consisting of the eigenvalues of M and P is orthogonal. Thus,

$$\text{tr}(M) = \text{tr}(PDP') = \text{tr}(P'PD) = \text{tr}(ID) = \text{tr}(D)$$

By (i.), the eigenvalues of D are 0 or 1 and the number of nonzero eigenvalues is the rank of a matrix, so $\text{tr}(D) = r(D)$. Since $r(M) = r$, we must have $r(D) = r$ and, hence $tr(D) = r$, thus giving $r(M) = tr(M) = r$.

Note: A more direct proof can be done by recalling that the trace of a matrix is the sum of its eigenvalues, but the above proof is a nice invocation of the spectral theorem.

(iii.) This is immediate from the fact that the eigenvalues of M are 0 or 1, and hence are nonnegative. (It is also immediate because A is symmetric and so has nonzero eigenvalues).

Theorem: (Construction of Orthogonal Projection Operator (1))
Let X be an $n \times p$ matrix of rank $r \leq \min(n, p)$. Suppose $\{a_1, \ldots, a_r\}$ is an orthonormal basis for $C(X)$ and let $A = (a_1, \ldots a_r)$. Then

$$AA' = \sum_{i=1}^{r} a_i a_i'$$

is the unique orthogonal projection operator onto $C(X)$.

Note: Showing that AA' is symmmetric and idempotent is NOT enough. This means that AA' is an orthogonal projection operator, but does not show what the projection is onto. You need also to show that $C(AA') = C(X)$ by showing each set is a subset of the other.

Proof:
First, note that $(AA')' = (A')'A' = AA'$. Thus, AA' is symmetric.

Now, note that the ij-th element of $A'A$ simply $a_i'a_j$. Since $\{a_1, \ldots, a_r\}$ form an orthonormal basis, we must have

$$\begin{aligned}(A'A)_{ii} &= a_i'a_i = 1 \\ (A'A)_{ij} &= a_i'a_j = 0 \text{ for } i \neq j\end{aligned}$$

Hence, $A'A = I$ and

$$(AA')^2 = AA'AA' = A(A'A)A' = A(I)A' = AA'$$

Thus, AA' is symmetric and idempotent, and hence is the orthogonal projection operator (onto some space).

Now, we show $C(AA') = C(X)$.

$(C(AA') \subset C(X))$
Let $x \in C(AA')$. Then $x = AA'b$ for some b by definition. Thus, $x = Ab^*$ for some $b*$, so that $x \in C(A)$. Since the columns of A form an orthonormal basis for $C(X)$, we must have $x \in C(X)$. Hence, $C(AA') \subset C(X)$.

$(C(AA') \supset C(X))$
Let $x \in C(X)$, then $x \in C(A)$ since the columns of A form an orthonormal basis for $C(X)$. Hence, $x = At$ for some t. Since A has full rank, any vector $t \in R^r$ can be written as $t = A'z$, where $z \in R^n$. Thus, $x = At = AA'z$ and so $x \in C(AA')$. Thus, $C(X) = C(AA')$.

A more direct proof:
Note that $C(X) = C(A)$ since the columns of A form an orthonormal basis for $C(X)$. Now, note that A is an $n \times r$ matrix of rank r, and that A' is an $r \times n$ matrix of rank r. By the miscellaneous results for column spaces (page 58, #2), we have $C(A) = C(AA')$. Thus, $C(X) = C(A) = C(AA')$, giving $C(X) = C(AA')$.

Theorem: (special case)
Suppose X is an $n \times p$ matrix of rank p. Define $M = X(X'X)^{-1}X'$. Then M is the orthogonal projection operator onto $C(X)$.

Proof: Note that

$$M^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = XI(X'X)^{-1}X' = X(X'X)^{-1}X' = M$$

and

$$M' = [X(X'X)^{-1}X']' = X(X'X)^{-1}X' = M$$

so M is idempotent and symmetric, and hence M is an orthogonal projection operator.

We now show that it is an orthogonal projection operator onto $C(X)$. There are two methods of doing this:

Method 1: (Show column spaces are equal)
Note that $C(X) = \{Xb : b \in R^n\}$.

$$C(M) = \{Mv : v \in R^n\} = \{X(X'X)^{-1}X'v : v \in R^n\} = \{Xv^* : v^* \in R^n\} = C(X).$$

Method 2: (Show the projections)
Let $v \in C(X)$. Then $v = Xb$ for some b. We have

$$Mv = X(X'X)^{-1}X'Xb = Xb = v$$

Let $w \in C(X)^{\perp} = \mathcal{N}(X')$. Then $X'w = 0$ and we have

$$Mw = X(X'X)^{-1}X'w = X(X'X)^{-1}0 = 0$$

Some basic theorems will be presented below without proof.

Theorem:
Let $M = X(X'X)^{-1}X'$. Then M can be written as $M = QQ' = V_1V_1'$ where Q and V_1 are defined as in the Q-R and SVD decompositions, respectively.

Theorem:
$I - M$ is the unique orthogonal projection operator onto $C(X)^{\perp}$.

Theorem:
Suppose M is an orthogonal projection operator. Let m_{ii} denote the i-th diagonal element of M. Then $0 \leq m_{ii} \leq 1$. Suppose $M = X(X'X)^{-1}X'$ and $1 \in C(X)$. Then $\frac{1}{n} \leq m_{ii} \leq 1$. If the number of rows of X exactly equal to x_i is c, then $\frac{1}{n} \leq m_{ii} \leq \frac{1}{c}$.

Proof:
Since $\text{tr}(M) = \sum_{i=1}^{n} \lambda_i$ where λ_i are the eigenvalues of M, and since M is a projection so its eigenvalues are exactly 0 or exactly 1, we have $0 \leq \sum_{i=1}^{n} m_{ii} \leq n$. Hence, $0 \leq m_{ii} \leq 1$.

The other proofs are currently unknown to the author.

Theorem:
Suppose X is an $n \times p$ matrix of rank p. Write $X = (X_1, X_2)$, where X_1 is $n \times k$ and X_2 is $n \times (p-k)$ and $r(X_1) = k$ and $r(X_2) = p - k$. Let $M_j = X_j(X_j'X_j)^{-1}X_j, j = 1, 2$ and let $M = X(X'X)^{-1}X'$. Further, let

$$\begin{aligned} X_j^* &= (I - M_{3-j}X_j, j = 1, 2 \\ M_j^* &= X_j^*(X_j^{*\prime}X_j^*)^{-1}X_j^* \end{aligned}$$

Then X_1^* consists of columns which are orthogonal to X_2 and X_2^* has columns which are orthogonal to X_1. Therefore, $M = M_1 + M_2^*$ and $M = M_2 + M_1^*$

Proof:
We will only show one of the two cases. Let $X_2^* = (I - M_1)X_2$. Let x_{2j}^* denote the j-th column of X_2^* and let x_{1i} denote the i-th column of X_1. Let e_k denote the k-th unit vector. Then

$$\begin{aligned} x_{1i}'x_{2j}^* &= (X_1e_i)'(X_2^*e_j) \\ &= e_i'X_1'(I - M_1)X_2e_j \\ &= e_i'X_1'X_2e_j - e_i'X_1'M_1X_2e_j \\ &= e_i'X_1'X_2e_j - e_i'X_1'X_2e_j \\ &= 0 \end{aligned}$$

The fourth equality follows because since M_1 is the orthogonal projection operator onto $C(X_1)$, we must have $M_1X_1 = X_1$, and $(M_1X_1)' = X_1'M_1' = X_1M_1$ so $X_1'M_1 = X_1'$.

Now, since X_2 is full rank, we have $r(X_2^*) = r((I - M_1)X_2) = r(I - M_1) = p - k$ and $r(M_1) = k$ and $X_1 and X_2$ have orthogonal columns so we must have $M = M_1 + M_2^*$.

2.i.1 Generalized Inverses

Definition: (Generalized Inverse)
Consider the linear transformation $A : R^p \to R^n$. A generalized inverse of A is a linear transformation A^- such that

$$AA^-y = y \text{ for all } y \in C(A)$$

equivalently, if A is an $n \times p$ matrix, then $A^-_{p\times n}$ is a generalized inverse of A if

$$AA^-A = A$$

Note that $(A^-A)(A^-A) = A^-(AA^-A) = A^-A$ so A^-A is a projection.

Generalized inverses are not unique in general, but a particular type of generalized inverse is unique.

Definition (Moore-Penrose)
Suppose A is an $n \times p$ matrix. The Moore-Penrose generalized invese of A is a matrix A^+ such that

1. $(AA^+)' = AA^+$ (AA^+ is symmetric)
2. $(A^+A)' = A^+A$ (A^+A is symmetric)
3. $AA^+A = A$ (A^+ is a generalized inverse of A)
4. $A^+AA^+ = A^+$ (A is a generalized inverse of A^+).

Thus, AA^+ and A^+A are orthogonal projection operators (they are symmetric and idempotent).

Theorem
Every matrix A has a Moore-Penrose generalized inverse

Proof:
Suppose A has rank r. From the SVD of A, we can write $A = V_1\Delta U_1'$, where Δ is an $r \times r$ matrix with positive elements, and V_1 and U_1 have orthonormal columns. Let $A^+ = U_1\Delta^{-1}V_1'$.

It is easy to simply verify the four properties above.

Theorem
The Moore-Penrose generalized inverse is unique.

Proof
Suppose A^- is another matrix satisfying (1)-(4) above. Then we have

$$\begin{aligned}
AA^+ &= (AA^-AA^+) \because \text{ def. of g-inv} \\
&= (AA^-)'(AA^+)' \because \text{ (1)} \\
&= (A^-)'A'(A^+)'A' \\
&= (A^-)'(AA^+A)' \\
&= (A^-)'A' \because \text{ def. of g-inv} \\
&= (AA^-)' \\
&= AA^- \because \text{ (1)}
\end{aligned}$$

Theorem:
$M = XX^+$ is the unique orthogonal projection operator onto $C(X)$.

Proof:
We have already shown that M is symmetric and idempotent. We need to show $C(M) = C(X)$.

$(\implies)$
$C(M) = C(XX^+) \subset C(X)$

$(\impliedby)$
Note that $r(M) = \text{tr}(M) = \text{tr}(XX+) = tr(V_1\Delta U_1'U_1\Delta^{-1}V_1') = \text{tr}(V_1V_1') = tr(I_{r\times r}) = r$

Theorem:
Let X^- be any generalized inverse of an $n \times p$ matrix X. Then

$$X^* = X^-XX^- + (I - X^-X)A + B(I - XX^-)$$

is also a generalized inverse of X for any $p \times n$ matrices A and B. If X is another generalize dinverse of X, then there is a choice of A and B for which $X^* = X$.

We will not show the proof as it is more important to remember the form of the new generalized inverse.

Theorem:
For any matrix A, there exists a generalized inverse of A.

Theorem:
If G_1 and G_2 are generalized inverses of A, then so is G_1AG_2.

Theorem:
If A is symmetric, there exists a generalized inverse of A that is symmetric (note: the MP generalized inverse is symmetric).

Theorem (Projection Invariant to Choice of Generalized Inverse)
Suppose X is $n \times p$ of rank r. Consider the matrix $X'X$. If G and H are generalized inverses of $X'X$, then

1. $XGX'X = XHX'X = X$
2. $XGX' = XHX'$ (invariant to choice of g-inv)

Proof:

1. Let $v \in R^n$ and write $v = v_1 + v_2$ where $v_1 \in C(X)$ and $v_2 \in C(X)^{\perp}$. Then $v_1 = Xb$ for some $b \in R^p$. Thus,

$$\begin{aligned} v'XGX'X &= (v_1' + v_2')XGX'X \\ &= v_1'XGX'X \text{ since } v_2'X = 0 \\ &= b'X'(XGX'X) \\ &= b'(X'X)G(X'X) \\ &= b'(X'X) \\ &= v_1'X \\ &= v'X \end{aligned}$$

   Since v and G were arbitrary, we have $XGX'X = X$ for any G.

2. Let $v \in R^n$ with $v = v_1 + v_2$. Then $v_1 = Xb$ for some b. Thus,

$$\begin{aligned} XGX'v &= XGX'(v_1 + v_2) \\ &= XG(X'v_1 + X'v_2) \\ &= XGX'Xb \\ &= XHX'Xb \text{ by part i of this theorem} \\ &= XHX'v \end{aligned}$$

   since v was arbitrary, we have $XGX' = XHX'$.

This theorem leads to the following important result.

Theorem:
Suppose X is an $n \times p$ matrix of rank r. Let $M = X(X'X)^{-}X'$, where $-$ denotes <u>any</u> generalized inverse. Then M is the unique orthogonal projection operator onto $C(X)$.

Proof:
We use the definition of orthogonal projection operator.

(1) Let $v \in C(X)$. Then $v = Xb$ for some b. Thus,

$$Mv = X(X'X)^{-}X'v = X(X'X)^{-}X'(Xb) = Xb = v$$

The second equality follows by part (i) of the previous theorem.

(2) Let $w \in C(X)^{\perp} = \mathcal{N}(X')$. Then $X'w = 0$. We have

$$Mw = X(X'X)^{-}X'w = X(X'X)^{-}0 = 0$$

Summary of ways to compute orthogonal projection operator

1. $M = QQ'$, where Q is the matrix from the QR decomposition of X.
2. $M = V_1V_1'$, where V_1 comes from the SVD of X.
3. $M = XX^+$ where X^+ is the MP generalized inverse of X.
4. $M = X(X'X)^-X'$ where $(X'X)^-$ is any g-inv of X'X.

Theorem:
Suppose M and M_0 are orthogonal projection operators with $C(M_0) \subset C(M)$. Then

1. $MM_0 = M_0M = M_0$
2. $M - M_0$ is an orthogonal projection operator
3. $C(M_0) \perp C(M - M_0)$
4. $C(M - M_0) = C(M) \cap C(M_0)^\perp$
5. $M - M_0$ is the OPO onto $C(M - M_0)$.

Proof:

1. (Different from notes)
   Let v be a vector and write $v = v_1 + v_2$, where $v_1 \in C(M_0)$ and $v_2 \in C(M_0)^\perp$. We have

   $$\begin{aligned} MM_0v &= MM_0(v_1 + v_2) \\ &= MM_0v_1 + MM_0v_2 \\ &= Mv_1 + 0 \\ &= v_1 \\ &= M_0v_1 + M_0v_2 \\ &= M_0(v_1 + v_2) \\ &= M_0v \end{aligned}$$

   Since v was an arbitrary vector, we have $MM_0 = M_0$. Now, note that

   $$(MM_0)' = M_0'M' = M_0M$$

   since M, M_0 are OPOs and hence symmetric. The result follows.

2. It is clear that $M - M_0$ is symmetric.

   $$(M - M_0)(M - M_0) = MM - M_0M - MM_0 + M_0M_0 = M - M_0$$

   where we use (i).

3. Let $x \in C(M_0)$ and let $y \in C(M - M_0)$. Then $x = M_0r$ for some r and $y = (M - M_0)s$ for some s. Hence,

   $$M_0(M - M_0) = M_0M - M_0M_0 = M_0 - M_0 = 0 \tag{2.1}$$

   where we use (i).

4. Let $x \in C(M - M_0)$. Then $x = (M - M_0)x = Mx - M_0x = Mx$. Thus, $x \in C(M)$. Hence, $x \in C(M) \cap C(M_0)^\perp$.

   Let $x \in C(M) \cap C(M_0)^\perp$. Then $x \in C(M)$ and $x \in C(M_0)^\perp$. Thus, $Mx = x$ and $M_0x = 0$. Hence,

   $$(M - M_0)x = Mx - M_0x = Mx - 0 = Mx = x$$

   Hence, $x \in C(M - M_0)$.

5. Use part (iv).

Theorem:
Suppose M and M_0 are orthogonal projection operators with $C(M_0) \subset C(M)$. Then $C(M - M_0)$ is the orthogonal complement of $C(M_0)$ with respect to C(M)

Proof:
From (iii) above, $C(M - M_0) \perp C(M_0)$. Thus, $C(M - M_0)$ is contained in the orthogonal complement of $C(M_0)$ with respect to $C(M)$. If $x \in C(M) cap C(M_0)^\perp$, then
$Mx = x = (M - M_0)x + M_0x = (M - M_0)x$ so $x \in C(M - M_0)$.

Hence, the orthogonal complement of $C(M_0)$ with respect to $C(M)$ is contained in $C(M - M_0)$.

Hence,

$$C(M) = C(M_0) + C(M - M_0)$$

and thus $r(M) = r(M_0) + r(M - M_0)$.

Theorem:
Suppose M_1 and M_2 are two orthogonal projection operators in R^n. Then $M_1 + M_2$ is the orthogonal projection operator onto $C(M_1, M_2)$ if and only if $C(M_1) \perp C(M_2)$.

Note: $C(M_1) \perp C(M_2) iff M_1M_2 = M_2M_1 = 0$.

Theorem:
If $M_1 and M_2$ are symmetric matrices with $C(M_1) \perp C(M_2)$, and $M_1 + M_2$ is an orthogonal projection operator, then M_1 and M_2 are orthogonal projection operators.

3 Random Vectors and Matrices

Theorem:
Let Y be an $s \times 1$ random vector and W an $r \times 1$ random vector with $\text{Cov}(W) = \Sigma_w$, $\text{Cov}(Y) = \Sigma_y$, $Cov(W, Y) = \Sigma_{wy}$. Then

$$\text{Cov}(AW + BY) = A\Sigma_w A' + B\Sigma_y B' + A\Sigma_{wy} B' + B\Sigma_{yw} A'$$

Theorem:
Covariance matrices are always positive semidefinite.

