Contents

1 Foundation **1**
1.1 Matrix algebra 1
1.2 Matrix decomposition 4

2 Linear regression **5**
2.1 Geometric approach 5
2.2 Algebraic approach 5
2.2.1 When X is full rank 5
2.2.2 When X is not full rank 5
2.3 Projection matrix P 6
2.3.1 When X is full rank 6
2.3.2 When X is not full rank 6
2.4 Residual Sums of Squares (RSS) 6
2.5 PROPERTIES OF LEAST SQUARES ESTIMATES 7
2.5.1 Best Linear Unbiased Estimator, BLUE 7
2.5.2 Unbiased estimation of σ^2 7

3 Linear regression with distribution assumption **8**
3.1 MLE 8
3.1.1 review on MLE properties 9
3.1.2 Orthogonal columns in the regression matrix 9
3.2 Estimation with linear constrain 9
3.3 Identifiability 11
3.4 Estimability 11

4 Generalized least square **12**

5 Hypothesis Testing **13**
5.1 Likelihood ratio test 13
5.2 F test 14

6 Some comments about ANOVA **15**

1 Foundation

1.1 Matrix algebra

Vectors dependency definition
A set of vectors $D = \{x_1, x_2, \ldots, x_r\}$ is called *linearly dependent* if there is a set of scalar $\alpha_1, \alpha_2, \ldots, \alpha_r$ not all zero such that

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

Conversely, if $\sum_{i=1}^{r} \alpha_i x_i = 0 \Rightarrow \alpha_i = 0,\ i = 0, 1, \ldots, r$, then $D = \{x_1, x_2, \ldots, x_r\}$ are *linearly independent*.

Column space
Suppose A is an n*p matrix. Then each column of A is a vector in $\mathbb{R}^n$. We can write $A = (x_1, \ldots, x_n)$, where each $x_i \in \mathbb{R}^n$, $i = 1, \ldots, p$. The space spanned by the columns of A is called the *column space* of A, written $C(A)$. That is $S(A) = C(A)$, where $S(A)$ is the space spanned by A.

Vector differentiation
Define the vector differentiation as follows

$$\frac{d}{d\beta} = \left(\frac{d}{d\beta_i}\right)$$

where β is a n*1 vector. Then we have the following properties

$$\frac{d\left(\beta' a\right)}{d\beta}=a$$

$$\frac{d\left(a'\beta\right)}{d\beta}=a$$

$$\frac{d\left(\beta' A\beta\right)}{d\beta}=2A\beta$$

Patterned matrices
If all inverses exist

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

where $B_{22}=A_{22}-A_{21}A_{11}^{-1}A_{12}, B_{12}=A_{11}^{-1}A_{12}$, $B21=A_{21}A_{11}^{-1}, C_{11}=A_{11}-A_{12}A_{22}^{-1}A_{21}, C_{12}=A_{12}A_{22}^{-1}$, and $C_{21}=A_{22}^{-1}A_{12}$

Nonsingular
Suppose A is an n*n square matrix. Then A is said to be nonsingular if there exists a matrix A^{-1} such that

$$A^{-1}A=AA^{-1}=I$$

Null space
The set of all x such that $Ax=0$ is a vector space and is called the null space of A, written $N(A)$.

Theorem 1.1. *Suppose A is n*n. If $r(A)=r$ then $r(N(A))=n-r$*

Trace
Suppose A is an n*n square matrix with ijth element a_{ij}. The trace of A is defined as

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

Property 1. $\operatorname{tr}(A+B)=\operatorname{tr}(A)+\operatorname{tr}(B)$

Property 2. The trace is invariant under cyclic

Property 3. Suppose A; B; C are n*n square matrices. Then

$$\operatorname{tr}(ABC)=\operatorname{tr}(BCA)=\operatorname{tr}(CAB)$$

Property 4. If A is an n*n matrix with eigenvaules λ_j, then $\operatorname{tr}(A)=\sum_{i=1}^{n}\lambda_j$ and $\det(A)=\prod_{i=1}^{n}\lambda_j$

Property 5. Assume that A is symmetric, then $\operatorname{tr}\left(A^s\right)=\sum_{i=1}^{n}\lambda_i^s$

Property 6. Assume that A is symmetric and nonsingular, then the eigenvalues of A^{-1} are $\lambda_i^{-1}(i=1,\ldots,n)$ and hence $\operatorname{tr}\left(A^{-1}\right)=\sum_{i=1}^{n}\lambda_i^{-1}$

Rank

Property 1. $\operatorname{rank}(AB)\leq\operatorname{minimum}(\operatorname{rank}A,\operatorname{rank}B)$

Property 2. If A is any matrix, and P and Q are any conformable nonsingular matrices, then $\operatorname{rank}(PAQ)=\operatorname{rank}(A)$

Property 3. $\operatorname{rank}(A)=\operatorname{rank}\left(A'\right)=\operatorname{rank}(A)=\operatorname{rank}\left(AA'\right)$

Property 4. Let A be any m*n matrix such that $r=\operatorname{rank}(A)$ and $s=\operatorname{nullity}(A)$, [the dimension of $N(A)$, the null space or kernel of A, i.e., the dimension of $\{x: Ax=0\}$]. Then $r+s=n$

Property 5. If $C(A)$ is the column space of A, then $C\left(A'A\right)=C\left(A'\right)$

Property 6. If A is symmetric, then $\operatorname{rank}(A)$ is equal to the number of nonzero eigenvalues

Eigenvalues and Eigenvectors of a matrix
Suppose A is an n*n square matrix.

$$Ax = \lambda x, \lambda \in \mathbb{R}^1$$

then λ is called an eigenvalue of A and x is called an eigenvector. Note that eigenvectors are not unique.

Property 1. Assume that A is symmetric, then the eigenvalues of $(I_n + cA)$ are $1 + c\lambda_i, (i = 1, \ldots, n)$

Property 2. Any n*n symmetric matrix A has a set of n orthonormal eigenvectors, and $C(A)$ is the space spanned by those eigenvectors corresponding to nonzero eigenvalues

Theorem 1.2. *if x_1 and x_2 are eigenvectors with the same eigenvalue, then any nonzero linear combination of x_1 and x_2 is also an eigenvector with the same eigenvalue.*

Theorem 1.3. *λ is an eigenvalue of A if and only if $A - \lambda I$ is singular.*

The eigenvalues of a matrix A are found by finding the solutions of the equation for λ

$$\det(A - \lambda I) = 0$$

Theorem 1.4. *Suppose A is n*n with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$*

- $\det(A) = \prod_{i=1}^{n} \lambda_i$
- *if A is singular, then* $\det(A) = 0$
- *if A is nonsingular then A^{-1} exists and the eigenvalues are given by $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$*
- *the eigenvalues of A' are the same as those of A*
- $\operatorname{tr}(A) = \sum_{i=1}^{n} \lambda_i$ *and* $\operatorname{tr}\left(A^{-1}\right) = \sum_{i=1}^{n} \lambda_i^{-1}$
- *if A is symmetric then* $\operatorname{tr}\left(A^r\right) = \sum_{i=1}^{n} \lambda_i^r$ *for any integer r*

Orthogonal matrix
A square matrix is orthogonal if $PP' = P'P = I$

Theorem 1.5. *An n*n matrix P is orthogonal if and only if the columns of P form an orthonormal basis for R^n, that is the columns of P are all unit vectors and orthogonal to each other.*

Positive-semi definite matrices
A symmetric matrix A is said to be positive-semidefinite (p.s.d.) if and only if $x'Ax \geq 0 \quad \forall x$

Property 1. The eigenvalues of a p.s.d. matrix are nonnegative

Property 2. If A is p.s.d., then $\operatorname{tr}(A) \geq 0$

Property 3. A is p.s.d. of rank r if and only if there exists an n*n matrix R of rank r such that $A = RR'$

Property 4. If A is an n*n p.s.d. matrix of rank r, then there exists an n*r matrix S of rank r such that $S'AS = I_r$

Property 5. A is p.s.d. then $x'Ax = 0 \Longrightarrow Ax = 0$

Positive-definite matrices
A symmetric matrix A is said to be positive- definite (p.d.) if $x'Ax > 0 \quad \forall x \neq 0$. We note that a p.d. matrix is also p.s.d.

Property 1. The eigenvalues of a p.d. matrix A are all positive; thus A is also nonsingular

Property 2. A is p.d. if and only if there exists a nonsingular R such that $A = RR'$

Property 3. If A is p.d. then so is A^{-1}

Property 4. If A is p.d. then $\operatorname{rank}(CAC') = \operatorname{rank}(C)$

Property 5. If A is an n*n p.d. matrix and C is p*n of rank p, then CAC' is p.d.

Property 6. If X is n*p of rank p, then $X'X$ is p.d.

Property 7. A is p.d. if and only if all the leading minor determinants of A [including $\det(A)$ itself] are positive.

Property 8. The diagonal elements of a p.d. matrix are all positive

Property 9. (Cholesky decomposition) If A is p.d., there exists a unique upper tri-angular matrix R with positive diagonal elements such that $A = R'R$

Property 10. (Square root of a positive-definite matrix) If A is p.d., there exists a p.d. square root $A^{1/2}$ such that $(A^{1/2})^2 = A$

Idempotent matrices
A matrix P is idempotent if $P^2 = P$. A symmetric idempotent matrix is called a **projection matrix**

Property 1. If P is symmetric, then P is idempotent and of rank r if and only if it has r eigenvalues equal to unity and n-r eigenvalues equal to zero.

Property 2. If P is a projection matrix then $\operatorname{tr}(P) = \operatorname{rank}(P)$

Property 3. If P is idempotent then so is I-P

Property 4. Projection matrices are positive-semidefinite

Property 5. If $P_i(i = 1, 2)$ is a projection matrix and $P_1 - P_2$ is p.s.d. then $P_1P_2 = P_2P_1 = P_2$ and $P_1 - P_2$ is a projection matrix

Generalized inverse
Consider the linear transformation $A : \mathbb{R}^p \longrightarrow \mathbb{R}^n$. A generalized inverse of A is the linear transformation A^- such that

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

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

$$AA^-A = A$$

from the definition we can get

$$\left(A^-A\right)\left(A^-A\right) = A^-\left(AA^-A\right) = A^-A$$

thus A^-A is idempotent and hence a projection. The generalized inverse is not unique, but always exists.
Moore-Penrose generalized inverse
Suppose A is an n*p matrix. If the generalized inverse $A^-_{p\times n}$ satisfies four conditions

- $AA^-A = A$
- $A^-AA^- = A^-$
- $(\mathrm{A}A^-)' = A\ \mathrm{A}^-$
- $(\mathrm{A}^-\mathrm{A})' = \mathrm{A}^-\mathrm{A}$

then $A^-_{p\times n}$ is called the Moore-Penrose generalized inverse. The Moore-Penrose generalized inverse is unique.

1.2 Matrix decomposition

Theorem 1.6. ***Spectral decomposition***
*Suppose A is an n*n* symmetric matrix. *Then there exists an orthogonal matrix P such that*

$$A = P \wedge P'$$

*where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is an n*n diagonal matrix of the eigenvalues of A with $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_n$ and P is the orthogonal matrix of orthonormal eigenvectors corresponding to the eigenvalues of A.*

Theorem 1.7. ***Singular value decomposition***
*Suppose A is an n*p matrix of rank r, where $r \leq \min(n, p)$. There exists orthogonal matrices $U_{p\times p}$ and $V_{n\times n}$ such that*

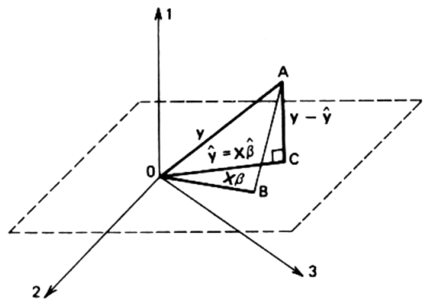
$$V'AU = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix} \longrightarrow A = VDU', \text{ where } D = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix}$$

*where $\Delta = \operatorname{diag}(\delta_1, \ldots, \delta_r)$ is an r*r diggonal matrix with $\delta_1 \geq \delta_2 \ldots \geq \delta_r > 0$. The δ_i are called the singular values of A.*

2 Linear regression

We are going to learn about the general linear model in the form of $E(Y) = X\beta$ and estimation of β using the least squared method and associated distribution theory. The LSE method consists of minimizing $\sum_i \epsilon_i^2$ with respect to β
Let $\theta = X\beta$, we minimize $\epsilon'\epsilon = ||Y - \theta||^2$ subject to $\theta \in C(X) = \Omega$, where Ω is the column space of X.