Proof: Let $Z = Y - \mu$ so that $\mathrm{Cov}(Y) = \mathrm{Cov}(Z)$. Then

$$x'\mathrm{Cov}(Y)x = x'\mathbb{E}ZZ'x = \mathbb{E}x'ZZ'x = \mathbb{E}||Z'x||^2 \geq 0$$

since the norm of a vector is always nonnegative and the expectation of nonnegative numbers is nonnegative. Hence, covariance matrices are always positive semidefinite.

4 Estimability

Definition: (Estimability)
$\lambda'\beta$ is estimable if there exists $n \times 1$ vector of constants ρ, such that

$$\mathbb{E}(\rho'Y) = \lambda'\beta$$

for any β.

Definition (Linear Estimate)
$f(Y)$ is a linear estimate of $\lambda'\beta$ if

$$f(Y) = a_0 + a'Y$$

for some nonrandom a_0 and a.

NB: Thus, $\lambda'\beta$ is estimable if there exists a linear unbiased estimate of it. This leads to the following theorem.

Theorem:
$a_0 + a'Y$ is unbiased for $\lambda'\beta$ if and only if $a_0 = 0$ and $a'X = \lambda'$. This implies that $\lambda'\beta$ is estimable if and only if $\lambda \in C(X')$.

Proof: ($\Longleftarrow$)
If $a_0 = 0$ and $a'X = \lambda'$ then $\mathbb{E}(a_0 + a'Y) = 0 + a'X\beta = \lambda'\beta$

($\Longrightarrow$)
If $a_0 + a'Y$ is unbiased for $\lambda'\beta$, then

$$\lambda'\beta = E(a_0 + a'Y) = a_0 + a'X\beta$$

for any β. Hence,

$$(\lambda' - a'X)\beta = a_0 \text{ for any } \beta$$

This can only be true if $a_0 = 0$ and $\lambda' = a'X$.

Corollary:
$\lambda'\beta$ is estimable if and only if there exits an $n \times 1$ vector ρ such that

$$\rho'X = \lambda'$$

i.e., $\lambda = X'\rho$, i.e., $\lambda'\beta$ is estimable if $\lambda \in C(X')$. We note that the concept of estimability is based entirely on the assumption that $\mathbb{E}(Y) = X\beta$, but not on $\mathrm{Cov}(Y)$.

The next definition is based on the estimability of several linear combinations of the β vector.

Definition:
Suppose Λ is a $p \times s$ matrix of constants. Then the $s \times 1$ vector of linear functions $\Lambda\beta$ is estimable if and only if there exists an $n \times s$ matrix of constants P such that

$$P'X = \Lambda'$$

<u>Remarks</u> $Lambda'\beta$ is estimable if each of its components are estimable. P is not unique, but MP is unique. If X is full rank, then β is estimable and every linear combination of the components is estimable. Moreover, if X is full rank, we can pick $P' = (X'X)^{-1}X'$ so $P'X\beta = \beta$

5 Least Squares Estimation

Definition (Least Squares Estimator)
The least squares estimate of β, denoted $\hat{\beta}$, satisfies

$$(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \min_{\beta}(Y - X\beta)'(Y - X\beta)$$

That is, the least squares estimator of β minimizes the Euclidean squared distance between Y and its mean $\mu = X\beta$. Note that this vector is the closest vector to Y that is in $C(X)$, which we know is MY. We are led to the following theorem.

Theorem:
$\hat{\beta}$ is a least squares solution to β if and only if $X\hat{\beta} = MY$.

Proof:
Let $\tilde{\beta}$ be an arbitrary estimate of β. We can write

$$\begin{aligned}(Y - X\tilde{\beta})'(Y - X\tilde{\beta}) &= (Y - MY + MY - X\tilde{\beta})'(Y - MY + MY - X\tilde{\beta}) \\ &= (Y - MY)'(Y - MY) + (Y - MY)'(MY - X\tilde{\beta})(MY - X\tilde{\beta})'(Y - MY) + (MY - \\ &= (Y - MY)'(Y - MY) + (MY - X\tilde{\beta})'(MY - X\tilde{\beta})\end{aligned}$$

The first term does not depend on $\tilde{\beta}$ and the second term is a norm, which is minimized when the vector is 0. This happens if and only if $MY = X\tilde{\beta}$.

Theorem:
$\lambda' = \rho'X$ iff $\lambda'\hat{\beta}_1 = \lambda'\hat{\beta}_2$ for any $\hat{\beta}_1, \hat{\beta}_2$ satisfying

$$X\hat{\beta}_1 = MY, \quad X\hat{\beta}_2 = MY$$

Proof: See pages 121-122.

Corollary:
The unique least squares estimate of $\rho' X\beta = \rho' MY$

Corollary:
The unique least squares estimate of $P' X\beta = P' MY$

Corollary:
The unique least squares estimator of $\mu = X\beta$ is MY.

Theorem:
If $\lambda' = \rho' X$ then $\mathbb{E}(\rho' MY) = \lambda'\beta$.

Proof:

$$\mathbb{E}(\rho' MY) = \rho' M\mathbb{E}(Y) = \rho' MX\beta = \rho' X\beta = \lambda'\beta$$

Theorem:
Suppose Y is a random $n \times 1$ $E(Y) = \mu$ and Cov(Y) *Sigma*. Moreover, suppose A is any $n \times n$ matrix. Then

$$E(Y'AY) = \mu' A\mu + tr(A\Sigma)$$

Proof:

$$\begin{aligned} E(Y'AY) &= E(\mathrm{tr}(Y'AY)) \\ &= E(\mathrm{tr}(AYY')) \\ &= \mathrm{tr}(E(AYY')) \\ &= \mathrm{tr}(A\mathbb{E}YY') \\ &= \mathrm{tr}(A(\mu\mu' + \Sigma)) \\ &= \mathrm{tr}(A\mu\mu' + A\Sigma) \\ &= \mathrm{tr}(A\mu\mu') + \mathrm{tr}(A\Sigma) \\ &= \mathrm{tr}(\mu' A\mu) + \mathrm{tr}(A\Sigma) \\ &= \mu' A\mu + \mathrm{tr}(A\Sigma) \end{aligned}$$

5.a Estimation of σ^2

Theorem:
Suppose $r(X) = r$. Then

$$\frac{||(I-M)Y||^2}{n-r} = \frac{Y'(I-M)Y}{n-r}$$

is unbiased for σ^2.

The proof is presented in full on page 126, but it is very easy. Simply use the theorem above on how to take expectations of quadratic forms.

Below, we present the most important theorem in linear models.

Theorem: (Gauss-Markov)
Consider the linear model

$$Y = X\beta + \epsilon$$

where $E(\epsilon) = 0$ and $\text{Cov}(\epsilon) = \sigma^2 I$. If $\lambda'\beta$ is estimable, then the (unique) least squares estimate of $\lambda'\beta$ is the unique best linear unbiased estimator (BLUE) of $\lambda'\beta$.

Proof: We must show that the least squares estimate is (1) a linear estimator, (2) unbiased, (3) it is "best" in the sense that it has minimum variance among the class of unbiased estimators, and (4) it is unique. (1) and (2) are trivial, so we will show (3) and (4).

Let $a'Y$ be another linear unbiased estimator of $\lambda'\beta$. Recall that the unique least squares estimator of $\lambda'\beta = \rho' MY$.

$$\begin{aligned} \text{Var}(a'Y) &= \text{Var}(\rho'MY + a'Y - \rho'MY) \\ &= \text{Var}(\rho'MY) + \text{Var}(a'Y - \rho'MY) + 2\text{Cov}(\rho'MY, a'Y - \rho'MY) \end{aligned}$$

The first term is the variance of the least squares estimator and the second term is nonnegative, so we only need to show that the covariance term is also nonnegative (it is in fact 0).

Now,

$$\begin{aligned} \text{Cov}(\rho'MY, a'Y - \rho'MY) &= \text{Cov}((\rho'M)Y, (a' - \rho'M)Y) \\ &= \rho'M\text{Cov}(Y)(a - M\rho) \\ &= \sigma^2\rho'M(a - M\rho) \\ &= \sigma^2\rho'(Ma - M\rho) \end{aligned}$$

Since $a'Y$ is unbiased, we have $\lambda'\beta = \mathbb{E}a'Y = a'\mathbb{E}Y = a'X\beta$. Since $\lambda'\beta$ is estimable, we have $\lambda \in C(X')$. Hence, $\lambda' = \rho'X$ for some ρ. Putting these together, we have $a'X\beta = \lambda'\beta = \rho'X\beta$ for any β. Thus, $a'X = \lambda' = \rho'X$.

Hence

$$\begin{aligned} a'M - \rho'M &= a'X(X'X)^{-}X' - \rho'X(X'X)^{-}X' \\ &= \lambda'(X'X)^{-}X'\lambda'X(X'X)^{-}X' \\ &= 0 \end{aligned}$$

Plugging this into the covariance expression above, we get that the covariance is indeed 0. Thus, $\text{Var}(a'Y) \geq \text{Var}(\rho'MY)$. This completes the proof of (3).

For (4), note that if an estimator is BLUE, then $\text{Var}(a'Y - \rho'MY)$ must be 0 (otherwise, it would not be minimum variance). Thus,

$$\begin{aligned} 0 &= \text{Var}(a'Y - \rho'MY) \\ &= \text{Var}((a' - rho'M)Y) \\ &= (a' - \rho'M)\sigma^2 I(a - M\rho) \\ &= \sigma^2(a' - \rho'M)(a - M\rho) \\ &= \sigma^2||a - M\rho||^2 \end{aligned}$$

The norm ($||.||$) is a nonnegative function and equals 0 if and only if the vector inside the norm is 0. Thus, we must have $a - M\rho = 0$, so $a = M\rho$, i.e., the two estimators are identical. This completes (4).

6 Weighted Least Squares

Consider the linear model

$$\begin{aligned} Y &= X\beta + \epsilon \\ \mathbb{E}\epsilon &= 0 \text{ and } \mathrm{Cov}(\epsilon) = \sigma^2 V \end{aligned}$$

where V is a known positive definite matrix.

Since V is positive definite, we can write $V = QQ'$ for a nonsingular matrix Q, so that $Q^{-1}VQ'^{-1} = I$

We can apply a transformation to the model above

$$Q^{-1}Y = Q^{-1}X\beta + Q^{-1}\epsilon$$

It is easy to check that $\mathbb{E}Q^{-1}\epsilon = 0$ and $\mathrm{Cov}(Q^{-1}\epsilon) = \sigma^2 I$. From here, we can apply the least squares theory learned in the previous section. The least squares estimate of β minimizes

$$\begin{aligned} (Q^{-1}Y - Q^{-1}X\beta)'(Q^{-1}Y - Q^{-1}X\beta) &= (Y - X\beta)'(Q'^{-1}Q^{-1})(Y - X\beta) \\ &= (Y - X\beta)'(QQ')^{-1}(Y - X\beta) \\ &= (Y - X\beta)'V^{-1}(Y - X\beta) \end{aligned}$$

This leads us to a theorem about the estimability of weighted least squares

Theorem: (Estimability of Weighted Least Squares)

1. $\lambda'\beta$ is estimable in untransformed model if and only if $\lambda'\beta$ is estimable in the transformed model.
2. $\hat{\beta}$ is a weighted least squares estimate of β if and only if
$$X(X'V^{-1}X)^{-}X'V^{-1}Y = X\hat{\beta}$$
3. For any estimable function $\lambda'\beta$, $\lambda' = \rho'X$, the unique weighted least squares estimate of $\lambda'\beta$ is $\rho'AY$, where $A = X(X'V^{-1}X)^{-}X'V^{-1}$. The unique weighted least squares estimate of $\mu = X\beta$ is AY.
4. For any estimable function $\lambda'\beta$, $\lambda' = \rho'X$, $\rho'AY$ is the BLUE of $\lambda'\beta$, where A is the matrix in (3).

Theorem:
Let $A = X(X'V^{-1}X)^{-}X'V^{-1}$. Then

1. A is invariant with respect to choice of generalized inverse.
2. A is a projection operator onto $C(X)$ along $\mathcal{N}(A)$.

Proof:
(1)

$$\begin{aligned} A &= X(X'V^{-1}X)^{-}X'V^{-1} \\ &= V^{1/2}V^{-1/2}X(X'V^{-1/2}V^{-1/2}X)^{-}X'V^{-1/2}V^{-1/2} \\ &= V^{1/2}B(B'B)^{-}BV^{-1/2} \qquad\qquad = V^{1/2}M_BV^{-1/2} \end{aligned}$$

where $B = V^{-1/2}X$. and hence M_B is the orthogonal projection operator onto $C(B)$. Since M_B is invariant to the choice of generalized inverse, so is A.

The proof for (2) is a little more involved. Let $V = QQ'$, where Q, Q' are nonsingular since V is positive definite. Consider the orthogonal projection operator onto $C(Q^{-1}X)$. Call this projection matrix P. Then

$$\begin{aligned} P &= (Q^{-1}X)[(Q^{-1}X)'(Q^{-1}X)]^{-}(Q^{-1}X)' \\ &= Q^{-1}X[X'(Q'Q)^{-1}X]^{-}X'Q^{-1} \\ &= Q^{-1}X[X'V^{-1}X]^{-}X'Q'^{-1} \end{aligned}$$

Since, P is the orthogonal projection operator onto $C(Q^{-1}X$, the following relationship must hold:

$$\begin{aligned} PQ^{-1}X = Q^{-1}X &\iff Q^{-1}X[X'V^{-1}X]^{-}X'Q'^{-1}Q^{-1}X = Q^{-1}X \\ &\iff Q^{-1}X[X'V^{-1}X]^{-}X'V^{-1}X = Q^{-1}X \\ &\iff Q^{-1}AX = Q^{-1}X \\ &\iff AX = X \end{aligned}$$

(Note: it is probably enough to end the proof here, but to be rigorous, we should show $Av = v$ for all $v \in C(X)$.

Now, let $v \in C(X)$. Then $v = Xb$ for some b. Thus,

$$Av = AXb = Xb = v$$

where the second equality follows by the projection matrix argument above.

Remark: A is a projection operator, but not an orthogonal projection operator (it is not symmetric). It is, however, an OPO with respect to the inner product $x'V^{-1}y$. For this inner product space, x and y are orthogonal if $x'V^{-1}y = 0$ for $x \in C(X)$ and $y \in C(X)^{\perp}$.

Theorem: (Column Space under Premultiplication)
Suppose X is $n \times p$ of rank r, and V is a positive definite matrix. Then $C(V^{-1}X) = C(X)$ if and only if $C(VX) = C(X)$

Proof:
We will prove one direction. The other direction is analogous.

($\Longrightarrow$)
Suppose $C(V^{-1}X) = C(X)$. We want to show that $C(VX) = C(X)$.

Now,

$$\begin{aligned} w \in C(VX) &\iff w = VXb \text{ for some } b \\ &\iff V^{-1}w = Xb \text{ for some } b \\ &\iff V^{-1}w \in C(X) = C(V^{-1}X) \\ &\iff V^{-1}w = V^{-1}Xb^* \text{ for some } b^* \\ &\iff w = Xb^* \text{ for some } b^* \\ &\iff w \in C(X) \end{aligned}$$

Since $w \in C(VX)$ was arbitrary, we have $C(VX) = C(X)$.

Corollary:
Suppose X has full rank p. Then the weighted least squares estimate of β is given by $\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$ and the OLS estimate of β is given by $\hat{\beta} = (X'X)^{-1}X'Y$. Then $\tilde{\beta} = \hat{\beta}$ if and only if $C(VX) = C(X)$.

Proof:
Proof of $\Longrightarrow$ is on p.139 of the notes.

Proof of ($\Longleftarrow$) is unknown to the author.

6.a Estimation of σ^2 in WLS

Let $M^* = Q^{-1}X(X'V^{-1}X)^-X'Q^{-1}$. Then an unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{||(I - M^*)Q^{-1}Y||^2}{n - r}$$

It can be shown that $(I - M^*)Q^{-1} = Q^{-1}(I - A)$ (p. 140). Thus, $||(I - M^*)Q^{-1}Y||^2 = ||Q^{-1}(I - A)Y||^2$, and

$$\hat{\sigma}^2 = \frac{Y'(I - A)'V^{-1}(I - A)Y}{n - r}$$

as shown on page 141.

Theorem:
$V^{-1}(I - A) = (I - A)'V^{-1}(I - A)$

Proof:
Note that

$$(I - A)'V^{-1}(I - A) = V^{-1}(I - A) - A'V^{-1}(I - A)$$

Thus, the proof will be complete if we can show the right term is 0, since this will imply that $A'V^{-1} = A'V^{-1}A$.

$$\begin{aligned} A'V^{-1}A &= [V^{-1}X(X'V^{-1}X)^{-}X']V^{-1} \times [X(X'V^{-1}X)^{-}X'V^{-1}] \\ &= V^{-1}X(X'V^{-1}X)^{-}(X'V^{-1}X)(X'V^{-1}X)^{-}X'V^{-1} \\ &= V^{-1}X(X'V^{-1}X)^{-}X'V^{-1} \\ &= A'V^{-1} \end{aligned}$$

Theorem: (page 142)
$AVA' = AV = VA'$

In the notes, it says $AVA = AV = VA'$, but the author does not believe this to be true. He has found that $AVA' = AV = VA'$.

Proof:
First, we can write

$$A = X(X'V^{-1}X)^{-}X'V^{-1} = V^{1/2}B(B'B)^{-}B'V^{-1/2}$$

where $B = V^{-1/2}X$. Thus,

$$\begin{aligned} AV &= V^{1/2}B(B'B)^{-}B'V^{-1/2}V \\ &= V^{1/2}B(B'B)^{-}B'V^{1/2} \end{aligned}$$

Now, note that $VA' = (AV')' = (AV)'$ since V is symmetric. Looking above, we see that AV is a symmetric matrix, so we have $VA' = (AV)' = AV$ and hence $VA' = AV$.

Finally, we have

$$\begin{aligned} AVA' &= (AV)A \\ &= V^{1/2}B(B'B)^{-}B'V^{1/2}V^{-1/2}B(B'B)^{-}B'V^{1/2} \\ &= V^{1/2}B(B'B)^{-}B'B(B'B)^{-}B'V^{-1/2} \qquad = V^{1/2}B(B'B)^{-}B'V^{1/2} = AV \end{aligned}$$

6.b Covariance Matrices of WLS Estimates

1. For WLS,

$$\begin{aligned} \text{Cov}(\rho'AY) &= (\rho'A)(\sigma^2V)(\rho'A)' \\ &= \sigma^2\rho'AVA'\rho \end{aligned}$$

2. The residuals sum to zero if an intercept term is included (as they do in the regular linear model. To see this, note that $1 \in C(X)$. Since A is a projection operator onto

$C(X)$, we have $1 \in C(A)$ since $C(A) = C(X)$ by definition. Hence,

$$\begin{aligned} 1'(I-A)Y &= 1'Y - 1'AY \\ &= 1'Y - (A'1)'Y \\ &= 1'Y - (A1)'Y \\ &= 1'Y - 1'Y \\ &= 0 \end{aligned}$$

Important Note:
The cautious reader will have noticed that no distributional assumptions other than $E(\epsilon) = 0$ and $\mathrm{Cov}(\epsilon) = \Sigma$ have been made in the discussion henceforth. Thus, the theory on least squares and BLUEs are valid regardless of the true distribution of Y. In order to construct hypothesis tests, confidence regions, and prediction regions, we need to make distributional assumptions. In particular, we will assume that $Y \sim N_n(X\beta, \Sigma)$.

7 Distribution Theory

7.a Chi-square Distribution

A random variable X is said to have a central chi-square distribution with n degrees of freedom, written $X \sim \chi^2(n)$ if X has density

$$f(x) = \frac{1}{\Gamma(n/2)}(\frac{1}{2})^{n/2} x^{n/2-1} e^{-x/2}$$

Note that the chi-squared density is the gamma density with parameter $(\alpha = n/2, \beta = 1/2)$.

Let $\psi_X(t)$ denote the MGF of a random variable X. Recall that $\psi_X(t) = E(e^{tX})$.

Theorem:
If $X \sim \chi^2(n)$, then $\psi_X(t) = (1-2t)^{-n/2}$.

Proof:

$$\begin{aligned} \psi_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\Gamma(n/2)}(\frac{1}{2})^{n/2} x^{n/2-1} e^{-x/2} dx \\ &= \frac{1}{\Gamma(n/2)2^{n/2}} \int_{-\infty}^{\infty} x^{n/2-1} e^{-x(\frac{1}{2}-t)^{-1}} dx \\ &= \frac{[(1/2-t)^{-1}]^{-n/2}}{2^{n/2}} \\ &= \frac{(\frac{1-2t}{2})^{n/2}}{2^{n/2}} \\ &= (1-2t)^{n/2} \end{aligned}$$

The third equality follows because the integrand is the kernel of a $\Gamma(n/2, 1/2-t)$ random variable.

7.b Normal Distribution

A random variable X is said to have a normal distribution with mean μ and variance σ^2, written $X \sim N(\mu, \sigma^2)$, if X has density

$$f(x) = (2\pi)^{-1/2}\sigma^{-1}\exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\}$$

Theorem:
If $X \sim N(\mu, \sigma^2)$, then $\psi_X(t) = \exp\{t\mu + \frac{1}{2}t^2\sigma^2\}$

Proof: See page 147 in the notes. Proof involves complete the square.

Theorem:
Suppose $Z_1, \ldots, Z_n$ are independently and identically distributed $N(0,1)$ random variables. Define

$$X = \sum_{i=1}^{n} Z_i^2$$

Then $X \sim \chi^2(n)$

Proof:
Let $X_i = Z_i^2$ for $i = 1, \ldots, n$ and let $\gamma^2 = (1-2t)^{-1}$. Then

$$\begin{aligned}\psi_{X_i}(t) &= E(e^{tX_i})\\ &= E(e^{tZ_i^2}\\ &= \int_{-\infty}^{\infty} e^{tz^2}(2\pi)^{-1/2}e^{-z^2/2}dz \quad = \int_{-\infty}^{\infty}(2\pi)^{-1/2}e^{-\frac{z^2}{2(1-2t)^{-1}}}dz = \gamma\int_{-\infty}^{\infty}(2\pi\sigma^2)^{-1/2}e^{-\frac{z^2}{2\gamma^2}}dz \quad = \gamma\\ &= (1-2t)^{-1/2}\end{aligned}$$

Since the observations are independent, we have

$$\begin{aligned}\psi_X(t) &= \prod_{i=1}^{n}\psi_{X_i}(t)\\ &= [(1-2t)^{-1/2}]^n\\ &= (1-2t)^{-n/2}\end{aligned}$$

which is the MGF of a $\chi^2(n)$ random variable. By the uniqueness theorem for MGFs, we have $X \sim \chi^2(n)$.

7.c Noncentral Chi-Square Distribution

Theorem:
Suppose $Y_1, \ldots, Y_n$ are independent and $Y_i \sim N(\mu_i, \sigma^2), i = 1, \ldots, n$. Define

$$X = \frac{1}{\sigma^2} \sum_{i=1}^{n} Y_i^2$$

Then $X \sim \chi^2(n, \gamma)$, where $\gamma = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \mu_i^2$.

7.c.1 Properties of Noncentral Chi-square

If $X \sim \chi^2(n, \gamma)$,

1. $$\psi_X(t) = (1 - 2t)^{-n/2} \exp\{\frac{2\gamma t}{1 - 2t}\}$$

2. $$\begin{aligned} E(X) &= n + 2\gamma \\ \text{Var}(X) &= 2n + 8\gamma \end{aligned}$$

3. If $\gamma = 0$, then this corresponds to a central chi-square random variable with n degrees of freedom.

7.d t Distribution

Suppose $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, and $X and Y$ are independent. Define the random variable

$$T = \frac{T}{\sqrt{Y/n}}$$

Then T is said to have a t distribution with n degrees of freedom, and we write $T \sim t(n)$.

7.e Noncentral t Distribution

Suppose $X \sim N(\mu, 1)$ and $Y \sim \chi^2(n)$, and X and Y are independent. Define the random variable

$$W = \frac{X}{\sqrt{Y/n}}$$

Then W is said to have a noncentral t distribution with n degrees of freedom and noncentrality parameter μ. We write $W \sim t(n, \mu)$. If $\mu = 0$, then W reduces to a central t distribution with n degrees of freedom.

7.f F Distribution

Suppose $X_1 \sim \chi^2(n_1, \gamma_1)$ and $X_2 \sim \chi^2(n_2, \gamma_2)$, and X_1 and X_2 are independent. Then the random variable

$$F = \frac{X_1/n_1}{X_2/n_2}$$

is said to have a doubly noncentral F distribution with $(n_1, n_2$ degrees of freedom and noncentrality parameters (γ_1, γ_2). We write $F \sim F(n_1, n_2, \gamma_2, \gamma_2)$.

1. If $\gamma_2 = 0$, F is said ot have a noncentral F distribution. We represent this as $F \sim F(n_1, n_2, \gamma_1)$
2. If $\gamma_1 = 0\ \gamma_2$, then F is said to have a central F distribution. We represent this as $F \sim F(n_1, n_2)$.

7.g Multivariate Moment Generating Functions

Suppose $X = (X_1, \ldots, X_n)'$ is an $n \times 1$ random vector with n dimensional density f. The multivariate moment generating function of X is defined as

$$\psi_X(t) = E(e^{t'X})$$

7.g.1 Properties of Multivariate MGF

1. $\psi_X(0) = 1$
2. If $X_1, \ldots, X_n$ are independent then

$$\psi_X(t) = \prod_{i=1}^{n} \psi_{X_i}(t_i)$$

3. Moments can be obtained by differentiating the multivariate MGF.
4. The MGF for any marginal distribution of X is obtained by setting equal to 0 those t_j's that correspond to the X_j's not in the marginal distribution.
5. The multivariate characteristic function is defined as

$$\phi_X(t) = E(e^{it'X})$$

Note that the MGF may not always exist (e.g., Cauchy distribution), but the CF always exists for any random variable.

7.h Multivariate Normal Distribution

Definition:
Suppose $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ random variables. Let $Z = (Z_1, \ldots, Z_n)'$. We have $E(Z) = 0$ and $\text{Cov}(Z) = I$. We say that Y has an r-dimensional multivariate normal distribution if Y has the same distribution as $AZ + b$ for some $r \times n$ matrix of constants A and an $r \times 1$ vector of constants b. We denote the distribution of Y by

$$Y \sim N_r(b, AA')$$

Note that if Σ is not positive definite, then Y is said to be singular normal and the density of Y does not exist.

Definition:
Suppose $X = (X_1, \ldots, X_n)'$. Then X is said to have an n-dimensional multivariate normal distribution with mean μ and covariance matrix Σ if X has density

$$f(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) \right\}$$

Obviously, this density does not exist if Σ is singular, so this definition is less general than the one above.

7.i Properties of MVN Distributions

1. Suppose $X \sim N_n(\mu, \Sigma)$, then

$$\psi_X(t) = \exp\left\{ t'\mu + \frac{1}{2}t'\Sigma t \right\}$$

   Note that the MGF exists even with Σ is singular, and so the MGF is a good way to define a multivariate normal random variable (e.g., $Y \sim MVN(.,..)$ if Y has MGF...

2. A linear transformation of MVNs is MVN. Suppose $X \sim N_n(\mu, \Sigma)$, and define $Y = AX + b$, where A is an $r \times n$ matrix of constants and b is an $r \times 1$ vector of constants. Then

$$Y \sim N_r(A\mu + b, A\Sigma A')$$

   This result can be proved easily using MGF.

3. A linear combination of independent MVN's is MVN.

4. Marginal distributions of MVN are MVN.

5. Conditional distributions of MVN are MVN. In particular,

$$X_1 | X_2 = x_2 \sim N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11.2})$$

   where $\Sigma_{11.2} = \Sigma_1 1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Theorem:
If $X \sim N_n\mu, \Sigma$, then all marginals, conditionals, and linear combinations of the components of X are MVN.

Note: the converse is false.

7.i.1 Independence of MVN

Definition (Independence)
Two random vectors are independent if their joint density $f(x, y)$ factors into $f(x, y) = f_1(x)f_2(y)$.

Theorem:
If X and Y are independent random vectors, then $G(X)$ and $H(Y)$ are independent, where G and H are arbitrary functions.

Theorem:
Suppose $X \sim N_n(\mu, \Sigma)$. Define $Y_1 = AX$ and $Y_2 = BX$, where A is an $r \times n$ matrix of constants and B is an $s \times n$ matrix of constants. Then Y_1 and Y_2 are independent if and only if $A\sigma B' = 0$. If $\Sigma = \sigma^2 I$, then Y_1 and Y_2 are independent if and only if $AB' = 0$.

Proof:
First, let $Y = (Y_1, Y_2)'$. By the theorem above, we know that $Y \sim N\left(\begin{bmatrix} A\mu \\ B\mu \end{bmatrix}, \begin{bmatrix} A\Sigma A' & A\Sigma B' \\ B\Sigma A' & B\Sigma B' \end{bmatrix}\right)$.
$(\implies)$
By the above, we have for any $t = (t_1, t_2)'$,

$$\begin{aligned} \psi_Y(t) &= \exp\{(A\mu B\mu)'t + \frac{1}{2}t' \begin{bmatrix} A\Sigma A' & A\Sigma B' \\ B\Sigma A' & B\Sigma B' \end{bmatrix} t\} \\ &= \exp\{(A\mu B\mu)'t + \frac{1}{2}(t_1' A\Sigma A' t_1 + t_1' A\Sigma B' t_2 + t_2' B\Sigma A' t_1 + t_2' B\Sigma B' t_2\} \end{aligned}$$

Now, by assumption, Y_1 and Y_1 are independent. Thus,

$$\begin{aligned} \psi_Y(t) &= E(e^{t'Y}) \\ &= E(e^{t_1'Y_1 + t_2'Y_2}) \\ &= E(e^{t_1'Y_1})E(e^{t_2'Y_2}) \\ &= e^{\left\{t_1'A\mu + \frac{1}{2}t_1'A\Sigma A' t_1\right\}} e^{\left\{t_2'B\mu + \frac{1}{2}t_2'B\Sigma B' t_2\right\}} \\ &= \exp\left\{(A\mu + B\mu)'t + \frac{1}{2}(t_1'A\Sigma A' t_1 + t_2' B\Sigma B' t_2)\right\} \end{aligned}$$

Since this holds for any t_1, t_2, we must have $A\Sigma B' = 0$.

$(\impliedby)$
Suppose $A\Sigma B' = 0$. Then

$$Y \sim N\left(\begin{bmatrix} A\mu \\ B\mu \end{bmatrix}, \begin{bmatrix} A\Sigma A' & 0 \\ 0 & B\Sigma B' \end{bmatrix}\right)$$

The rest is easy to verify. Assume $A\Sigma A'$ and $B\Sigma B'$ are positive definite so that the PDF of the joint distribution exists. Note that the determinant of a block diagonal matrix is the product of the determinants of the blocks. From here, it is easy to rearrange terms so that you get a product of two multivariate PDFs. Thus, Y_1 and Y_2 are independent.

Theorem: Suppose $X \sim N_n(\mu, \Sigma)$. Partition X into $X = (X_1', X_2')'$, where X_1 is $r \times 1$ and X_2 is $(n-r) \times 1$. Partition μ as $\mu = (\mu_1, \mu_2)'$. Similarly, partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

Proof:
We can simply apply the theorem above with

$$A = \begin{pmatrix} I_{r\times r} & 0_{r\times(n-r)} \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0_{n-r\times r} & I_{(n-r)\times(n-r)} \end{pmatrix}$$

Theorem:
If $X \sim N_n(\mu_x, \Sigma_x)$ and $Y \sim N_m(\mu_y, \Sigma_y)$, and X and Y are independent, then

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_{n+m}(\mu, \Sigma)$$

where $\mu = (\mu_x', \mu_y')'$ and $\Sigma = \text{blkdiag}\{\Sigma_x, \Sigma_y\}$

Proof: Can be easily shown using MGF.

7.j Distribution of Quadratic Forms

Suppose Y is an n dimensional random vector and let A be an $n \times n$ matrix of constants. A quadratic form is a random variable defined by $Y'AY$ for some Y and A.

Theorem:
Suppose $Y \sim N_n(0, \sigma^2 I)$, then

$$\frac{1}{\sigma^2}(Y'MY \sim \chi^2(r)$$

if and only if M is an <u>orthogonal</u> projection operator of rank r.

Theorem:
Suppose $Y \sim N_n(\mu, \sigma^2 I)$, then

$$\frac{1}{\sigma^2}(Y'MY \sim \chi^2(r, \gamma)$$

if and only if M is an <u>orthogonal</u> projection operator of rank r and $\gamma = \dfrac{\mu' M \mu}{2\sigma^2}$

Theorem:
Suppose $Y \sim N_n(\mu, \sigma^2 M)$ where M is an orthogonal projection operator of rank r and $\mu \in C(M)$. Then

$$Y'Y \sim \chi^2(r, \gamma)$$

where $\gamma = \dfrac{\mu'\mu}{2\sigma^2}$.

Theorem:
Suppose $Y \sim N_n(\mu, \Sigma)$ where Σ is positive definite. Then

$$Y'AY \sim \chi^2(r, \gamma)$$

where $\gamma = \dfrac{\mu' A \mu}{2}$ if and only if *any* of the following conditions are satisfied.

1. $A\Sigma$ is a projection operator of rank r
2. ΣA is a projection operator of rank r
3. $\Sigma$is a generalized inverse of A and A has rank r.

Note: of the aforementioned theorem, this one is the best to remember since it is the most general. The professors have been known to test on these general theorems (i.e., the other theorems do not apply to a problem), so it's best to have them memorized.

Theorem:
Suppose $Y \sim N_n(\mu, \Sigma)$, then $Y'AY \sim \chi^2(\text{tr}(A\Sigma, \gamma)), \gamma = \dfrac{\mu' A \mu}{2}$, if

1. $\Sigma A \Sigma A \Sigma = \Sigma A \Sigma$, and
2. $\mu' A \Sigma A \mu = \mu' A \mu$, and
3. $\Sigma A \Sigma A \mu = \Sigma A \mu$.

Theorem:
Suppose $Y \sim N_n(\mu, \Sigma)$, where Σ is positive definite. Then $Y'AY$ has the same distribution as the random variable

$$U = \sum_{i=1}^{n} d_{ii} U_i$$

where d_{ii} are the eigenvalues of $A\Sigma$ and $U_1, \ldots U_n$ are independent non-central chi-square random variables with one degree of freedom.