2.1 Geometric approach



From the image we know that $||Y - \hat{\theta}||^2$ minimized when $\Omega \perp (y - \hat{\theta})$, which is $X^{'}(Y - \hat{\theta}) = 0 \Rightarrow X'\hat{\theta} = X'Y$ thus $\hat{\theta} = (X')^{-1}X'Y$. Here $\hat{\theta}$ is *uniquely determined* being the unique orthogonal projection of Y onto Ω, but β is not necessarily unique.
We have

$$X'\hat{\theta} - X'Y = 0$$

defined as **normal equation**

2.2 Algebraic approach

To derive $\hat{\beta}$ algebraically. Write

$$\begin{aligned}\epsilon'\epsilon &= (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y'Y} - 2\boldsymbol{\beta'X'Y} + \boldsymbol{\beta'X'X\beta}\end{aligned}$$

By the vector differentiation we have

$$-2X'Y + 2X'X\beta = 0$$

$$\boldsymbol{X'X\beta} = \boldsymbol{X'Y}$$

To prove that we get the minimal β from this equation, we still need to take the second derivative, from which we get $2X'X \geq 0$.

2.2.1 When X is full rank

When the columns of X are linearly independent i.e. X is full rank, then there exists a unique vector

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

Cause when X is full rank, $X^{'}X$ is positive-definite and therefore non-singular

2.2.2 When X is not full rank

When the columns of X are linearly dependent i.e. X is not full rank, then the solution is given by

$$\hat{\beta} = \left(\boldsymbol{X'X}\right)^{-} \boldsymbol{X'Y}$$

where $\left(\boldsymbol{X'X}\right)^{-}$ is any generalized inverse of $\left(\boldsymbol{X'X}\right)$
Proof: $\boldsymbol{X'X\beta} = \boldsymbol{X'Y}$. Consider a g-inverse $\left(\boldsymbol{X'X}\right)^{-}$. We know that $(X'X)(X'X)^{-}(X'X) = (X'X)$. Then we have $(X'X)(X'X)^{-}(X'X)\hat{\beta} = (X'X)(X'X)^{-}X'Y = X'Y$, by comparing this equation with $\boldsymbol{X'X\beta} = \boldsymbol{X'Y}$. We get that $\hat{\beta} = \left(\boldsymbol{X'X}\right)^{-} \boldsymbol{X'Y}$ is a solution.

2.3 Projection matrix P

From the normal equation

$$\hat{\theta} = X\hat{\beta} = X\left(X'X\right)^{-}X'Y = PY$$

we define the projection matrix P as

$$\boldsymbol{P} = X\left(X'X\right)^{-}X'$$

P is unique and does not depend on the g-inverse used. When the inverse of $X^{'}$ exists

$$\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}X'$$

2.3.1 When X is full rank

Suppose that $\boldsymbol{X}$ is $n \times p$ of rank $p,$ so that $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ then following hold.
(i) $\boldsymbol{P}$ and $\boldsymbol{I_n} - \boldsymbol{P}$ are symmetric and idempotent.
(ii) $\operatorname{rank}\left(\boldsymbol{I_n} - \boldsymbol{P}\right) = \operatorname{tr}\left(\boldsymbol{I_n} - \boldsymbol{P}\right) = n - p.$
(iii) $\boldsymbol{PX} = \boldsymbol{X}$

Proof
(i) $PP = X\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X' = X\left(X'X\right)^{-1}X'$
(ii) $(I-P)(I-P) = I - P - P + P \cdot P = 1 - P$
(iii) $PX = X\left(X'X\right)^{-1}X'X = X$

2.3.2 When X is not full rank

If X has rank $r < p$, then the above result still holds but with p replaced by r
Theorem: Suppose that $\boldsymbol{X}$ is $n \times p$ of rank r so that $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{X}'$ then following hold.
(i) $\boldsymbol{P}$ and $\boldsymbol{I_n} - \boldsymbol{P}$ are symmetric and idempotent.
(ii) $\operatorname{rank}\left(\boldsymbol{I_n} - \boldsymbol{P}\right) = \operatorname{tr}\left(\boldsymbol{I_n} - \boldsymbol{P}\right) = n - r$
(iii) $\boldsymbol{PX} = \boldsymbol{X}$

Proof
X has rank r, let X_1 be the n*r matrix with r linearly independent column then $C[X_i] = C[X]$, then

$$P = X_1\left(X_1'X_1\right)^{-1}X_1$$

cause the linear space is the same, then it's easily got that $P^2 = P \quad, (I-P)^2 = (I-P)$.
Also $\exists L$ such that $X = X_1 L$
thus $PX = X_1\left(X_1'X_1\right)^{-1}X_1' \cdot X_1 L = X_1 L = X$

2.4 Residual Sums of Squares (RSS)

We denote the fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{Y}} = \left(\hat{Y}_i, \ldots, \hat{Y}_n\right)'$. The elements of the vector

$$\begin{aligned}\boldsymbol{Y} - \hat{\boldsymbol{Y}} &= \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \left(\boldsymbol{I_n} - \boldsymbol{P}\right)\boldsymbol{Y},\end{aligned}$$

then

$$\begin{aligned}RSS &= [(I-\boldsymbol{P})\boldsymbol{Y}]'[(I-\boldsymbol{P})\boldsymbol{Y}] \\ &= Y'(I-\boldsymbol{P})Y\end{aligned}$$

Another way of doing this is

$$\begin{aligned}\epsilon'\boldsymbol{\epsilon} &= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{\beta}}'\left[\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{Y}\right] \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

thus
$RSS = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$

2.5 PROPERTIES OF LEAST SQUARES ESTIMATES

$\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$. That is

$E(\hat{\beta}) = \beta$

The variance of the Least Square Estimator of β is given by

$\operatorname{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$

Proof

$\operatorname{Var}(\hat{\beta}) = (X'X)^{-1} X' \operatorname{Var}(Y) X \left[(X'X)^{-1}\right]' = \sigma^2 (X'X)^{-1}$

$E(\hat{\beta}) = (X'X)^{-1} X' E[Y] = (X'X)^{-1} X'X\beta = \beta$

Similar result holds for $\hat{\theta}$

$$E(\hat{\theta}) = PE(Y) = P\underbrace{X\beta}_{\theta} = X\beta = \theta$$
$$\operatorname{Var}(\hat{\theta}) = P \operatorname{Var}(Y) P' = \sigma^2 PP' = \sigma^2 P$$