Proof: Assume WLOG that A is symmetric. Then

$$\begin{aligned}
Y'AY &= Y'\Sigma^{-1/2}\Sigma^{1/2}A\Sigma^{1/2}\Sigma^{-1/2}Y \\
&= X'\Sigma^{1/2}A\Sigma^{1/2}X \\
&= X'PDP'X \\
&= Z'DZ \\
&= \sum_{i=1}^{n} d_{ii} Z_i^2
\end{aligned}$$

where $\Sigma^{-1/2}$ exists since Σ is positive definite, $X = \Sigma^{-1/2}Y$, D and P are from the spectral decomposition of $\Sigma^{1/2}A\Sigma^{1/2}$ (note this matrix is symmetric, so spectral theorem applies), and $Z = P'X$.

Now,

$$X = \Sigma^{-1/2}Y \sim N(\Sigma^{-1/2}\mu, I)$$

and

$$Z = P'X \sim N(P'\Sigma^{-1/2}\mu, I)$$

since P is orthogonal.

Note that $\text{Cov}(Z)$ is diagonal, so the Z_i's are independent normal random variables with variance 1 and nonzero mean. Hence, each Z_i^2 is an independent noncentral chi-squared random variable.

To complete the proof, we must show that each d_{ii} is an eigenvalue of $A\Sigma$. From the spectral theorem, we know that $d_i i$ is an eigenvalue of $\Sigma^{1/2}A\Sigma^{1/2}$.

Let λ be an eigenvalue for $\Sigma^{1/2}A\Sigma^{1/2}$. Then

$$\begin{aligned}
\det\bigl(\Sigma^{1/2}A\Sigma^{1/2} - \lambda I\bigr) = 0 & \\
&\iff \det\bigl(\Sigma^{1/2}(A\Sigma^{1/2} - \lambda\Sigma^{-1/2}I)\bigr) = 0 \\
&\iff \det\bigl(\Sigma^{1/2}\bigr)\det\bigl(A\Sigma^{1/2} - \lambda\Sigma^{-1/2}\bigr) = 0 \\
&\iff \det\bigl(A\Sigma^{1/2} - \lambda\Sigma^{-1/2}\bigr) = 0 \\
&\iff \det(A\Sigma - \lambda I) = 0 \\
&\iff \lambda \text{ is an eigenvalue of } A\Sigma
\end{aligned}$$

The result follows.

7.k Independence of Quadratic Forms

If $Y \sim N_n(\mu, \sigma^2 I)$, then

1. $Y'AY$ and BY are independent iff AB' = 0, where A is a symmetric matrix.
2. $Y'AY$ and $Y'BY$ are independent if and only if $AB = 0$, where A and B are symmetric.

Proof:
Note that AY and BY are independent if and only if $0 = \text{Cov}(AY, BY) = A\text{Cov}(Y)B' = \sigma^2 AB'$, but this happens if and only if $AB' = 0$.

1. Note that $Y'AY$ is a function of AY, and since AY and BY are independent if and only if $AB' = 0$, we have $Y'AY$ and BY are independent if and only if $AB' = 0$.

2. Note that $Y'AY$ and $Y'BY$ are functions of AY and BY, respectively. Thus,

$$\begin{aligned} &Y'AY, Y'BY \text{ are independent} \\ \iff\ & AB' = 0 \\ \iff\ & AB = 0 \text{ since B is symmetric} \end{aligned}$$

The result follows.

Theorem:
Suppose $Y \sim N_n(\mu, \Sigma)$ and suppose that A, B, and Σ are all positive semidefinite. Then $Y'AY$ and $Y'BY$ are independent if $\Sigma A \Sigma B \Sigma = 0$. If Σ is positive definite, then $Y'AY$ and $Y'BY$ are independent if $A\Sigma B = 0$.

Write $A = RR'$, $B = SS'$ and $\Sigma = QQ'$. Then $Y'AY = (R'Y)'(R'Y)$ and $Y'BY = (S'Y)'(S'Y)$, so $Y'AY$ and $Y'BY$ are independent if and only if $0 = \text{Cov}(S'Y, R'Y) = S'\Sigma R = S'QQ'R = (Q'S)'(Q'R)$. Recall that for any matrix A, $C(A) = C(AA')$. We have

$$\begin{aligned} (Q'S)'(Q'R) = 0 &\iff C(Q'S) \perp C(Q'R) \\ &\iff C(Q'SS'Q) \perp C(Q'RR'Q) \\ &\iff Q'SS'QQ'RR'Q = 0 \\ &\iff Q'A\Sigma BQ = 0 \\ &\iff C(A\Sigma BQ) \perp C(Q) \\ &\iff C(A\Sigma BQ) \perp C(QQ') \\ &\iff C(A\Sigma BQ) \perp C(\Sigma) \\ &\iff Q'B\Sigma A\Sigma = 0 \\ &\iff C(B\Sigma A\Sigma) \perp C(Q) \\ &\iff C(B\Sigma A\Sigma) \perp C(QQ') = C(\Sigma) \\ &\iff \Sigma A\Sigma B\Sigma = 0 \end{aligned}$$

as was to be shown.

If Σ is positive definite, then Σ is nonsingular, so Σ is invertible. Pre- and post-multiply $\Sigma A\Sigma B\Sigma$ by Σ^{-1} to get the result.

There is a more general result for when A and B are not symmetric. The details can be found on page 184 of the notes.

Remarks:

1. If $Y \sim N_n(\mu, \Sigma)$ and Σ is positive semidefinite, then AY and BY are independent $\iff$ $A\Sigma B' = 0$.
2. if Σ is full rank and $Y'AY$ and BY are independent, then AY and BY are independent. Also, if $Y'AY$ and $Y'BY$ are independent, then AY and BY are independent.
3. If Σ is not full rank, then the result in (2) does NOT hold.

7.l Minimum Variance Unbiased Estimation

Definition: (Completeness)
Suppose $T(Y)$ is a vector-valued statistic in Y. $T(Y)$ is said to be a complete sufficient statistic for the family of distributions indexed by $\theta \in \Theta$ if $T(Y)$ is sufficient and

$$E[f(T(Y))] = 0 \implies f(T(Y)) = 0 \text{ w.p.1}$$

for all $\theta in \Theta$ where Θ denotes the paramter space.

Theorem: (Lehmann-Scheffe)
If $T(Y)$ is a complete sufficient statistic, then $f(T(Y))$ is the unique UMVUE of $E(f(T(Y)))$.

Theorem: (Complete Sufficient Statistics for Exponential Families
Let $\theta = (\theta_1, \dots, \theta_p)'$ and let Y be a random vector with density

$$f(Y) = h(Y)c(\theta)\exp\left\{\sum_{i=1}^{p} \theta_i T_i(Y)\right\}$$

then $T(Y) = (T_1(Y), \dots, T_p(Y))'$ is a $\underline{\text{complete sufficient statistic}}$ for θ, if $\theta \in A$, where A is an open subset in R^p.

The requirement for an open subset always exists if there are no restrictions on the parameter vector θ. If there are constraints on the components of θ, then an open subset does not exist. This is the case in linear models when X is less than full rank.

Theorem: (UMVUE for linear models)
Suppose $Y = X\beta + \epsilon$ where $\epsilon \sim N_n(0, \sigma^2 I)$ and $r(X) = r$. Then

$$\frac{Y'(I-M)Y}{n-r}$$

is the unique UMVUE of σ^2 and $\rho' MY$ is the unique UMVUE of $\rho' X\beta$

7.m Sampling Distribution of Estimates

Suppose $Y = X\beta + \epsilon$ and $\epsilon \sim N_n(0, \sigma^2 I)$. Suppose $\Lambda'\beta$ is an estimable vector of linear functions of β, where Λ is a $p \times s$ matrix. By definition, there exists some $n \times s$ matrix P such that $\Lambda' = P'X$. It is easy to see that $E(P'MY) = P'X\beta$ and $\text{Cov}(P'MY) = \sigma^2(P'MP)$. Thus,

$$P'MY \sim N_s(P'X\beta, \sigma^2 P'MP)$$

Note that $P'MY = P'X(X'X)^{-}X'P = \Lambda'(X'X)^{-}\Lambda$. Thus, we can write

$$P'MY \sim N_s(\Lambda'\beta, \sigma^2\Lambda'(X'X)^{-}\Lambda)$$

If X has full rank, then β is estimable and the UMVUE of β is $\hat{\beta} = (X'X)^{-1}X'Y)$ and

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$$

Note that $I - M$ is an orthogonal projection operator and $\hat{\sigma}^2 = \frac{Y'(I-M)Y}{n-r}$. Since $Y \sim N_n(X\beta, \sigma^2 I)$ and $I - M$ is an orthogonal projection operator, by a theorem we have

$$\frac{1}{\sigma^2} Y'(I - M)Y \sim \chi^2(n - r, \gamma)$$

where

$$\gamma = \frac{\beta' X(I - M)X\beta}{2\sigma^2} = 0$$

Thus,

$$\frac{1}{\sigma^2} Y'(I - M)Y \sim \chi^2(n - r)$$

7.n Hypothesis Testing

Consider the usual linear model $Y = X\beta + \epsilon$ with the usual assumptions on ϵ. Note that this model implies $E(Y) \in C(X)$. We are interested in testing this model against a reduced model $Y = X_0\gamma_0 + \epsilon$ where $C(X_0) \subset C(X)$.

We assume that the full model is correct, so if the reduced model is correct, then so is the full model. Since $C(X_0) \subset C(X)$, we say that the two models are nested.

Let $M = X(X'X)^{-}X'$ denote the orthogonal projection operator onto $C(X)$ and $M_0 = X_0(X_0'X_0)^{-}X_0'$ denote the orthogonal projection operator onto $C(X_0)$. Since $C(M_0) = C(X_0) \subset C(X) = C(M)$, we have by a previous theorem that $M - M_0$ is the orthogonal projection operator onto $C(M - M_0)$–the subspace within M that is orthogonal to M_0 (see page 100).

Theorem:
Consider the usual linear model $Y = X\beta + \epsilon$ where $\epsilon \sim N_n(0, \sigma^2 I)$. Consider the reduced model

$$Y = X_0\beta + \epsilon$$

where $C(X_0) \subset C(X)$. We wish to test the hypothesis

$$\begin{aligned} &H_0 : E(Y) in C(X_0) \\ &H_a : E(Y) \in C(X) \cap C(X_0)^C \end{aligned}$$

Let $M_0 = X_0(X_0'X_0)^{-}X_0$ be the orthogonal projection operator onto $C(X_0)$ and $M = X(X'X)^{-}X'$ the OPO onto C(X). Suppose X is $n \times p$ with $r(X) = r$ and $r(X_0) = r_0 < r$. Define the F-statistic as

$$F = \frac{Y'(M - M_0)Y/(r - r_0)}{Y'(I - M)Y/(n - r)}$$

Then under H_a,

$$F \sim F(r - r_0, n - r, \gamma)$$

where $\gamma = \dfrac{||(I - M_0)X\beta||^2}{2\sigma^2} = \dfrac{||(M - M_0)X\beta||^2}{2\sigma^2}$

If the reduced model is correct, then

$$F \sim F(r - r_0, n - r)$$

Proof:
By the distribution theory section, we need to show the numerator is $\chi^2(r - r_0, \gamma)$ and the denominator is $\chi^2(n - r)$, and the numerator and denominator are independent.

Note that $M - M_0$ is an orthogonal projection operator of rank $r - r_0$. Thus, $\dfrac{Y'(M - M_0)Y}{\sigma^2} \sim \chi^2(r - r_0, \gamma)$, where

$$\gamma = \frac{(X\beta)'(M - M_0)(X\beta)}{2\sigma^2} = \frac{||(M - M_0)X\beta||^2}{2\sigma^2}$$

Note that if the reduced model is correct, we have $E(Y) = X_0\beta$ and we replace X with X_0 in the expression for γ above. Thus,

$$(M - M_0)X_0\beta = MX_0\beta - M_0X_0\beta = X_0\beta - X_0\beta = 0$$

So that the F statistic has a central F distribution.

Moreover, $I - M$ is also an orthogonal projection operator and its rank is $n - r$. By the same theorem, we have

$$\frac{Y'(I - M)Y}{\sigma^2} \sim \chi^2(n - r, \eta)$$

where

$$\begin{aligned} \eta &= \frac{\beta' X'(I - M)X\beta}{2\sigma^2} \\ &= 0 \end{aligned}$$

(since $I - M$ is the orthogonal projection operator onto $C(X)^\perp$, we have $(I - M)X = 0.$).

Now, we want to show the two quadratic forms are independent. By a previous theorem, it suffices to show that $(M - M_0)(I - M) = 0$. By a previous theorem, it suffices to show that $(M - M_0)(I - M) = 0$.

$$\begin{aligned} (M - M_0)(I - M) &= MI - M_0I - MM + M_0M \\ &= M - M_0 - M + M_0 \\ &= 0 \end{aligned}$$

The second equality follows because of a previous theorem establishing $M_0 = M_0M = MM_0$. Hence, the numerator and denominator in the F statistic are independent.

7.0 Testing Linear Parametric Functions

Consider the usual linear model. Suppose we want to test the hypothesis concerning an estimable function $\Lambda'\beta$, where $\Lambda' = P'X$ and P' is an $s \times n$ matrix of constants. Suppose we want to test the hypothesis $H_0 : P'X\beta = 0$. This condition imposes a linear constraint on β. We need to find the reduced model X_0 that corresponds to this linear constraint. The restriction is $E(Y) \in \mathcal{N}(P') = C(P)^{\perp}$. Thus, the null hypothesis is $E(Y) \in C(X) \cap C(P)^{\perp}$.

Theorem:
Let $M_{MP} = MP(P'MP)^{-}P'M$, so that M_{MP} is the orthogonal projection operator onto $C(MP)$. Then

$$C((I - M_{MP})X) = C(X) \cap C(MP)^{\perp}$$

See the proof on pages 215-16.

As a result of this theorem and since $C(M) = C(X)$, we have a choice for X_0 is $X_0 = M - M_{MP}$. Note that the choice for X_0 is not unique.

Now, since $Y'(M - M_0)Y = Y'(M - (M - M_{MP}))Y = ||M_{MP}Y||^2$, we have that the F statistic is

$$F = \frac{||M_{MP}Y||^2/r(M_{MP})}{||(I-M)Y||^2/r(I-M)} \sim F(r(M_{MP}), r(I-M), \gamma)$$

where $\gamma = \dfrac{||M_{MP}X\beta||^2}{2\sigma^2}$. Under H_0, $\gamma = 0$ so F has a central F distribution.

Since the original hypothesis is $\Lambda'\beta = 0$ and $\Lambda'\beta$ is estimable (so $\Lambda' = P'X$), we can write the F statistic in terms of Λ and $\hat{\beta}$. We have

$$Y'M_{MP}Y = \hat{\beta}'\Lambda(\Lambda'(X'X)^{-}\Lambda)^{-}\Lambda'\hat{\beta}$$

Since $r(M_{MP}) = r(\Lambda)$, the F statistic can be written as

$$F = \frac{\hat{\beta}'\Lambda(\Lambda'(X'X)^{-}\Lambda)^{-}\Lambda'\hat{\beta}/r(\Lambda)}{MSE}$$

and we can write the noncentrality parameter was

$$\gamma = \frac{\beta'\Lambda(\Lambda'(X'X)^{-}\Lambda)^{-}\Lambda'\beta}{2\sigma^2}$$

A third and final way to represent the F test is noting that $\text{Cov}(\Lambda'\hat{\beta}) = \sigma^2\Lambda'(X'X)^{-}\Lambda$, so

$$F = \frac{(\Lambda'\hat{\beta})'[\text{Cov}(\Lambda'\hat{\beta})]^{-}(\Lambda'\hat{\beta})}{r(\Lambda)}$$

We summarize the ways of writing the F test below:

1. $$F = \frac{||(M-M_0)Y||^2/r(M-M_0)}{MSE} \sim F(r-r_0, n-r, \gamma)$$

   where $\gamma = \frac{||(I-M_0)X\beta||^2}{2\sigma^2} = \frac{||(M-M_0)X\beta||^2}{2\sigma^2}$

2. $$F = \frac{||M_{MP}Y||^2/r(M_{MP})}{MSE} \sim F(r(M_{MP}, r(I-M), \gamma)$$

   where $\gamma = \frac{||M_{MP}X\beta||^2}{2\sigma^2}$

3. $$F = \frac{(\Lambda'\hat{\beta})'(\Lambda'(X'X)^-\Lambda)^-(\Lambda'\hat{\beta})/r(\Lambda)}{MSE} \sim F(r(\Lambda), r(I-M), \gamma)$$

   where $\gamma = \frac{\beta'\Lambda(\Lambda'(X'X)^-\Lambda)^-\Lambda'\beta}{2\sigma^2}$

4. $$F = \frac{(\Lambda'\hat{\beta})'(\widehat{\text{Cov}}(\Lambda'\hat{\beta}))^-(\Lambda'\hat{\beta})}{r(\Lambda)} \sim F(r(\Lambda), r(I-M), \gamma)$$

   where $\gamma = \frac{\beta'\Lambda(\Lambda'(X'X)^-\Lambda)^-\Lambda'\beta}{2\sigma^2}$

7.p Generalized Hypothesis Test Procedure

Suppose we wish to test
$$H_0 : \Lambda'\beta = d$$
where $\Lambda' = P'X$ for some $s \times n$ matrix P', and d is a known $s \times 1$ vector. Since d is possibly nonzero, it is difficult to express the null space since the null hypothesis describes a flat, which is not a subspace. We must translate the flat back to the origin so we can write the null hypothesis in terms of subspaces.