2.5.1 Best Linear Unbiased Estimator, BLUE

THEOREM 3.2: Let $\hat{\boldsymbol{\theta}}$ be the least squares estimate of $\boldsymbol{\theta} = \boldsymbol{X\beta}$ where $\boldsymbol{\theta} \in \Omega = C(\boldsymbol{X})$ and $\boldsymbol{X}$ may not have full rank. Then among the class of linear unbiased estimates of $\boldsymbol{c}'\boldsymbol{\theta}, \boldsymbol{c}'\hat{\boldsymbol{\theta}}$ is the unique estimate with minimum variance. We say that $c'\hat{\theta}$ is the best linear unbiased estimate *BLUE* of $\boldsymbol{c}'\theta$

Proof

$$\begin{aligned}
&\hat{\theta} = PY \qquad (LSE) \\
&E\left(c'\hat{\theta}\right) = c'E(\hat{\theta}) = c'\theta \quad , \forall \theta \in \Omega = c[x] \Rightarrow \text{ unbiasness}
\end{aligned}$$

Let $d^{'}Y$ be another estimator which is linear and unbiased then

$$E(d^{'}Y) = d^{'}E(Y) = d^{'}X\beta = d^{'}\theta \xrightarrow{\text{unbiasness}} (d' - c')\,\theta = 0 \quad \Rightarrow \quad (d - c) \perp \Omega \Rightarrow P(c - d) = 0 \Rightarrow Pc = Pd$$

then

$$\begin{aligned}
\operatorname{Var}(d'Y) - \operatorname{Var}\left(c'\hat{\theta}\right) &= d' \operatorname{Var}(Y) d - \operatorname{Var}\left(c'\hat{\theta}\right) \Leftarrow c'\hat{\theta} = c'PY = (Pc)'Y = (Pd)'Y \\
&= d'\left(\sigma^2 I\right) d - \operatorname{Var}\left[(Pd)'Y\right] = \sigma^2 d'd - (Pd)'\sigma^2 (Pd) \\
&= \sigma^2 d'd - \sigma^2 d'Pd \\
&= \sigma^2 d'(I - P)d \\
&= \sigma^2 d'(I - P)'(I - P)d \\
&= \sigma^2 [(I - P)d]'[(1 - P)d] \geq 0
\end{aligned}$$

equality holds when

$$(I - P)d = 0 \Rightarrow d = Pd = Pc$$

If X is full rank, then $a'\hat{\beta}$ is the BLUE of $a'\beta$ for every vector a.

2.5.2 Unbiased estimation of σ^2

THEOREM 3.3 $\quad E[\boldsymbol{Y}] = \boldsymbol{X\beta}$ where X is an $n \times p$ matrix of rank $r (r \leq p)$ and $\operatorname{Var}[\boldsymbol{Y}] = \sigma^2 \boldsymbol{I_n}$ then

$$S^2 = \frac{(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})'(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})}{n - r} = \frac{RSS}{n - r}$$

is an unbiased estimate of σ^2

Proof

$$\begin{aligned}\text{residual } &= (Y - X\hat{\beta}) = (I-P)Y \\ RSS &= [(I-P)Y]'[(I-P)Y] = Y'(1-P)Y \\ E(RSS) &= E\left\{Y'(I-P)Y\right\} = tr\left[(I-P)*\sigma^2 I\right] + (X\beta)'\underbrace{(1-P)(x\beta)}_{X-PX=0} \\ &\Rightarrow E\left(\frac{RSS}{n-r}\right) = \sigma^2 \qquad \text{unbiased estimator}\end{aligned}$$

3 Linear regression with distribution assumption

Until now the only assumptions we have made about the ϵ_i are that $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I_n$. If we assume that the ϵ_i are also normally distributed, i.e. $\epsilon \sim N_n\left(0, \sigma^2 I_n\right)$ and hence $Y \sim N_n\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I_n}\right)$. A number of distributional results then follow.
THEOREM 3.5 if $Y \sim N_n\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I_n}\right)$, where X is n*p of rank p then

- $\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$
- $(\hat{\beta} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$
- $\hat{\beta}$ is independent of S^2
- $\mathrm{RSS}/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$

Proof

- $\hat{\beta} = \underbrace{(X'X)^{-1}X'}_{c}Y$ and also $Y \sim N_n\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I_n}\right)$.
  then $cY \sim MVN(cX\beta, c\Sigma c')$ where $\mathrm{rank}(c) = \mathrm{rank}(X) = \mathrm{rank}(X')$
  $cX\beta = \beta, c\Sigma C' = \sigma^2 cc' = \sigma^2\left(X'X\right)^{-1}X'X\left[(X'X)^{-1}\right]' = \sigma^2\left(X'X\right)^{-1}$
- $(\hat{\beta} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$ since $\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$
- $\hat{\beta} = \underbrace{(X'X)^{-1}X'}_{c}Y$ and $(n-p)S^2 = Y'(I-P)Y = [(I-P)Y]'[(I-P)Y]$ then $(X'X)^{-1}X'[I-P] =$
  $(X'X)^{-1}X'\left[I - X(X'X)^{-1}X'\right] = (X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X' = 0$
- $\frac{RSS}{\sigma^2} = \frac{Y'(1-P)Y}{\sigma^2}$ since $(1-P)$ is idempotent with $\mathrm{rank}(1-P) = n-r$ based on THEOREM 2.7 we have $\mathrm{RSS}/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$

3.1 MLE

Assuming full rank of X, the likelihood is

$$L\left(\beta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2\right]$$

then the log-likelihood is

$$\ell\left(\beta, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) = -\frac{n}{2}\log\left(2\pi\mu\right) - \frac{1}{2\mu}(Y - X\beta)'(Y - X\beta)$$

where $\sigma^2 = \mu$
then

$$\frac{\partial l}{\partial \beta} = -\frac{1}{2\mu}\left(-2X'Y + 2X'X\beta\right) \overset{\text{set}}{=} 0 \Rightarrow \hat{\beta}_{\text{mle}} = \left(X'X\right)^{-1}X'Y \Rightarrow \text{lse} = \text{mle}$$

and

$$\frac{\partial l}{\partial \mu} = \frac{-n}{2\mu} + \frac{1}{2\mu^2}(Y - X\beta)'(Y - X\beta) \overset{\text{set}}{=} 0 \Rightarrow \hat{\mu} = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n} = \frac{\text{RSS}}{n} \neq \hat{\mu}_{\text{lse}} = \underbrace{\frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n-p}}_{\text{unbiased estimator for } \sigma^2}$$