See pages 221-225 for details. We can write the F test in multiple ways

1. Let b be <u>any</u> solution to $P'X\beta = d$. Then

$$\begin{aligned} F &= \frac{||M_{MP}(Y-Xb)||^2/r(M_{MP})}{MSE} \\ &\sim F(r(M_{MP}), r(I-M), \gamma) \end{aligned}$$

2. $$\begin{aligned} F &= \frac{(\Lambda'\hat{\beta}-d)'(\Lambda'(X'X)^-\Lambda)^-(\Lambda'\hat{\beta}-d)}{MSE} \\ &\sim F(r(\Lambda), r(I-M), \gamma) \end{aligned}$$

7.q Confidence Regions

Consider the problem of finding a confidence region for an estimable vector $\Lambda'\beta$ where $\Lambda' = P'X$. Let M_{MP} denote the OPO onto C(MP). Using the theory developed for tests of hypotheses, we have

$$\frac{(Y - X\beta)'M_{MP}(Y - X\beta)/r(M_{MP})}{(Y - X\beta)'(I - M)(Y - X\beta)/r(I - M)} \sim F(r(M_{MP}), r(I - M))$$

The noncentrality parameter is 0 since the random variables have mean 0. We can write

$$(Y - X\beta)'M_{MP}(Y - X\beta) = (\Lambda'\hat{\beta} - \Lambda'\beta)'(\Lambda'(X'X)^{-}\Lambda)^{-}(\Lambda'\hat{\beta} - \Lambda'\beta)$$

and

$$(Y - X\beta)'(I - M)(Y - X\beta) = Y'(I - M)Y$$

so that the denominator is simply the MSE. Thus, a $(1 - \alpha) \times 100\%$ confidence region for $\Lambda'\beta$ is

$$\left\{\beta : \frac{(\Lambda'\hat{\beta} - \Lambda'\beta)'(\Lambda'(X'X)^{-}\Lambda)^{-}(\Lambda'\hat{\beta} - \Lambda'\beta)/r(\Lambda)}{MSE} \leq F(1 - \alpha, r(\Lambda), r(I - M))\right\}$$

where $F(1-\alpha, r(\Lambda), r(I-M))$ is the upper $(1-\alpha)\times 100\%$ point of a central F distribution.

8 One-Way ANOVA

A one-way ANOVA model can be written as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where $i = 1, \dots, t$, $j = 1, \dots, n_i$. We will assume that the ϵ_{ij}'s are *i.i.d.* and $\epsilon_{ij} \sim N_1(0, \sigma^2)$. Let $n = \sum_{i=1}^{t} n_i$.

We can write the model above as $Y = X\beta + \epsilon$ where $\epsilon = N_n(0, \sigma^2 I)$. We want to construct the OPO onto $C(X)$. The design matrix is

$$X = (J, X_1, \dots X_t)$$

where J is a vector of ones and X_k is a vector of zeros and ones. We can write X_k as

$$X_k = (t_{ij})$$

where $t_{ij} = \delta_{ik}$ with

$$\delta_{ik} = 1_{i=k}$$

Thus, if the observation in the ijth row got the kth treatment, the ijth row of X_k is 1. Otherwise, it is 0. Thus, X_k has exactly n_k ones and $n - n_k$ zeros. Clearly, X is not full rank since $J = X_1 + \dots + X_t$. However, the X_k's are orthogonal.

Let $Z = (X_1, \dots, X_t)$. Then Z is full rank because the X_k's are orthogonal. Moreover, $C(Z) = C(X)$ and hence $M = Z(Z'Z)^{-1}Z'$. Note that $Z'Z = \text{diag}(n_1, \dots, n_t)$. It is easy to show that

$$M = Z(Z'Z)^{-1}Z' = \text{blkdiag}(n_k^{-1}J_{n_k}^{n_k})_{k=1}^t$$

where J_k^k is the $k \times k$ matrix of ones.

Now, M is the orthogonal projection operator for the "full" model–that is M projects vectors onto $C(J, X_1, \dots, X_t)$. Suppose we wish to test for no treatment effect. The null hypothesis can be written as

$$H_0 = \alpha_1, \dots, \alpha_t$$

The model under the null hypothesis is given by

$$Y_{ij} = \mu + \epsilon_{ij}$$

The design matrix for this model is simply $X_0 = J$ and the orthogonal projection operator onto $C(J)$ is

$$M_\mu = J(J'J)^{-1}J' = \frac{1}{n}JJ' = \frac{1}{n}J_n^n$$

Thus, the orthogonal projection operator onto the treatment space M_α can be found by subtraction

$$M_\alpha = M - M_\mu$$

Note: while not expressly stated in the notes, this is because of the previous theory written about linear models. Namely, that of testing full and reduced subspaces.

We have the decomposition

$$M = M_\mu + M_\alpha = M_\mu + (M - M_\mu)$$

Moreover,

$$\begin{aligned} M_\mu M_\alpha &= M_\alpha M_\mu \\ &= (M - M_\mu)M_\mu \\ &= MM_\mu - M_\mu \\ &= M_\mu - M_\mu \\ &= 0 \end{aligned}$$

so $C(M_\alpha) \perp C(M_\mu)$. We also have $r(M_\mu) = 1$ and $r(M) = t$ so since they are orthogonal we have $r(M_\alpha) = r(M) - r(M_\mu = n - t$.

From the one-way ANOVA table, we see that ANOVA is decomposition of R^n into a sum of orthogonal subspaces. For one-way ANOVA, we have

$$R^n = C(M) = C(I - M) = C(M_\mu) + C(M_\alpha) + C(I - M)$$

we can also write

$$I = M_\mu + M_\alpha + (I - M)$$

8.a Estimation and Testing of Contrasts

Definition:
A contrast in a one-way ANOVA is a function $\sum_{i=1}^{t} \lambda_i \alpha_i$ where $\sum_{i=1}^{t} \lambda_i = 0$

Write $\lambda = (0, \lambda_1, \ldots, \lambda_t)'$. If $\lambda'\beta$ is estimable, then we know that $\lambda' = \rho' X$. The BLUE of $\lambda'\beta$ is $\rho' MY$. The vector $M\rho$ is always unique, and for one-way ANOVA, it has the form

$$M\rho = (t_{ij})$$

where $t_{ij} = \lambda_i / n_i$. Thus,

$$\rho' = \left(\frac{\lambda_1}{n_1} J'_{n_1}, \ldots, \frac{\lambda_t}{n_t} J'_{n_t} \right)$$

Contrasts of the α_i's

Theorem:
$\rho' X\beta$ is a contrast if and only if $\rho' J = 0$.

Proof:
Suppose $\rho' X\beta$ is a contrast. Then the elements of $\rho' X$ must sum to zero. That is, $\rho' XJ = 0$. Since the first column of X is J, this means that $\rho' JJ = 0$ which means $\rho' J = 0$.

UNKNOWN proof for other direction.

Contrasts are the estimable functions that impose a constraint on $C(M_\alpha)$. To se this, recall that $\rho' X\beta = 0$ implies that $E(Y) \in C(X) \cap C(M\rho)^\perp$. By definition, $\rho' X\beta$ imposes a constraint on $C(M_\alpha)$. if $M\rho \in C(M_\alpha)$. We are led to the following theorem.

Theorem:
$\rho' X\beta$ is a contrast if and only if $M\rho = C(M_\alpha)$.

Proof:
Since $J \in C(X)$, by the previous theorem, $\rho' X\beta$ is a contrast iff $0 = \rho' J = \rho' MJ$. Now, $C(M_\alpha)$ is the subspace of $C(X)$ that is orthogonal to J. Hence, $\rho' MJ = 0$ iff $M\rho \in C(M_\alpha)$.

We can now characterize $C(M_\alpha)$.

Theorem:

$$C(M_\alpha) = \left\{ \rho : \rho = (t_{ij}), t_{ij} = \lambda_i / n_i, \sum_{i=1}^{t} \lambda_i = 0 \right\}$$

This theorem says that the set of all possible contrasts makes up $C(M_\alpha)$.

Definition: (Orthogonal Contrasts)
Consider the one-way ANOVA model and two contrasts $\lambda_1'\beta$ and $\lambda_2'\beta$. Then $\lambda_1'\beta$ and $\lambda_2'\beta$ are said to be orthogonal contrasts if

$$\sum_{k=1}^{t} \frac{\lambda_{1k} \lambda 2k}{n_k} = 0$$

In a one-way ANOVA, the only way to break up the treatment subspace into a sum of $t-1$ orthogonal subspaces is ot find $t-1$ mutually orthogonal contrasts. Each contrast will have an orthogonal projection operator M_i and $M_\alpha = \sum_{i=1}^{t-1} M_i$.

Theorem (ρ you should use for contrast)
If we choose

$$\rho' = (\frac{\lambda_1}{n_1}J'_{n_1}, \ldots, \frac{\lambda_t}{n_t}J_{n_t};)$$

then $\rho \in C(X)$ so that $M\rho = M_\alpha \rho = \rho$

An implication of this theorem is that the orthogonal projection operator M_i for a contrast $\rho_i' X\beta$ is

$$M_i = \frac{\rho_i\rho_i}{'}\rho_i'\rho_i$$

9 Two-Way ANOVA

The model is

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk}$$

for $i = 1, \ldots a, j = 1, \ldots, b, k = 1, \ldots, N$. We can write the design matrix as

$$X = (J, X_1, \ldots, X_a, X_{a+1}, \ldots, X_{a+b})_{n \times a+b+1}$$

where $X_r = (t_{ijk})$ and where $t_{ijk} = \delta_{ir}, r = 1, \ldots, a$, $X_s = (t_{ijk}), t_{ijk} = \delta_{j(s-a)}, s = a+1, \ldots, a+b$. We define $\delta_{gh} = 1$ if $g = h$ and 0 otherwise.

We want to break up the estimation space into a sum of orthogonal subspaces, with each subspace corresponding to one of the effects. We want to write

$$C(M) = C(M_\mu) + C(M_\alpha) + C(M_\eta)$$

Define

$$Z = (J, Z_1, \ldots, Z_a, Z_{a+1}, \ldots Z_{a+b})$$

where

$$Z_r = X_r - \frac{X_r'J}{J'J}J, \quad r = 1, \ldots a+b$$

Thus, we have used Gram-Schmidt on the X's to eliminate the effect of μ. Note then that

$$Z_r = X_r - \frac{1}{a}J, \quad r = 1, \ldots, a$$
$$Z_s = X_s - \frac{1}{b}J, \quad s = a+1, \ldots, a+b$$

One can show that the Z_r's, Z_s's, and J are mutually orthogonal, and that the J's and Z's make up the $C(X)$. Thus, we have decomposed $C(X) = C(M)$ into three orthogonal subspaces

1. $C(M_\mu) = C(J)$
2. $C(M_\alpha) = C(Z_1, \ldots, Z_a)$
3. $C(M_\eta) = C(Z_{a+1}, \ldots, Z_{a+b})$

letting $Z_\alpha = (Z_1, \ldots, Z_a)$ and $Z_\eta = (Z_{a+1}, \ldots, Z_{a+b})$, we have

1. $M_\alpha = Z_\alpha(Z'_\alpha Z_\alpha)^- Z_\alpha$
2. $M_\eta = Z_\eta(Z'_\eta Z_\eta)^- Z'_\eta$
3. $M = M_\mu + M_\alpha + M_\eta$

9.a Contrasts for Two-way ANOVA

Estimation and testing of contrasts in two-way ANOVA is done in exactly the same way as in one-way ANOVA. This will follow from the fact that a contrast in the i's involves a constraint on $C(M) and C(M)$ does not involve the j's.

Theorem: (Contrasts in Two-Way ANOVA)
Let $\lambda'\beta$ be estimable and let $\lambda' = \rho' X$. Then $\lambda'\beta$ is a contrast in the α_i's if and only if $M\rho = M_\alpha \rho$.

In this case, the estimate of $\lambda'\beta$ is $\rho' M_\alpha Y$, which is the estimate from the one-way ANOVA ignoring the η_j's.

Suppose we have a contrast $\sum_{i=1}^a c_i \alpha_i$, where $\sum_{i=1}^a c_i = 0$. Consider writing this contrast in the form $\rho' X \beta$ where

$$\rho' = (J'_{bN} c_1, \ldots, J'_{bN} c_a)$$

then

1. $M\rho = M_\alpha \rho = \rho$ since $\rho \in C(M_\alpha)$
2. The estimate of the contrast is $\rho' M Y = \rho' M_\alpha Y = \rho' Y = \sum_{i=1}^a c_i \bar{Y}_{i..}$
3. $\mathrm{Var}(\rho' M Y) = \sigma^2 \rho' M \rho = \sigma^2 \rho' \rho = \dfrac{\sigma^2}{bN} \sum_{i=1}^a c_i^2$

9.b Balanced Two-way ANOVA with Interaction

This is very similar to the above, but we note that the sum of the interaction columns gives J and

$$X_r = \sum_{j=1}^{b} X_{r,j} \quad r = 1, \ldots, a$$

$$X_s = \sum_{i=1}^{a} X_{i,s-a} \quad s = a+1, \ldots, a+b$$

so that the columns of the interaction actually make up the entire $C(X)$.

It turns out we can obtain M_α and M_η from the two-way ANOVA and get the interaction M_γ by subtraction since the spaces are orthogonal, and since the interaction columns form $C(X)$ we have

$$\begin{aligned} M &= Z_\gamma(Z_\gamma' Z_\gamma)^{-1} Z_\gamma' \\ &= \text{blkdiag}(N^{-1} J_N^N) \end{aligned}$$

The expected values of $Y' M_\alpha Y$ and $Y' M_\eta Y$ are <u>different</u> from those found in the main effects (no interaction) model. It can be shown that

$$E(Y' M_\alpha Y) = \sigma^2(a-1) + bN \sum_{i=1}^{a} (\alpha_i + \bar{\gamma}_{i.} - \bar{\alpha}_{.} - \bar{\gamma}..)^2$$

which depends on the γ_{ij}'s. For the no interaction model, we had

$$E(Y' M_\alpha Y) = \sigma^2(a-1) + bN \sum_{i=1}^{a} (\alpha_i - \bar{\alpha}_{.})^2$$

this difference implies that the F test for testing no α treatment effect is NOT a test that the α_i's are all equal. Rather, it is a test of the hypothesis

$$H_0 : \alpha_1 + \bar{\gamma}_{1.} = \ldots, = \alpha_a + \bar{\gamma}_{a.}$$

Thus, in the two-way ANOVA model with interaction the hypothesis of no α treatment effect is a test of the hypothesis that all the $\alpha_i + \bar{\gamma}_{i.}$'s are equal. The F test for this hypothesis is

$$F = \frac{||M_\alpha Y||^2 / r(M_\alpha)}{MSE}$$

Since the interaction space makes up $C(X)$, ALL estimable functions of the parameters are functions of the γ_{ij}'s. Thus, ANY estimable function MUST involve the γ_{ij}'s in some way. To see this, note that if $\lambda'\beta$ is not a function of the γ_{ij}'s but is estimable, then $\rho' X_i = 0$ for $i = a+b+1, \ldots, a+b+ab$. This imples $\rho' X = 0$, since $C(X) = C(X_{a+b+b}, \ldots, X_{a+b+ab}))$. This would imply that the estimable function is identically equal to 0.

9.b.1 Contrasts for Two-way ANOVA with Interaction

To examine contrasts in the α space, we need to examine the nature of

$$\lambda'\beta = \rho' X\beta = \rho' M X\beta = \rho' M_\alpha X\beta$$

$M_\alpha X\beta$ is an $n \times 1$ vector whose elements are of the form

$$\alpha_i + \gamma_{i.} = \bar{\alpha}_{.} - \bar{\gamma}_{..}$$

Now, $\rho' M_\alpha X\beta$ will be a contrast in these terms, or equivalently, a contrast in $\alpha_i + \bar{\gamma}_{i.}$ Thus, a contrast in terms of $\alpha_i + \bar{\gamma})i.$ is a contrast in the α space.

Contrasts in the Interaction Space To find the set of contrasts in the interaction space, we need to know how to put a constraint on $M_\gamma = M - M_\mu - M_\alpha - M_\eta$. The hypothesis $H_0 : \rho' X\beta = 0$ puts a constraint on the interaction space if and only if $M\rho = M_\gamma \rho$. Thus, $\rho' X\beta = 0$ implies

$$\rho \in C(M_\mu + M_\alpha + M_\eta)^\perp \implies \rho' X_i = 0 \text{ for } i = 0, \ldots, a+b$$

where $X_0 = J$.

Definition (Kronecker Product) Suppose A is an $r \times c$ matrix and B is an $s \times d$ matrix. The Kronecker Product of A and B, written $A \otimes B$ is an $rs \times cd$ matrix of the form

$$A \otimes B = (a_{ij} B)$$

Theorem: (Construction of Contrasts in Interaction Space) Let $d = (d_1, \ldots, d_a)_{1\times a}$ denote a set of contrast coefficients for the α space and let $c = (c_1, \ldots, c_b)_{1\times b}$ denote a set of contrast coefficients for the η space. We have $\sum_{i=1}^a d_i = 0$ and $\sum_{i=1}^b c_i = 0$. Then a set of contrast coefficients for the interactions pace are of the form $d \otimes c$.