then for the distribution

$$\frac{\partial^2 l}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2}\left(X'X\right) \quad \Rightarrow -E\left[-\frac{1}{\sigma^2}\left(X'X\right)\right] = \frac{X'X}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial\beta\partial\mu} = \frac{1}{2\mu^2}\left(-2X'Y + 2X'X\beta\right) \text{ ,sine } \hat{\beta} = \left(X'X\right)^{-1}X'Y \quad , \frac{\partial^2 l}{\partial\beta\partial\mu} = \frac{\partial^2 l}{\partial\mu\partial\beta} = 0$$

$$\frac{\partial^2 l}{\partial\mu^2} = \frac{n}{2\mu^2} - \frac{1}{\mu^3}(Y-X\beta)'(Y-X\beta) \Rightarrow -E\left[-\frac{\partial^2 l}{\partial\mu^2}\right] = \frac{-n}{2\mu^2} + \frac{1}{\mu^3}\underbrace{E\left[(Y-X\beta)'(Y-X\beta)\right]}_{(Y-X\beta)'(\sigma^2 I)^{-1}(Y-X\beta)\sim\chi^2_n} = \frac{-n}{2\mu^2} + \frac{n\sigma^2}{\mu^3} = \frac{-n}{2\mu^2} + \frac{n\mu}{\mu^3} = \frac{n}{2\mu^2}$$

then

$$I = \begin{bmatrix} \frac{1}{\mu}X'X & 0 \\ 0 & \frac{n}{2\mu^2} \end{bmatrix} \Rightarrow I^{-} = \begin{bmatrix} \mu\left(X'X\right)^{-1} & 0 \\ 0 & \frac{2\mu^2}{n} \end{bmatrix}$$

then we have

$$\begin{pmatrix} \hat{\beta}_{\text{mle}} \\ \hat{\sigma}^2_{\text{mle}} \end{pmatrix} \overset{\text{asymptotically normal}}{\sim} \left(\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \left(X'X\right)^{-1}\sigma^2 & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}\right)$$

3.1.1 review on MLE properties

Score: The partial derivative with respect to θ of the natural logarithm of the likelihood function is called the score

$$Z = l' = \frac{\partial}{\partial\theta}\log f(X;\theta)$$

$$E(Z) = 0 \text{ and } Z \xrightarrow{d} N\left(0, I(\theta_0)\right)$$

under θ_0
Fisher information: The variance of the score is defined to be the Fisher information

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 \mid \theta\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) \mid \theta\right]$$

Property 1. If $\hat{\theta}$ is the MLE estimate of θ_0, then it has the following property:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \frac{1}{I(\theta_0)})$$

3.1.2 Orthogonal columns in the regression matrix

Suppose that in the full-rank model $E[\boldsymbol{Y}] = \boldsymbol{X\beta}$ the matrix X has a column representation where the columns are all mutually orthogonal

$$\boldsymbol{X} = \left(\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(p-1)}\right)$$

then we will have

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y = \begin{bmatrix} [x^{(0)'}x^{(0)}]^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & [x^{(p-1)'}x^{(p-1)}]^{-1} \end{bmatrix} \begin{bmatrix} x^{(0)\prime}Y \\ \vdots \\ x^{(p-1)\prime}Y \end{bmatrix} = \left[\frac{x^{(j)\prime}Y}{x^{(j)\prime}x^{(j)}}\right]$$

this implies that under the orthogonal condition, if we only want to estimate certain β_j then we don't need to fit the whole model instead we can only fit Y on $x^{(j)}$

3.2 Estimation with linear constrain

The conclusion from this part is applicable under both MLE and LSE cause we only estimate $\hat{\beta}$
Let $Y = X\beta + \epsilon$ where X is n*p of full rank p. Suppose that we wish to find the minimum of $\epsilon'\epsilon$ subject to the linear restrictions $A\beta = c$ where A is a known q*p matrix of rank q and c is a known q*1 vector then with **Lagrange multiplier** we can get

$$\hat{\beta}_H = \hat{\beta} + \left(X'X\right)^{-1}A'\left[A\left(X'X\right)^{-1}A'\right]^{-1}(c - A\hat{\beta})$$

where $\hat{\beta}$ is the estimation without constrain, i.e. $\hat{\beta} = (X'X)^{-1} X'Y$.

Proof let $\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_q \end{pmatrix}$ then we define Lagrange multiplier as

$$\text{Lagrange multiplier} = (\beta' A' - C')\, \lambda$$

let

$$S = \min_{\beta} (Y - X\beta)'(Y - X\beta) + (\beta' A' - c')\, \lambda$$

then

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta + A'\lambda \overset{\text{set}}{=} 0 \Rightarrow (X'X)\, \beta = X'Y - \frac{1}{2}A'\lambda \Rightarrow \hat{\beta}_H = (X'X)^{-1} X'Y - \frac{1}{2}(X'X)^{-1} A'\hat{\lambda_H}$$

suppose $\hat{\beta_H}$ and $\hat{\lambda_H}$ are the solution under constrain $H : A\beta = c$ then we have

$$c = A\hat{\beta}_H = A \underbrace{(X'X)^{-1} X'Y}_{\hat{\beta}\text{ under no constrain}} - \frac{1}{2}A\,(X'X)^{-1} A'\hat{\lambda}_H = A\hat{\beta} - \frac{1}{2}A(XX)^{-1}A'\hat{\lambda}_H$$

since $\operatorname{rank}(A) = q$, $A\,(X'X)^{-1} A'$ is non-singular. Thus

$$-\frac{1}{2}\hat{\lambda}_H = \left[A\,(X'X)^{-1} A'\right]^{-1} (c - A\hat{\beta})$$

therefore

$$\hat{\beta}_H = \hat{\beta} + (X'X)^{-1} A' \left[A\,(X'X)^{-1} A'\right]^{-1} (c - A\hat{\beta})$$

we also need to prove the $\hat{\beta}_H$ minimize $\epsilon'\epsilon$ under constrain