Thus, a corresponding ρ vector for a contrast in the interaction space is the form

$$\rho' = \frac{1}{N}(d \otimes c) \otimes J_N'$$

Note that this theorem does not characterize the set of all possible contrasts in the interaction space–it just tells us how to construct a contrast in the interaction space.

Theorem: (Characterization of Contrasts in Interaction Space)
Suppose Q is an $a \times b$ matrix such that $J_a' Q = 0$ and $QJ_b = 0$. Thus, Q is a matrix with each row and column summing to 0. Then ALL contrasts in the interaction space have ρ of the form

$$\rho' = (\rho_{ijk})$$

where $\rho_{ijk} = q_{ij}/N$ and where $\bar{q}_{i.} = 0$ and $\bar{q}_{.j} = 0$. That is

$$\rho' = \left(q_1 1 \otimes \frac{J_N'}{N}, \ldots, q_{ab} \otimes \frac{J_N'}{N} \right)$$

This class of ρ's satisfy

1. $\rho \in C(X)$, and
2. $\rho' X_i = 0$ for $i = 0, \ldots, a+b$.

10 The Exponential Family

Definition: (Exponential Family)
If the density of a random variable Y with respect to a σ-finite measure $\Lambda(y)$ has the form

$$p(y|\xi) = \exp\left\{\phi[y\theta - b(\theta) - c(y)] - \frac{1}{2}s(y,\phi)\right\}$$

where $\xi = (\theta, \phi)$ $c(.), b(.)$ and $s(.,.)$ are some known functions, then Y or its distribution belongs to an exponential family, denoted by $Y \sim D(\theta, \phi)$. In addition, θ is the natural parameter of the exponential family and ϕ is the dispersion family.

Theorem: (Cumulant Generating Function of Exponential Families)
For the exponential family,

$$K_Y(t) = \phi[b(\theta + t/\phi) - b(\theta)]$$

Proof:
Simple. Find the MGF by using the pdf integrates to 1 trick. Then take the log.

10.a Multivariate Exponential Family

Many multivariate distributions belong to exponential families, because, in general, exponential families can be defined for multi-dimensional θ and $\mathbf{y}$. A multivariate exponential family (MEF) is defined as

$$p(\mathbf{y}|\xi) = \exp\left\{Q(\mathbf{y})^T T(\theta) - b(\theta) - c(\mathbf{y})\right\}$$

where $\xi = \theta$, $T(\theta) = (t_1(\theta), \ldots t_k(\theta))^T$, and $Q(\mathbf{y}) = (q_1(\mathbf{y}), \ldots, q_k(\mathbf{y}))^T$. If all components of $T(\theta)$ and all components of $Q(\mathbf{y})$ are linearly independent, then the exponential family is said to be full rank.

11 Likelihood Theory

11.a Bartlett Identities

Let $p(\mathbf{u}, \xi)$ be the joint density function of $\mathbf{U} = (U_1, \ldots, U_n)$, and therefore

$$\int p(\mathbf{u}, xi) d\Lambda(\mathbf{u}) = 1$$

where ξ contains all the unknown parameters. By differentiating with respect to $\xi \in \Xi$, a subset of R^q, we are led to the following theorem.

Theorem: (Bartlett Identities)
Suppose that differentiation and integration are exchangeable and all the necessary expectations are finite. We have the following results.

1. $E_\xi(\partial_{\xi_j}\ell_n) = 0$
2. $E_\xi(\partial^2_{\xi_j,\xi_k}\ell_n) + E_\xi(\partial_{\xi_j}\ell_n\partial_{\xi_k}\ell_n) = 0$

11.b Maximum Likelihood Estimation

In this section, we will assume that $U_1, \ldots, U_n$ are independent random variables with probability density function $p(u;\xi)$. The log-likelihood function is then given by

$$\ell_n(\xi) = \sum_{i=1}^{n} \log p(u_i, \xi)$$

Definition (Maximum Likelihood Estimate)
The maximum likelihood estimate of ξ is defined as

$$\hat{\xi} = \arg\max_{\xi \in \Xi} \ell_n(\xi)$$

For a smooth $\ell_n(\xi)$, a necessary condition for the existence of $\hat{\xi}$ is that $\partial_\xi \ell_n(\hat{\xi}) = \mathbf{0}$ and $-\partial^2_\xi \ell_n(\hat{\xi})$ is positive definite.

11.b.1 Newton-Raphson

Although the score equation may not have an explicit closed form, the Newton-Raphson algorithm can be used to numerically calculate $\hat{\xi}$ by using the iterative scheme

$$\xi^{t+1} = \xi^t + \left\{\partial^2_\xi \ell_n(\xi^t)\right\}^{-1} \partial_\xi \ell_n(\xi^t)$$

The key idea behind the Newton-Raphson algorithm is a first-order Taylor series expansion of the score function at a trial value ξ^t.

11.c Estimation Theory

Theorem:
Let $U_1, \ldots, U_n$ be *i.i.d.* with pdf $p(U_i, \xi_*)$. Suppose that the following conditions hold

1. $\sum_{i=1}^{n} \partial_\xi \log p(U_i, \hat{\xi}) = \mathbf{0}$.
2. $\hat{\xi}$ is consistent for ξ_*
3. For each ξ in an open subset of Euclidean space, $\partial_{xi} \log p(U, \xi)$ is twice continuously differentiable for every x and the norm for the third derivative is at most $f(U)$, where $f(U)$ is a fixed integrable function.

4. $E[\partial_\xi \log p(U, \xi_*)] = 0$, $E||\partial_\xi||^2 < \infty$, and the matrix $E[\partial_\xi^2 \log p(U_i, \xi_*)]$ exists and is nonsingular.

Then

$$\sqrt{n}(\hat{\xi} - \xi_*) = [-E\partial_\xi^2 \log(p(U, \xi_*))]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_\xi \log p(U_i, \xi_*) + o_p(1)$$

In particular,

$$\sqrt{n}(\hat{\xi} - \xi_*) \to_d N(0, -E[\partial_\xi^2 \log p(U, \xi_*)]^{-1})$$

11.d Hypothesis Testing

We consider testing linear or non-linear hypotheses of the form

$$H_0 : h_0(\xi) = b_0 \text{ vs. } h_0(\xi) \neq b_0$$

We will also consider testing a linear hypothesis

$$H - 0 : R\xi = b_0 \text{ vs. } R\xi \neq b_0$$

where R is an $r \times q$ matrix of full row rank and b_0 is an $r \times 1$ specified vector. WLOG, we assume that $R\xi = \xi(1)$ where $R = (I_r, 0), \xi^T = \left(\xi(1)^T, \xi(2)^T\right)^T$ and $\xi(1)$ and $\xi(2)$ are $r \times 1$ and $(q - r) \times 1$ subvectors of ξ. Thus, the linear hypothesis can be re-written as

$$H_0 : \xi(1) = b_0 \text{ vs. } H_1 : \xi(1) \neq b_0$$

Write the maximum likelihood estimate of ξ as $\hat{\xi} = (\hat{\xi}(1)^T, \hat{\xi}(2)^T)^T$ Under H_0, the constrained ML estimate of ξ is denoted by $\tilde{\xi} = (b_0^T, \tilde{\xi}(2)^T)^T$. We define $\hat{\xi}(2)[\xi(1)]$ as a function of $\xi(1)$ which maximizes the log likelihood for each $\xi(1)$. Thus, $\tilde{\xi}(2) = \hat{\xi}(2)(b_0)$

11.d.1 Wald Test

The Wald Test is defined by

$$W_n = [h_0(\hat{\xi}) - b_0]^T \left\{ H(\hat{\xi}) E[-\partial_\xi^2 \ell_n(\hat{\xi})]^{-1} H(\hat{\xi})^T \right\}^{-1} [h_0(\hat{\xi}) - b_0]$$

where $H(\xi) = \frac{\partial h_0(\xi)}{\partial \xi}$ is an $r \times q$ matrix. In practice, we may replace $E[\partial_\xi^2 \ell_n(\hat{\xi})]$ by $\partial_\xi^2 \ell_n(\hat{\xi})$ Since the average difference of these two converges in probability to 0.

For a linear test, the Wald test statistic is

$$\begin{aligned} W_n &= (\hat{\xi}(1) - b_0)^T [\text{Cov}(\hat{\xi}(1))]^{-1} (\hat{\xi}(1) - b_0)^T \\ &= (\hat{\xi}(1) - b_0)^T [R I_n(\hat{\xi})^{-1} R^T]^{-1} (\hat{\xi}(1) - b_0)^T \end{aligned}$$

where $I_n(\hat{\xi}) = E[-\partial_\xi^2 \ell_n(\hat{\xi})]$

We can estimate $\text{Cov}(\hat{\xi}(1))$ using $R\text{Cov}(\hat{\xi})R^T = -L^{11}$, where L^{11} is the first $r \times r$ submatrix of the inverse Hessian matrix of the log likelihood.

11.d.2 Score Test

Rao's score test is defined as

$$\mathrm{SC}_n = \partial_\xi \ell_n(\tilde{\xi})^T \left\{ E[\partial_\xi \ell_n(\xi)^{\otimes 2}] \right\}^{-1}|_{\xi=\hat{\xi}} \partial_\xi \ell_n(\tilde{\xi})$$

for which we can replace $E[\partial_\xi \ell_n(\tilde{\xi})]$ by $E[-\partial_\xi^2 \ell_n(\tilde{\xi})]$ or just $-\partial_\xi^2 \ell_n(\tilde{\xi})$

For a linear test, the Score test statistic is

$$\mathrm{SC}_n = -\dot{L}_1(\tilde{\xi})^T L^{11}(\tilde{\xi}) \dot{L}_1(\tilde{\xi})$$

where $\dot{L}_1(\xi)$ is the first r components of $\partial_\xi \ell_n(\xi)$ and L^{11} is the upper $r \times r$ submatrix of the <u>inverse</u> second derivative of the log likelihood.

An advantage of the Score test is that it avoids the calculation of an estimator under the alternative hypothesis.

To summarize, below are the steps to calculate the score test:

1. Compute $\tilde{\xi} = (b_0, \tilde{\xi}(2))$, where $\xi(2)$ is the last $p - r$ components of ξ. Specifically, $\tilde{\xi}(2)$ maximizes $\ell_n(b_0, \xi(2))$ when $\xi(1) = b_0$.
2. Compute $\partial_\xi \ell_n(\tilde{\xi})$
3. Compute $I_n(\tilde{\xi}) = E[-\partial_\xi^2 \ell_n(\tilde{\xi})]$
4. Compute $SC_n = \partial_\xi \ell_n(\tilde{\xi})^T I_n(\tilde{\xi})^{-1} \partial_\xi \ell_n(\tilde{\xi})$

11.d.3 Likelihood Ratio Test

The likelihood ratio test is defined as

$$LRT_n = 2[\ell_n(\hat{\xi}) - \ell_n(\tilde{\xi})]$$

where $\tilde{\xi}$ is the maximum likelihood estimate under the restriction $H_0 : h(\xi) = b_0$.

All of the tests are asymptotically $\chi^2(r)$ under the null hypothesis. An <u>asymptotically valid test</u> can be obtained by comparing the sample values of the test statistic W_n with the critical value of the right-hand tail of a $\chi^2(r)$ distribution at a pre-specified significance level α. We reject H_0 when the test statistic is larger than $\chi^2_\alpha(r)$, the upper α-percentile of a $\chi^2(r)$ distribution.

11.e The Delta Method

Suppose that $g(\xi_*)$ is a parameter of interest. A natural estimator for $g(\xi_*)$ is $g(\hat{\xi})$.

Theorem 2.6

1. If $g : R^q \to R^{q'}$ is continuous at ξ_* and $\hat{\xi}$ converges to ξ_* in probability, then $g(\hat{\xi})$ converges to $g(\xi_*)$ in probability.
2. If g is differentiable at ξ_* and $\sqrt{n}(\hat{\xi} - \xi_*) \to_d N(0, I(\xi_*)^{-1})$, then $\sqrt{n}(g(\hat{\xi}) - g(\xi_*)) \to_d N(0, \partial_\xi g(\xi_*) I(\xi_*)^{-1} g(\xi_*)^T)$

12 Generalized Linear Models (Ch. 3)

Definition: (Generalized Linear Models)
GLMs are defined as follows:

- The components of $\mathbf{y} = (y_1, \ldots, y_n)^T$ are mutually independent, and the conditional density of y_i given $\mathbf{x}_i$ is $D(\theta_i, \phi/\omega_i)$, which is a member of the exponential family, where θ_i is a function of $\mathbf{x}_i$ and ω_i is a weight.
- μ_i is related to x_i by
$$g(\mu_i) = \mathbf{x}_i^T \beta = \eta_i$$
for $i = 1, \ldots, n$ where $\beta = (\beta_1, \ldots \beta_p)^T$ is defined in a subset $\mathcal{B}$ of R^p and $g(\cdot)$ is a known monotonic link function.

Since $\mu_i = \dot{b}(\theta_i)$ and $g(\mu_i) = g(\dot{b}(\theta_i)) = \xi^T \beta$, we have

$$\begin{aligned} \mu_i &= g^{-1}(x_i^T \beta) \\ \theta_i &= \dot{b}^{-1} \circ g^{-1}(x_i^T \beta) = k(x_i^T \beta) \end{aligned}$$

If $\theta_i = \eta_i = x_i^T \beta$, then $\dot{b}^{-1} \circ g^{-1} = k$ is an identity function.

When $g(\cdot) = \dot{b}^{-1}(\cdot)$, we call it a canonical link function.

12.a Estimation Methods

Lemma 3.1 (GLM Score and Information in Matrix Form) For the GLM, the score function and the Fisher Information of β can be written as

$$\begin{aligned} \dot{\ell}_n(\beta) &= \phi D_\theta(\beta)^T e(\beta) = \phi D(\beta)^T V(\beta)^{-1} e(\beta) \\ E[-\ddot{\ell}_n] &= \phi D_\theta(\beta)^T V D_\theta(\beta) = \phi D(\beta)^T V^{-1} D(\beta) \end{aligned}$$

where

$$\begin{aligned} V(\beta) &= \text{diag}(v_1(\beta), \ldots, v_n(\beta)) \\ e(\beta) &= (y_1 - \mu_1(\beta), \ldots, y_n - \mu_n(\beta)) \\ D_\theta(\beta)^T &= (\partial_\beta \theta_1(\beta), \ldots, \partial_\beta \theta_n(\beta))_{p \times n} \\ D(\beta)^T &= (\partial_\beta \mu_1(\beta), \ldots, \partial_\beta \mu_n(\beta))_{p \times n} \end{aligned}$$

where v_i is the variance function

$$v_i = \ddot{b}(\theta(\beta)) = \ddot{b}(\dot{b}^{-1}(\mu(\beta))) = \ddot{b}(k(x_i^t\beta))$$

where k is the theta-link function, $k = \dot{b}^{-1} \circ g^{-1}$

12.a.1 Newton-Raphson

The Newton-Raphson algorithm for obtaining the ML estimate $\hat{\beta}$ in a GLM is given by

$$\begin{aligned}\beta^{k+1} &= \beta^k + \Big\{E[-\ddot{\ell}_n(\beta^k)]\Big\}^{-1} \dot{\ell}_n(\beta^k)\\ &= \beta^k + \big\{(D^TV^{-1}D)^{-1}D^TV^{-1}e\big\}_{\beta_k}\end{aligned}$$

12.b Likelihood Theory

Consistency
$\hat{\xi} = (\hat{\beta}, \hat{\phi}) \to \xi_*$ the true value in probability or almost surely.

Asymptotic Normality

$$\sqrt{n}(\hat{\xi} - \xi_*) \to_d N(\mathbf{0}, I(\xi_*)^{-1})$$

as $n \to \infty$, where $I(\xi_*) = \lim_{n\to\infty} n^{-1}I_n(\xi_*)$ denotes the limit of the average Fisher information matrix based on $y_1, \ldots, y_n$ at ξ_*, and we have

$$I_n(\xi_*) \approx \begin{pmatrix} \phi D_\theta(\hat{\beta})^T V(\hat{\beta}) D_\theta(\hat{\beta}) & \mathbf{0} \\ \mathbf{0} & 0.5\sum_{i=1}^n \ddot{s}(y_i, \hat{\phi}) \end{pmatrix}$$

12.b.1 Hypothesis Tests

We consider the following hypotheses:

$$H_0 : \beta_1 = b_0 \text{ vs. } H_1 : \beta_1 \neq b_0$$

where β_1 is the first $r \times 1$ subvector of β.