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

where

$$\begin{aligned}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) &= \left(\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta\right)' X'X \left(\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta\right)\\ &= \left(\hat{\beta} - \hat{\beta}_H\right)' X'X \left(\hat{\beta} - \hat{\beta}_H\right) + \left(\hat{\beta}_H - \beta\right)' X'X \left(\hat{\beta}_H - \beta\right) + 2\left(\hat{\beta} - \hat{\beta}_H\right)' X'X \left(\hat{\beta}_H - \beta\right)\\ &= \left(\hat{\beta} - \hat{\beta}_H\right)' X'X \left(\hat{\beta} - \hat{\beta}_H\right) + \left(\hat{\beta}_H - \beta\right)' X'X \left(\hat{\beta}_H - \beta\right)\end{aligned}$$

since

$$\left(\hat{\beta} - \hat{\beta}_H\right)' X'X \left(\hat{\beta}_H - \beta\right) = \left[\frac{1}{2}(X'X)^{-1} A'\hat{\lambda}_H\right]' X'X \left(\hat{\beta}'_H - \beta\right) = \frac{1}{2}\hat{\lambda}'_H A \left(\hat{\beta}_H - \beta\right) = \frac{1}{2}\hat{\lambda}_H(c - c) = 0$$

we get

$$(Y - X\beta)'(Y - X\beta) = (Y - X\hat{\beta})'(Y - X\hat{\beta}) + \left[X\left(\hat{\beta} - \hat{\beta}_H\right)\right]' \left[X\left(\hat{\beta} - \hat{\beta}_H\right)\right] + \left[X\left(\hat{\beta}_H - \beta\right)\right]' \left[X\left(\hat{\beta}_H - \beta\right)\right]$$

all three terms are positive, thus minimum achieved when

$$\left[X\left(\hat{\beta}_H - \beta\right)\right] = 0$$

i.e.

$$\beta = \hat{\beta}_H$$

then we get

$$\left\|\mathbf{Y} - \hat{\mathbf{Y}}_H\right\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \left\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H\right\|^2$$

3.3 Identifiability

The conclusion from this part is applicable under both MLE and LSE cause we only estimate $\hat{\beta}$
A model is identifiable if it is theoretically possible to learn the true values of this model's underlying parameters after obtaining a sample. Mathematically, this is equivalent to saying that different values of the parameters must generate different probability distributions of the observable variables.
Let

$$\boldsymbol{P} = \{\boldsymbol{P}_\theta : \theta \in \Theta\}$$

then

$$\boldsymbol{P}_{\theta_1} = \boldsymbol{P}_{\theta_2} \Longrightarrow \theta_1 = \theta_2$$

for $\forall \theta_1, \theta_2 \in \Theta$
We have $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$. The design matrix X is n*p and is not full rank. So $\hat{\beta}$ is not unique. But $\theta = X\beta$ and $\hat{\theta}$ is unique. Also $\mathrm{RSS} = Y'(I-P)Y$ is unique. We can always have solution to β based on **g inverse**. The other way is that we impose constraints H so that models are identifiable.

$$H\beta = 0$$

let

$$G = \begin{bmatrix} X \\ H \end{bmatrix}$$

then we have

$$\hat{\beta} = (G'G)^{-1} X'Y$$

Proof

$$G = \begin{pmatrix} X_{n\times p} \\ H_{(p-r)\times p} \end{pmatrix} \Rightarrow G' = (X', H') \Rightarrow G'G = X'X + H'H$$

then G^{-1} exists cause G is full rank. We know

$$(X'X)\hat{\beta} = X'Y, \quad H\hat{\beta} = 0$$

thus

$$(G'G - H'H)\hat{\beta} = X'Y \Rightarrow G'G\hat{\beta} = X'Y \Rightarrow \hat{\beta} = (G'G)^{-1} X'Y$$

We have that $\hat{\beta}$ is unbiased.
Proof

$$\begin{aligned} E(\hat{\beta}) = E\left[(G'G)^{-1} X'Y\right] = (G'G)^{-1} X'E(Y) &= (G'G)^{-1} X'X\beta \\ &= (G'G)^{-1}(G'G - H'H)\beta \\ &= \beta - (G'G)^{-1} H' \underbrace{H\beta}_{0} \\ &= \beta \end{aligned}$$

3.4 Estimability

Since $\hat{\beta}$ is not unique, β is not estimable. We consider function of elements of β, i.e. $a'\beta$
Definition The parametric function $a'\beta$ is said to be estimable if it has a linear unbiased estimate, say $b'Y$
This implies that

$$\begin{aligned} &E(b'Y) = b'E(Y) = b'X\beta \quad ,\forall\beta \\ &b'X\beta = a'\beta \quad ,\forall\beta \\ &a' = b'X \text{ or } a = X'b \end{aligned}$$

thus THEOREM 1

$$\begin{aligned} a'\beta \text{ is estimatable } &\Leftrightarrow a = X'b \\ &\Leftrightarrow a \in C[X'] \text{ or} \end{aligned}$$

THEOREM 2 if $a'\beta$ is estimable and $\hat{\beta}$ is any solution of the normal equation then (1) $a'\hat{\beta}$ is unique and (2) $a'\hat{\beta}$ is the BLUE of $a'\beta$

THEOREM 3 $a'\beta$ is estimable if and only if $a'(X'X)^{-} X'X = a'$

Proof
$\Rightarrow a'\beta$ is estimable then

$$\begin{aligned}&\exists c \text{ s.t. } a' = c'X\\ &\Rightarrow a'\left(X'X\right)^{-}X'X = c'X\left(X'X\right)^{-}X'X = c'PX = c'X = a'\end{aligned}$$

$\Rightarrow$ suppose $a'\left(X'X\right)^{-}X'X = a'$ then

$$E\left[a'\hat{\beta}\right] = E\left[a'\left(X'X\right)^{-}X'Y\right] = a'\left(X'X\right)^{-}X'X\beta = a'\beta$$

4 Generalized least square

Having developed a least squares theory for the full-rank model $Y = X\beta+\epsilon$, where $E(\epsilon) = 0$ and $\operatorname{Var}(\epsilon) = \sigma^2 I$, we now consider what modifications are necessary if we allow the ϵ_i to be correlated. In particular, we assume the $\operatorname{Var}[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{V}$, where V is a known n*n positive-definite matrix.

Theorem 4.1. *Under the above setting, we have*

$$\boldsymbol{\beta}^{\star} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$$

$$\begin{aligned}\operatorname{Var}\left[\boldsymbol{\beta}^{\star}\right] &= \sigma^2\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\\ RSS &== \left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^{\star}\right)'V^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^{\star}\right)\end{aligned}$$

Proof $Y = X\beta+\epsilon$, where $\epsilon \sim N(0,\sigma^2 V)$. Since V is a positive-definite

$$\Rightarrow \exists K_{n*n} \text{ s.t. } V = KK', K^{-1} \text{ exist}$$

then write

$$\underbrace{k^{-1}Y}_{Z} = \underbrace{K^{-1}X}_{B}\beta + \underbrace{K^{-1}\varepsilon}_{\eta} \Rightarrow z = B\beta+\eta$$

, where $E(\eta) = 0, \operatorname{Var}(\eta) = \sigma^2 I$ since

$$E(\eta) = E\left(K^{-1}\varepsilon\right) = 0 \quad ;\operatorname{Var}(\eta) = \operatorname{Var}\left(K^{-1}\varepsilon\right) = K^{-1}\operatorname{Var}(\varepsilon)\left(K^{-1}\right)' = K^{-1}\sigma^2 KK'\left(K'\right)^{-1} = \sigma^2 I$$

then

$$\beta^* = \left(B'B\right)^{-1}B'Z = \left(X'K^{-1\prime}K^{-1}X\right)^{-1}X'K^{-1\prime}K^{-1}Y = \left(X'V^{-1}X\right)^{-1}X'V^{-1}Y$$

then

$$E\left(\beta^*\right) = \left(X'V^{-1}X\right)^{-1}X'V^{-1}X\beta = \beta \quad \text{unbiased}$$

$$\begin{aligned}\operatorname{Var}\left(\beta^*\right) = \left(X'V^{-1}X\right)^{-1}X'V^{-1}\operatorname{Var}(Y)\left[\left(X'V^{-1}X\right)^{-1}X'V^{-1}\right]' &= \left(X'V^{-1}X\right)^{-1}X'V^{-1}\sigma^2 V \quad \left[\left(X'V^{-1}X\right)^{-1}X'V^{-1}\right]'\\ &= \left(X'V^{-1}X\right)^{-1}\sigma^2\end{aligned}$$

$$E\left[\boldsymbol{\beta}^{\star}\right] = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$E\left[\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^{\star}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^{\star}\right)\right] = (n-p)\sigma^2$$

$$\begin{aligned}RSS = \left(Z-B\beta^*\right)'\left(Z-B\beta^*\right) &= \left(K^{-1}Y-K^{-1}X\beta^*\right)'\left(K^{-1}Y-K^{-1}X\beta^*\right)\\ &= \left(Y-X\beta^*\right)'\left(K^{-1}\right)'K^{-1}\left(Y-X\beta^*\right)\\ &= \left(Y-X\beta^*\right)'V^{-1}\left(Y-X\beta^*\right)\end{aligned}$$

Alternative method for deriving the β is minimizing $\eta'\eta$ with respect to β

$$\begin{aligned}\boldsymbol{\eta}'\boldsymbol{\eta} &= \boldsymbol{\epsilon}'\boldsymbol{V}^{-1}\boldsymbol{\epsilon}\\ &= (\boldsymbol{Y}-\boldsymbol{X}\beta)'V^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta)\\ &= \boldsymbol{Y}'\boldsymbol{V}^{-1}\boldsymbol{Y} - 2\beta'\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} + \beta'\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\beta\end{aligned}$$

$$\frac{\partial\boldsymbol{\eta}'\boldsymbol{\eta}}{\partial\boldsymbol{\beta}} = -\mathbf{2}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} \overset{\text{set}}{=} 0$$

we also get

$$\boldsymbol{\beta}^{\star}=\left(\boldsymbol{X}^{\prime} \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\prime} \boldsymbol{V}^{-1} \boldsymbol{Y}$$

A special case is that when

$$\boldsymbol{V}=\operatorname{diag}\left(w_1^{-1}, w_2^{-1}, \cdots, w_n^{-1}\right)\left(w_i>0\right)$$

we have

$$\boldsymbol{\beta}^{\star}=\frac{\sum_i w_i Y_i x_i}{\sum_i w_i x_i^2}$$

$$\left(X^{\prime} V^{-1} X\right)^{-1}=\left(x^{\prime} V^{-1} x\right)^{-1}=\left(\sum w_i x_i^2\right)^{-1}$$

Theorem 4.2. *$a'\beta^*$ is the best linear unbiased estimate (BLUE) of $a'\beta$ under the generalized linear model.*

Proof

$$a^{\prime} \beta^*=\underbrace{a^{\prime}\left(X^{\prime} V^{-1} X\right)^{-1} X^{\prime} V^{-1}}_{b^{\prime}} Y=b^{\prime} Y$$

is lienar in Y and also unbiased. Let $b'_{-1}Y$ be another unbiased linear estimator of $a'\beta$

$$a^{\prime} \beta^*=\alpha^{\prime}\left(B^{\prime} B\right)^{-1} B^{\prime} Z$$

$$b_1^{\prime} Y=b_1^{\prime} K K^{-1} Y=\left(K^{\prime} b_1\right)^{\prime} Z$$

by previous theorem BLUE:

$$\operatorname{Var}\left(a^{\prime} \beta^*\right) \leqslant \operatorname{Var}\left(\left(K^{\prime} b_1\right)^{\prime} Z\right)=\operatorname{Var}\left(b_1^{\prime} Y\right)$$

equality holds if and only if

$$\left(K^{\prime} b_1\right)^{\prime}=a^{\prime}\left(B^{\prime} B\right)^{-1} B^{\prime} \Rightarrow b_1^{\prime} K=a^{\prime}\left(B^{\prime} B\right)^{-1} B^{\prime} \Rightarrow b_1^{\prime}=a^{\prime}\left(B^{\prime} B\right)^1 B^{\prime} K^{-1}=a^{\prime}\left(X^{\prime} V^{-1} X\right)^{-1} X^{\prime} V^{-1}=b^{\prime}$$

5 Hypothesis Testing

We are interested in the form of $H_0 = A\beta = 0$

5.1 Likelihood ratio test

$$\begin{aligned}
&G: Y_i=\beta_0+\beta_1 x_{i 1}+\cdots+\beta_{p-1} x_{i, p-1}+\epsilon_i, \text{ full model} \\
&H: Y_i=\beta_0+\beta_1 x_{i 1}+\cdots+\beta_{r-1} x_{i, r-1}+\epsilon_i, \text{ reduced model}
\end{aligned}$$

$$\text{Hypothesis of interest is } H_0: \beta_r=\beta_{r+1}=\ldots\ldots=\beta_{p-1}=0$$