Wald Test:
The Wald test statistic is given by

$$W_n = (\hat{\beta}_1 - b_0)^T[RI_n(\hat{\xi})R^T]^{-1}(\hat{\beta}_1 - b_0)$$

where $R = [\mathbf{I}_r, \mathbf{0}]$

Score Test:

1. Compute $\hat{\xi} = (b_0, \tilde{\beta}_2, \tilde{\phi})$ where β_2 is the last $p-r$ subvector of β and $\tilde{\beta}_2$ is the maximizer of the log likelihood when $\beta_1 = b_0$.
2. Calculate $\partial_\xi \ell_n(\tilde{\xi}) = \begin{pmatrix} \tilde{\phi} D_\theta(\tilde{\beta})^T e(\tilde{\beta}) \\ 0 \end{pmatrix}$
3. Compute $I_n(\tilde{\xi}) = \begin{pmatrix} \tilde{\phi} D_\theta(\tilde{\beta}) V(\tilde{\beta}) D_\theta(\tilde{\beta}) & 0 \\ 0 & * \end{pmatrix}$
4. From the theory of the Score statistic, it follows that

$$SC_n = \tilde{\phi}^2 e(\tilde{\beta})^T D_\theta(\tilde{\beta}) I_n(\tilde{\xi})^{-1} D_\theta(\tilde{\beta})^T e(\tilde{\beta})$$

12.c Deviance

The deviance for all n observations is defined as

$$Dv(\mathbf{y}; \hat{\mu}) = \phi^{-1} LRT_n = \sum_{i=1}^{n} Dv_i$$

where $Dv_i = 2\{y_i q(y_i) - b(q(y_i))\} - 2\{y_i q(\mu_i(\hat{\beta})) - b(q(\mu_i(\hat{\beta}))\}$

Essentially, the deviance is a scaled LRT statistic where we substitue in y for the full model for μ and $\hat{\mu} = \mu(\hat{\beta})$ as the reduced model.

Another important application of the deviance is to select an 'optimal' model from a sequence of nested models. Consider two nested models defined by $\mu = \mu_A(\beta)$ and $\mu = \mu_B(\beta)$, respectively. For model $\mu = \mu_A(\beta)$, the deviance equals $Dv(y; \hat{\mu}_A)$, while the deviance equals $Dv(y; \hat{\mu}_B)$ for model $\mu = \mu_B(\beta)$. Therefore, if $A \subset B$, then

$$Dv(\mathbf{y}; \hat{\mu}_A) - Dv(y; \hat{\mu}_B) = 2\phi^{-1}\{\ell_n(\mu_B(\hat{\beta}), y) - \ell_n(\mu_A(\hat{\beta}), y\}$$

which is close to the likelihood ratio test statistic except for the factor ϕ.

If ϕ is unknown, we cannot interpret the deviance difference between the two models as a scaled likelihood ratio statistic. Instead, we use the estimate from model B $\hat{\phi}$.

13 Eliminating Nuisance Parameters (Ch. 7)

Statistical models involves a set of parameters $\xi = (\psi, \lambda)$ in which ψ is the parameter of interest and λ is a nuisance parameter. Specific methods are needed to eliminate nuisance parameters.

13.a Hypergeometric Distributions

For a fixed sample size n, the joint distribution of the cell counts in a 2×2 contingency table is

$$\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\pi_{11}^{n_{11}}\pi_{12}^{n_{12}}\pi_{21}^{n_{21}}\pi_{22}^{n_{22}}$$

where $\pi_{ij} = P(X = i-1, Y = j-1)$

Since $n_{12} = n_{1.} - n_{11}$ and $n_{22} = n - n_{11} - n_{12} - n_{21}$, we have

$$\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}^{n_{11}}\pi_{12}^{n_{1.}}\pi_{21}^{n_{.1}}\pi_{22}^{n-n_{1.}-n_{.1}}$$

From this, we see that $(n_{1.}, n_{.1}, n)$ is sufficient for $\lambda = (\pi_{12}, \pi_{21})$. To eliminate the nuisance parameters, we can calculate the conditional distribution of n_{11} given $(n_{.1}, n_{1.}, n)$

Note that $n_{.1} = n_{11} + n_{21} \geq n_{11}$ and $n_{1.} \geq n_{11}$, so $n_{11} \leq \min\{n_{.1}, n_{1.}\}$. Moreover,

$$n_{22} = n - n_{11} - (n_{1.} - n_{11}) - (n_{.1} - n_{11}) = n - n_{1.} - n_{.1} + n_{11} \geq 0$$

so $n_{11} \geq \max\{0, n_{1.} + n_{.1} - n\}$

Thus, the marginal distribution of $(n_{1.}, n_{.1}|n)$ is

$$p(n_{1.}, n_{.1} \mid n) = \sum_{n_{11}=a}^{b} \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}^{n_{11}}\pi_{12}^{n_{1.}}\pi_{21}^{n_{.1}}\pi_{22}^{n-n_{1.}-n_{.1}}$$

where $a = \min\{n_{1.}, n_{.1}\}$ and $b = \max 0, n_{1.} + n_{.1} - n$

This leads to the non-central hypergeometric distribution

$$p(n_{11} \mid n_{1.}, n_{.1}, n) = \frac{p(n_{11}, n_{1.}, n_{.1} \mid n)}{p(n_{1.}, n_{.1} \mid n)} = \frac{\binom{n_{1.}}{n_{11}}\binom{n-n_{1.}}{n_{.1}-n_{11}}\psi^{n_{11}}}{P_0(\psi)}$$

where $P_0(\psi) = \sum_{x=a}^{b}\binom{n_{1.}}{x}\binom{n-n_{1.}}{n_{.1}-x}\psi^x$

We can write the non-central hypergeometric distribution as an exponential family with canonical parameter $\theta = \log\psi$, $\phi = 1$, and $b(\theta) = \log P_0(\psi) = \log P_0(\exp(\theta))$ Using properties of exponential families, the cumulant generating function is given by

$$\begin{aligned}
K(t) &= b(\theta + t/\phi) - b(\theta) \\
&= b(\theta + t) - b(\theta) \\
&= \log P_0(e^{\theta+t}) - \log P_0(e^{\theta}) \\
&= \log P_0(e^{\theta}e^{t}) - \log\big(P_0(e^{\theta})\big) \\
&= \log P_0(\psi e^{t}) - \log(P_0(\psi))
\end{aligned}$$

We can find the mean and variance of the non-central hypergeometric random variable by taking the first two derivatives at $t = 0$. Note we must plug in $\psi^* = \psi e^t$ into P_0 when taking the derivatives. Doing so, we get

$$\mu = \frac{P_1(\psi)}{P_0(\psi)}$$
$$\sigma^2 = \frac{P_2(\psi)}{P_0(\psi)} - \mu^2$$

where

$$P_j(\psi) = E(X^j) = \sum_{x=a}^{b} \binom{(}{n}_{1.})x \binom{n - n_{1.}}{n_{.1} - x} \psi^x x^j$$

By the invariance property of the MLE, we have that the conditional maximum likelihood estimate of ψ, denoted $\hat{\psi}_C$, is the solution to

$$n_{11} = \frac{P_1(\hat{\psi}_C)}{P_0(\hat{\psi}_C)}$$

The variance of $\hat{\psi}_C$ can be approximated by the inverse of the Fisher information matrix.

13.b Conditional Logistic Regression

In the traditional logistic regression setup, let α denote the intercept parameter and $\beta = (\beta_1, \ldots, \beta_p)^T$. The likelihood is given by

$$L(\alpha, \beta | Y_1 = y_1, \ldots Y_n = y_n) = \frac{\exp\Big[(\sum_i Y_i)\alpha + \sum_{j=1}^{p}(\sum_{i=1}^{n} y_i x_i j)\beta_j\Big]}{\prod_i [1 + \exp\Big(\alpha + \sum_{j=1}^{p} \beta_j x_{ij}\Big)]}$$

It followos that $s_j := \sum_i y_i x_{ij}$ is sufficent for β_j and $s_0 = \sum_i y_i$ is sufficient for α. We can eliminate α by conditioning on $\sum_{i=1}^{n} y_i$ to obtain

$$P(y_1, \ldots, y_n | s_0) = \frac{\exp\Big[\sum_{j=1}^{p}(\sum_i y_i x_{ij})\beta_j\Big]}{\sum_{S(s_0)} \exp\Big[\sum_{j=1}^{p}(\sum_i y_i x_{ij})\beta_j\Big]}$$

where $S(s_0)$ denotes the conditional reference set of samples having the same value of s_0. Finally, we get

$$P(y_1, \ldots, y_n, s_0 | \xi) = P(y_1, \ldots, y_n | s_0, \beta_1, \ldots \beta_p) P(s_0 | \xi)$$

If we wish to focus on β_p in the logistic regression model, we must eliminate the other parameters $(\alpha, \beta_1, \ldots, \beta_{p-1})$ by conditioning on their sufficient statistics $s_0 = \sum_i y_i$ and $s_j = \sum_i y_i x_{ij}$ for $j = 1, \ldots, p-1$. We obtain

$$P(y_1, \ldots, y_n | s_j, j = 0, \ldots, p-1) = \frac{\exp(s_p \beta_p)}{\sum_{s_p^* \in S(s_0, \ldots s_{p-1})} \exp\big(s_p^* - \beta_p\big)}$$

where

$$S(s_0,\ldots,s_{p-1}) = \left\{ (y_1^*,\ldots,y_n^*) : \sum_i y_i^* = s_0, \sum_i y_i^* x_{ij} = s_j, j = 1,\ldots,p-1 \right\}$$

and so we have

$$P(y_1,\ldots y_n, s_0|\xi) = P(y_1,\ldots,y_n|s_0,\ldots,s_{p-1},\beta_p)P(s_0,\ldots,s_{p-1}|\xi)$$

Conditional inference about β_p can be carried out using $P(y_1,\ldots,y_n|s_0,\ldots,s_{p-1},\beta_p)$ or $P(s_p = t|s_0,\ldots,s_{p-1},\beta_p)$. Let $c(s,t)$ be the number of data vectors in $S(s_0,\ldots,s_{p-1})$ for which $s_p = t$. Then

$$P(s_p = t|s_0,\ldots,s_{p-1},\beta_p) = \frac{c(\mathbf{s},t)\exp(t\beta_p)}{\sum_u c(\mathbf{s},u)\exp(u\beta_p)}$$

The CMLE of β_p, denoted by $\hat{\beta}_{p,c}$, satisfies the score function

$$s_{p,obs} = \frac{\sum_u c(s,u)u\exp\left(u\hat{\beta}_{p,c}\right)}{\sum_u c(s,u)\exp\left(u\hat{\beta}_{p,c}\right)}$$

The covariance matrix of $\hat{\beta}_c$ can be approximated by

$$\partial^2_{\beta_p}\left(\log\left[\sum_u c(\mathbf{s},u)\exp\left(u\hat{\beta}_{p,c}\right)\right]\right)$$

13.c Binary Matched Pairs

Let (Y_{i1}, Y_{i2}) be the ith pair of observations, $i = 1,\ldots,n$ and $g[p(Y_{it} = 1)] = \alpha_i + \beta x_t$ for $i = 1,\ldots,n$ and $t = 1, 2$ where $g(\cdot)$ is the link function. For the logit link, we have

$$P(Y_{i1} = 1) = \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}, \quad P(Y_{i2} = 1) = \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}$$

The number of parameters in this model is greater than n. We can use conditional likelihood to eliminate the α_i's. If we assume independence of responses for different subjects and for the two observations on the same subject (the latter being a strong, unrealistic assumption), the likelihood is given by

$$\prod_{i=1}^n \frac{\exp(\alpha_i y_{i1})}{1+\exp(\alpha_i)} \frac{\exp((\alpha_i+\beta)y_{i2})}{1+\exp(\alpha_i+\beta)}$$

and $s_i = y_{i1} + y_{i2}$ is the sufficient statistic for $\alpha_i, i = 1,\ldots,n$. To estimate β, we eliminate the α_i's by conditioning on all the s_i's. Note that $P(Y_{i1} = Y_{i2} = 0|s_i = 0) = P(Y_{i1} = Y_{i2} = 1|s_i = 2) = 1$. and

$$P(Y_{i1}, Y_{i2}|s_i = 1) = \begin{cases} \exp(\beta)/[1+\exp(\beta)], & y_{i1} = 0, y_{i2} = 1 \\ 1/[1+\exp(\beta)], & y_{i1} = 1, y_{i2} = 0 \end{cases}$$

The conditional likelihood is then given by

$$\prod_{s_i=1} \left(\frac{1}{1+\exp(\beta)}\right)^{y_{i1}} \left(\frac{\exp(\beta)}{1+\exp(\beta)}\right)^{y_{i2}} = \frac{\exp(\beta n_{21})}{[1+\exp(\beta)]^{n^*}}$$

where $n_{21} = \#\{i : y_{i1} = 0, y_{i2} = 1\}$ and $n^* = \#\{i : y_{i1} + y_{i2} = 1\}$

$\hat{\beta} = \log(n_{21}/(n^* - n_{21}))$ and its standard error is given by $\text{se}(\hat{\beta}) = \sqrt{1/n_{21} + 1/n_{12}}$ It can be shown that

$$P(Y_{i1} = 0, Y_{i2} = 1|S_i = 1) = \exp\big(x_{i2}^T\beta\big)/[\exp\big(x_{i1}^T\beta\big) + \exp\big(x_{i2}^T\beta\big)]$$

and

$$P(Y_{i1} = 1, Y_{i2} = 0|S_i = 1) = \exp\big(x_{i1}^T\beta\big)/[\exp\big(x_{i1}^T\beta\big) + \exp\big(x_{i2}^T\beta\big)]$$

which combines to form

$$P(Y_{i1} = 0, Y_{i2} = 1|S_i = 1) = \exp\big((x_{i2} - x_{i1})^T\beta\big)/[(\exp(x_{i2} - x_{i1})^T\beta) + 1]$$

which is the form of a logistic regression with no intercept and with predictor values $x_i^* = x_{i2} - x_{i1}$. Thus, wee can obtain conditional ML estimates by fitting a logistic regression model to the artificial response $y_i^* = 1$ when $(y_{i1} = 0, y_{i2} = 1)$, $y^* = 0$ when $(y_{i1} = 1, y_{i2} = 0)$, no intercept, and x_i^*.

13.d Conditional Likelihood

If the distribution of $\mathbf{s}$ does not contain any information about ψ, $\mathbf{s}$ is said to be ancillary for ψ in the presenc
A statistic $\mathbf{s}$ is said to be S-ancillary for ψ in the presence of λ if the family of $\{p(\mathbf{s};, \psi, \lambda) : \lambda \in \Lambda\}$ is the same for each ψ.

A statistic is said to be P-ancillary for ψ in the presence of λ if the partial information for ψ based on the
That is

$$I_\psi(\xi; \mathbf{s}) = I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi} = 0$$

where

$$I(\psi, \lambda) = \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix}$$

and $I_{\psi\psi}$, $I_{\psi,\lambda}$, and $I_{\lambda\lambda}$ are the appropriate elements of the Fisher information matrix of (ψ, λ) based on $\mathbf{s}$.

Note that $I_\psi(\xi, \mathbf{s}) := I(\psi|\lambda)$ is the Fisher information for ψ given λ is known. It takes the same form as the conditional covariance in a MVN. $I(\psi|\lambda)^{-1}$ is the Cramer-Rao lower bound for the asymptotic covariance matrix of unbiased estimators of ψ when λ is known, and $I(\psi|\lambda)^{-1} = (I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi})^{-1}$.

Consider a conditional score statistic $U_\psi(\xi) = \left.\frac{\partial \ell_c(\xi, \psi_0)}{\partial \psi}\right|_{\psi_0 = \psi}$. Then

$$U_\psi(\xi) = \partial_\psi \log p(Y|\xi) - E[\partial_\psi \log p(Y|\xi)|s_\lambda(\psi_0)] \mid_{\psi_0 = \psi}$$

Since $p(Y|\xi) = p(Y|s_\lambda(\psi_0), \xi)p(s_\lambda(\psi_0)|\xi)$,

$$E[\partial_\psi \log p(Y|\xi)|s_\lambda(\psi_0)] = \mathbb{E}[\partial_\psi \log p(Y|s_\lambda(\psi_0), \xi)] + \mathbb{E}[\partial_\psi \log p(s_\lambda(\psi_0); , xi)|s_\lambda(\psi_0)]$$

The first term of the above expression is 0, whereas the second term equals $\partial_\psi \log p(s_\lambda(\psi_0); , \xi)$. The conditional score statistic U_ψ can be regarded as the residual of $\partial_\psi \log p(Y|\xi)$ under its best prediction $s_\lambda(\psi)$.

Let $U_\psi(\xi) = \partial_\mu \ell_c(\mu, \sigma^2; \mu_0) \mid_{\mu_0=\mu}$, where

$$\ell_c(\mu, \sigma^2; \mu_0) = \log p(Y|\mu, \sigma^2) - \log p(s(\mu_0); \mu, \sigma^2)$$

we have that $E(U_\psi(\xi)) = 0$ and $\text{Var}(U_\psi(\xi) = -E[\partial_\mu^2 \ell_c(\mu, \sigma^2; \mu_0) \mid_{\mu_0=\mu}$

14 Over- and Varying-Dispersion Models

Analyzing discrete data using standard GLMs often exhibits overdispersion, which reflects the fact that the actual variability of the discrete data exceeds their nominal variances predicted by standard GLMs. For example, if $Y_i \sim Bin(n_i, \pi(x_i))$, the theoretical variance is $n_i\pi_i(1-\pi_i)$, but if there is overdispersion, the true/empirical variance is greater than this.

The consequences of failing to account for overdispersion are

1. A failure of appropriately modeling the variance-mean relationship in the exponential family.
2. Underestimating the standard error of the parameter estimates in GLMs.
3. Produce misleading results.

There are two statistical methods for handling overdispersion. These include assuming a more general form of the variance function and two-level hierarchical models.

14.a Quasi-likelihood

We will illustrate the first type of overdispersion model with a binomial GLM. Suppose $y_i \sim B(n_i, \pi(x_i))$ and $g(\pi(x_i)) = x_i^T\beta$. We have $E(y_i) = n_i\pi(x_i)$ and $\text{Var}(y_i) = \phi n_i \pi(x_i)(1-\pi(x_i)$ where $\phi = \sigma^2$ is a free parameter, and we have overdispersion if $\phi > 1$.

The quasi-likelihood estimate $\hat{\beta}_p$ is found by solving the quasi-likelihood score equation:

$$\sum_{i=1}^n \partial_\beta(E(y_i)^T)\text{Var}(y_i)^{-1}(y_i - E(y_i)) = \sum_{i=1}^n n_i \frac{\partial\pi(x_i)}{\partial\beta^T} \frac{y_i - n_i\pi(x_i)}{\sigma^2 n_i \pi(x_i)(1-\pi(x_i))} = 0$$

The n_i in the numerator and denominator cancel out, and taking the σ^2 out from the denominator gives the quasi-likelihood equations.

$$S_n(\beta) = \sum_{i=1}^n \frac{y_i - n_i\pi(x_i)}{\pi(x_i)(1-\pi(x_i))} \frac{\partial\pi(x_i)}{\partial\beta} = \mathbf{0}$$

To find the fisher information, we put the $1/\sigma^2$ back into the quasi-score and take the derivative of the quasi-score with respect to β.

$$\frac{1}{\sigma^2}S_n(\beta) = \frac{1}{\sigma^2}\sum_{i=1}^n \partial_\beta(y_i - n_i\pi(x_i))\frac{\partial_\beta \pi(x_i)^T}{\pi(x_i)(1-\pi(x_i))} + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - n_i\pi(x_i))\partial_\beta \frac{\partial_\beta \pi(x_i)}{\pi(x_i)(1-\pi(x_i))}$$

If we take expectations, the second term in the sum becomes 0 because the only random term is y_i and $E(y_i - \pi(x_i)) = 0$. Moreover, $\partial_\beta(y_i - n_i\pi(x_i)) = -\partial_\beta\pi(x_i)$. Thus, the Fisher information becomes

$$-\frac{1}{\sigma^2}E(\partial_\beta S_n(\beta)) = \frac{1}{\sigma^2}\sum_{i=1}^n \frac{n_i}{\pi(x_i)(1-\pi(x_i))}\partial_\beta\pi(x_i)^{\otimes 2}$$

Thus, the Newton-Raphson equation is

$$\beta^{t+1} = \beta^t + \left\{ \left[\sum_{i=1}^n \frac{n_i}{\pi(x_i)(1-\pi(x_i))}\partial_\beta\pi(x_i)^{\otimes 2} \right]^{-1} S_n(\beta) \right\}_{\beta^t}$$

Estimates for σ^2 can be obtained from moment estimate. Note that

$$E\tilde{\sigma}^2 = E\left[\sum_{i=1}^n \frac{(y_i - n_i\pi(x_i))^2}{\pi(x_i)(1-\pi(x_i))} \right] = n\sigma^2$$

Thus,

$$\hat{\sigma}^2 = \frac{n}{n-p}\tilde{\sigma}^2$$

is a consistent estimator for σ^2.

14.b Two-level Hierarchical Model

The two levels of the hierarchical model are

1. $Y_i|P_i \sim B(n_i, P_i)$
2. P_i are i.i.d. random variables with $E(P_i) = \pi(x_i)$ and $\text{Var}(p_i) = \sigma^2\pi(x_i)(1-\pi(x_i))$

Thus, $E(Y_i) = n_i\pi(x_i)$ and $\text{Var}(Y_i) = n_i\pi(x_i)(1-\pi(x_i))[1+\sigma^2(n_i-1)]$. Both of these can be found by the Law of Total Expectation / Tower Property. Note that the marginal distribution of Y_i is

$$p(y_i) = \int p(y_i|p_i)p(p_i)dp_i$$

14.b.1 Beta-Binomial Model

In level 2, $P_i \overset{i.i.d.}{\sim} Beta(\alpha, \beta)$ and

$$p(P_i;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}P_i^{\alpha-1}(1-P_i)^{\beta-1}$$

Now, make the 1-1 reparameterization $\pi = \alpha/(\alpha+\beta)$ and $\rho = 1/(\alpha+\beta)$

14.b.2 Negative Binomial Model

The negative binomial model has $Y_i|\theta_i \sim Poisson(\theta_i)$ and $\theta_i \sim Gamma(\mu_i, k)$.

15 Generalized Estimating Equations

15.a Quasi-Likelihood

Quasi-likelihood provides an important method for making statistical inference without making parametric assumption.Quasi-likelihood can be applied to independent and dependent observations.

Suppose that the components of Y are independent and satisfy $E(Y) = \mu$ and $\text{Cov}(Y) = \sigma^2 V(\mu) = \text{diag}(V_1(\mu_1), \ldots, V_n(\mu_n))$, where μ is a known mean function and $V(\mu)$ is a known variance function. Moreover, $\mu_i = \mu(x_i; \beta) = \mu_i(\beta), i = 1, \ldots, n$.

The quasi log-likelihood for μ basd on Y is given by

$$\ell_q(\mu, y) = \sum_{i=1}^{n} q_i(\mu_i, y_i) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\sigma^2 V_i(t)} dt$$

The random variable $U_i := \frac{y_i - \mu_i}{\sigma^2 V_i(\mu_i)}$ has the following properties:

- $E(U_i) = 0$
- $\text{Var}(U_i) = 1/[\sigma^2 V_i(\mu_i)] = -E(\partial_{\mu_i} U_i)$

The maximum quasi-likelihood estimator is given by $\hat{\beta} = \arg\max \ell_q(\mu(\beta), y)$. The score function can be expressed as

$$\begin{aligned} \frac{\partial \ell_q}{\partial \mu} &= \phi V^{-1}(\mu)(Y - \mu) \\ \implies S_n(\beta) := \frac{\partial \ell_q}{\partial \beta} &= \frac{\partial \mu}{\partial \beta}^T \frac{\partial \ell_q}{\partial \mu} = \phi D^T V^{-1} e(\beta) \end{aligned}$$

where $D = \frac{\partial \mu}{\partial \beta}^T$ and $e = Y - \mu(\beta)$. Of course, $S_n(\hat{\beta}) = 0$ by definition of $\hat{\beta}$.

The asymptotic covariance matrix for $\hat{\beta}$ is given by

$$\text{Cov}(\hat{\beta}) \approx \sigma^2 (D^T V^{-1} D)^{-1}$$

We will now derive this result. From the first order Taylor series expansion of the quasi-score equation about $\hat{\beta}$, we have

$$0 = S_n(\hat{\beta}) \approx S_n(\beta_*) + \partial_\beta S_n(\beta_*)(\hat{\beta} - \beta_*)$$

Rearranging some terms, we get

$$\sqrt{n}(\hat{\beta} - \beta_*) = [-n^{-1}\partial_\beta S_n(\beta_*)]^{-1}\frac{1}{\sqrt{n}}S_n(\beta_*)$$

Using properties of covariances, we have

$$\text{Cov}(\sqrt{n}\hat{\beta}) = [-n^{-1}\partial_\beta S_n(\beta_*)]^{-1}\text{Cov}\left(\frac{1}{\sqrt{n}}S_n(\beta_*)\right)[-n^{-1}\partial_\beta S_n(\beta_*)]^{-1}$$

Now, the outer terms can be expressed as

$$\begin{aligned} -n^{-1}\partial_\beta S_n(\beta_*) &= -n^{-1}\partial_\beta\left[\sum_{i=1}^n \partial_\beta\mu_i(\beta_*)V_i(\beta_*)^{-1}e_i(\beta_*)\right] \\ &= n^{-1}\sum_{i=1}^n \partial_\beta[\partial_\beta\mu_i(\beta_*)V_i(\beta_*)^{-1}]e_i(\beta_*) + n^{-1}\sum_{i=1}^n \partial_\beta\mu_i(\beta_*)V_i(\beta_*)^{-1}\partial_\beta\mu_i(\end{aligned}$$

$$\approx n^{-1}\sum_{i=1}^n \partial_\beta\mu_i(\beta_*)V_i(\beta_*)^{-1}\partial_\beta\mu_i(\beta_*)^T$$

since the first term converges to 0 in probability by the SLLN. It follows that

$$\begin{aligned} \text{Cov}\left[S_n(\beta_*)/\sqrt{n}\right] &= \text{Cov}\left[\sum_{i=1}^n \partial_\beta\mu_i(\beta_*)V_i(\beta_*)^{-1}e_i(\beta_*)/\sqrt{n}\right] &= \sigma^2\sum_{i=1}^n \partial_\beta\mu_i(\beta_*)V_i(beta_*)^{-1}\partial_\beta\mu_i(\beta_*)/n \\ &= \sigma^2 D^T V^{-1} D/n \end{aligned}$$

since $\text{Var}(e_i) = \sigma^2 V_i$. Thus, we have

$$\text{Cov}(\hat{\beta}) = \text{Cov}(\sqrt{n}\hat{\beta})/n = \sigma^2(D^T V^{-1} D)^{-1}$$

To gain an estimate for σ^2, note that

$$\mathbb{E}\frac{(y_i - \mu_i)^2}{V(\mu_i)} = \sigma^2, \quad i = 1, \ldots, n$$

Hence, by the weak law of large numbers, the average of these random variables converge in probability to σ^2. It follows that a consistent estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

It turns out (Theorem 9.1) that the quasi-likelihood estimator of β is Godambe efficient for β, which is related to the notion of BLUE. Essentially this theorem states that for any linear estimator $\tilde{\beta}$ that solves the equation $S_n(y, \beta) = H^T(y - \mu(\beta))$ for some matrix H, the difference in covariance matrices $\text{Cov}\tilde{\beta} - \text{Cov}(\hat{\beta})$ is positive semidefinite. The proof of this is on pages 918 - 921 in the notes. Essentially, it involves two steps:

1. Do a Taylor Series expansion on the score equation above with the H matrix. Find an expression for $\text{Cov}(\sqrt{n}\tilde{\beta})$
2. To show the difference in covariance matrices is positive semidefinite, we can show instead that
$$\{\text{Cov}(\hat{\beta})\}^{-1} - \{\text{Cov}(\tilde{\beta})\}^{-1}$$
is positive semi-definite.

15.b Z-Estimators

Let $U_1, \ldots, U_n$ be i.i.d. as $g(U_i)$, which is unknown. Consider the Z-estimator $\hat{\xi}$, which satisfies
$$S_n(\hat{\xi}) = \sum_{i=1}^{n} h(\hat{\xi}, U_i) = \mathbf{0}$$

Let $S(\xi) = \int h(\xi, U)g(U)dU$ and $S(\xi_*) = 0$ Then

- $n^{-1}S_n(\xi) \to S(\xi)$
- $\hat{\xi} \to \xi_*$
- $n^{-1}S_n(\hat{\xi}) = 0$ and $S(\xi_*) = 0$.
- $\hat{\xi} = \arg\max_{\xi} -[n^{-1}S_n(\xi)]^2$ and $\xi_* = \arg\max_{\xi} -[S(\xi)]^2$

Theorem 9.2
If for every $\epsilon > 0$,
$$\sup_{\xi \in \Xi} |n^{-1}S_n(\xi) - S(\xi) \to_p 0, \quad \sup_{\xi : ||\xi - \xi_*|| \geq \epsilon} ||S(\xi)|| > ||S(\xi_*)|| = 0$$
then $\hat{\xi} \to_p \xi_*$.

Theorem 9.3 (Asymptotic Normality of Z-Estimator) Under some conditions (see page 931),
$$\sqrt{n}(\hat{\xi} - \xi_*) = [-E(\partial_{\xi} h(\xi_*, U))]^{-1} n^{-1/2} \sum_{i=1}^{n} h(\xi_*, U_i) + o_p(1)$$

In particular, $\sqrt{n}(\hat{\xi} - \xi_*)$ converges to a $N(0, \Sigma(\xi_*))$ distribution as $n \to \infty$, where
$$\begin{aligned}\Sigma(\xi_*) &= B^{-1}CB^{-1} \\ &= [-E(\partial_{\xi} h(\xi_*, U))]^{-1}[E(h(\xi_*, U))^{\otimes 2}][-E(\partial_{\xi} h(\xi_*, U))]^{-1}\end{aligned}$$

15.b.1 Hypothesis Testing

Since we are estimating without a likelihood,, the likelihood ratio test is not valid. We do have a Wald and Score test.

Consider testing the nonlinear hypotheses

$$H_0 : h_0(\xi) = b_0 \text{ vs } .H_1 : h_0(\xi) \neq b_0$$

The Wald test statistic is given by

$$W_n = (h_0(\hat{\xi}) - b_0)^T \left[\text{Cov}(h_0(\hat{\xi}))\right]^{-1} (h_0(\hat{\xi}) - b_0)$$

where it can be shown by the Delta Method hat

$$\text{Cov}(h_0(\hat{\xi})) = H(\hat{\xi})\text{Cov}(\hat{\xi})H(\hat{\xi})^T$$

Score Test
We only consider testing the linear hypothesis $H_0 : R\xi = b_0$. Let $R = (I_r, 0)$, $R\xi = \xi(1)$ and $\xi^T = (\xi(1)^T, \xi(2)^T)$. Let $h(\xi, U) = h_1(\xi, U)^T, h_2(\xi, U)^T)$

Let $\tilde{\xi} = (b_0^T, \xi(2)(b_0)^T)^T$ be the MLE under the null hypothesis. Write $S_n(\xi) = (S_{n,1}(\xi)^T, S_{n,2}(\xi)^T)^T$. The Score statistic is given by

$$SC_n = S_{n,1}(\tilde{\xi})^T \Sigma^{11}(\tilde{\xi}) S_{n,1}(\tilde{\xi})$$

where

$$\Sigma^{11}(\tilde{\xi}) = \{(I_r, -\tilde{S}_{n,12}\tilde{S}_{n,22})\}^{-1}\{\sum_{i=1}^{n}[h(\tilde{\xi}, U_i) - \bar{h}(\tilde{xi})]^{\otimes 2}\}\{(I_r, -\tilde{S}_{n,12}\tilde{S}_{n,22})\}^{-1}$$

and where $\bar{h}(\tilde{\xi}) = \sum_{i=1}^{n} h(\tilde{\xi}, U_i)/n$ and $\tilde{S}_{n,12} = S_{n,12}(\tilde{\xi})$ and $\tilde{S}_{n,22} = S_{n,22}(\tilde{\xi})$ are the components of the derivative of the score equation.

15.c Generalized Estimating Equations

GEEs are used when we no longer have the independence assumption, for example, if we have repeated measures in which we measure independent individuals across time. There is correlation within each individual, although the between correlation is 0.

The key idea of GEEs is to use the same link functions and linear predictor set-up as for GLMs in the independence case, but GEE methods account for the correlation structure of the covariance matrix of the responses in the estimation process.

For the j-th time point from the i-th subject, we observe a scalar y_{ij} that is the response and a $p \times 1$ vector of covariates x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, T_i$.

The model is as follows:

- $E(y_{ij}) = \dot{b}(\theta_{ij}) = \mu_{ij}$ and
  $\text{Var}(y_{ij}) = \phi^{-1}\ddot{b}(\theta_{ij})$.
- $g(\mu_{ij}) = x_{ij}^T\beta$
- <u>Working Covariance Matrix:</u>
  $\phi^{-1}V_i = \phi^{-1}B_i^{1/2}R_i(\alpha)B_i^{1/2}$
  where $B_i = \text{diag}\{\text{Var}(y_{i1}), \ldots, \text{Var}(y_{i,T_i})\}$ and $R_i(\alpha)$ is a working covariance matrix for y_i.
- $G_n(\alpha, \beta) = \sum_{i=1}^n D_i^T V_i^{-1}[y_i - \mu_i(\beta)] = 0$, where $D_i = \frac{\partial \mu_i}{\partial \beta}$.
- The GEE estimator $\hat{\beta}$ is the solution of these GEEs.

The working correlation matrix $R_i(\alpha)$ is fully specified by a vector of parameters α. α is usually measured by the Pearson residual $e_{ij} := (y_{ij} - \mu_{ij})/\sqrt{\phi^{-1}\ddot{b}(\theta_{ij})}$

The GEE Newton-Raphson algorithm to obtain estimates is given as follows:

1. Compute an initial estimate $\beta^{(0)}$ based on an independent working correlation matrix.
2. Compute the working correlation matrix $R_i(\alpha)$ based on the Pearson residuals and the current $\beta^{(r)}$.
3. Compute an estimate of the covariance $V_i = \phi B_i^{1/2} R_i(\hat{\alpha}) B_i^{1/2}$.
4. Update β by the Newton-Raphson equation:

$$\beta^{(r+1)} = \beta^{(r)} + \left[\sum_{i=1}^n D_i^T V_i^{-1} D_i\right]^{-1} \sum_{i=1}^n D_i^T V_i^{-1}(y_i - \mu_i)\Bigg|_{\beta^{(r)}}$$

   Repeat steps 2-4 until convergence.

The resulting estimator is consistent $\hat{\beta} \to \beta_*$ in probability as $n \to \infty$. In addition, we have

$$\sqrt{n}(\hat{\beta} - \beta_*) \to_d N(0, \Sigma)$$

and an estimator for Σ is given by

$$\hat{\Sigma} = K_0^{-1}\left[n^{-1}\sum_{i=1}^n D_i^T V_i^{-1}\text{Cov}(Y_i) D_i\right] K_0^{-1}$$

where $K_0^{-1} = n^{-1}D_i^T V_i^{-1} D_i$ and $\text{Cov}(Y_i)$ can be replaced by $(y_i - \mu_i(\hat{\beta}))^{\otimes 2}$