Given the linear model $G: Y = X\beta + \epsilon$, where X is n*p of rank p and $\epsilon \sim N_n\left(0, \sigma^2 I_n\right)$, we wish to test the hypothesis $H_0 : \mathbf{A}\beta = 0$, where A is q*p of rank q. the likelihood function for G is

$$L\left(\boldsymbol{\beta}, \sigma^2\right)=\left(2 \pi \sigma^2\right)^{-n / 2} \exp \left[-\frac{1}{2 \sigma^2}\|\boldsymbol{Y}-\boldsymbol{X} \boldsymbol{\beta}\|^2\right]$$

under the MLE

$$\hat{\sigma}^2=\|\boldsymbol{Y}-\boldsymbol{X} \hat{\boldsymbol{\beta}}\|^2 / n$$

then the likelihood becomes

$$L\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right)=\left(2 \pi \hat{\sigma}^2\right)^{-n / 2} e^{-n / 2}$$

let $\hat{\beta}_H$ and $\hat{\sigma^2}_H$ be the estimator of $Y = X\beta + \epsilon$ when $\mathbf{A}\beta = 0$, then under the hypothesis

$$L\left(\hat{\boldsymbol{\beta}_{\boldsymbol{H}}}, \hat{\sigma}_H^2\right)=\left(2 \pi \hat{\sigma}_H^2\right)^{-n / 2} e^{-n / 2}$$

then the likelihood ratio test of H is given by

$$\Lambda=\frac{L\left(\hat{\beta_H}, \hat{\sigma}_H^2\right)}{L\left(\hat{\beta}, \hat{\sigma}^2\right)}=\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_H^2}\right)^{n / 2}$$

We have learned that $-2log\Lambda$ has a chi-squared distribution.

5.2 F test

In summary

$$F=\frac{n-p}{q}\left(\Lambda^{-2/n}-1\right)$$

has an $F_{q,n-p}$ distribution when H_0 is true. And we also have

- $$RSS_H-RSS=\left\|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_{\boldsymbol{H}}\right\|^2=(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{c})'\left[\boldsymbol{A}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}-\boldsymbol{c})$$

- $$\begin{aligned}E\left[RSS_H-RSS\right]&=\sigma^2q+(\mathbf{A}\beta-\boldsymbol{c})'\left[\boldsymbol{A}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{A}\beta-\boldsymbol{c})\\&=\sigma^2q+(RSS_H-RSS)_{Y=E[\boldsymbol{Y}]}\end{aligned}$$

- $$F=\frac{\overbrace{\left(RS\widetilde{S}_H-RSS\right)}^{\text{improvement}}/q}{RSS/(n-p)}=\frac{(A\hat{\beta}-c)'\left[A\left(X'X\right)^{-1}A'\right]^{-1}(A\hat{\beta}-c)}{qS^2}$$

  is distributed as $F_{q,n-p}$ (the F-distribution with q and n-p degrees of freedom, respectively)

- when $c=0$, F can be expressed in the form

  $$F=\frac{n-p}{q}\frac{Y'\left(P-P_H\right)Y}{Y'\left(I_n-P\right)Y}$$

  where P_H is symmetric and idempotent and $\boldsymbol{P}_H\boldsymbol{P}=\boldsymbol{P}\boldsymbol{P}_H=\boldsymbol{P}_H$

Proof $y=X\beta+\varepsilon$, where $\operatorname{rank}(X)=p,\ \operatorname{rank}(A)=q$ and $H_0:\ A\beta=c$. Under the constrain, $H_0:\ A\beta=c$ we know

$$\hat{\beta}_H=\hat{\beta}+\left(X'X\right)^{-1}A'\left[A\left(X'X\right)^{-1}A'\right]^{-1}(c-A\hat{\beta})\tag{1}$$

also since we have $\left\|\mathbf{Y}-\hat{\mathbf{Y}}_H\right\|^2=\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2+\left\|\hat{\mathbf{Y}}-\hat{\mathbf{Y}}_H\right\|^2$ we get

$$RSS_H=RSS+\left(\hat{\beta}-\hat{\beta}_H\right)'X'X\left(\hat{\beta}-\hat{\beta}_H\right)$$

from 1 we know that

$$\hat{\beta}_H-\hat{\beta}=\left(X'X\right)^{-1}A'\left[A\left(X'X\right)^{-1}A'\right]^{-1}(c-A\hat{\beta})$$

and

$$\begin{aligned}RSS_H-RSS&=\left(\hat{\beta}-\hat{\beta}_H\right)'X'X\left(\hat{\beta}-\hat{\beta}_H\right)\\&=(c-A\hat{\beta})'\left[A(XX)^{-1}A'\right]^{-1}A(XX)^{-1}(XX)(XX)^{-1}A'\left[A\left(X'X\right)^{-1}A'\right]^{-1}(c-A\hat{\beta})\\&=(c-A\hat{\beta})'\left[A(X'X)^{-1}A'\right]^{-1}(c-A\hat{\beta})\\&=(A\hat{\beta}-c)'\left[A(XX)^{-1}A'\right]^{-1}(A\hat{\beta}-c)\end{aligned}$$

the first equation is proved
We also know that

$$A\hat{\beta}\sim N(A\beta,\ \underbrace{A\left(X'X\right)^{-1}A'}_{B}\sigma^2)$$

let $Z=\hat{A\beta}-c$ then $\operatorname{Var}(Z)=B\sigma^2$ we have

$$\begin{aligned}E\left(RSS_H-RSS\right)&=E\left[(A\hat{\beta}-c)'\left[A(XX)^{-1}A'\right]^{-1}(A\hat{\beta}-c)\right]=E\left[Z'B^{-1}Z\right]\\&=\operatorname{tr}(\sigma^2B^{-1}B)+(A\beta-c)'B^{-1}(A\beta-c)\\&=\operatorname{tr}(\sigma^2I_{q\times q})+(A\beta-c)'B^{-1}(A\beta-c)\\&=\sigma^2q+(A\beta-c)'\left[A(X'X^{-1})A'\right]^{-1}(A\beta-c)\end{aligned}$$

the second equation is proved
$H_0:\ A\beta = c$, under H_0, $A\hat{\beta} \sim N\left(c = A\beta,\ \sigma^2 A\left(X'X\right)^{-1}A'\right)$ thus

$$(A\hat{\beta} - c)'\left[\sigma^2 A\left(X'X\right)^{-1}A'\right]^{-1}(A\hat{\beta} - c) = \frac{RSSH - RSS}{\sigma^2} \sim \chi^2_q$$

we also have

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}$$

and previously we learned that RSS is independent of $\hat{\beta}$, thus $\text{RSS}_H - \text{RSS} \perp \text{RSS}$. Therefore,

$$\frac{\text{RSS}_{\ H} -\ \text{RSS}\ /(\sigma^2/q) \sim \chi^2_q}{\text{RSS}\ /(\sigma^2/(n-p)) \sim \chi^2_{n-p}} >\ \text{independence} \implies \sim \text{F}_{q,\,n-p}$$

the third equation is proved
when $c = 0$,

$$\begin{aligned}\hat{Y}_H = X\hat{\beta}_H &= X\left[\hat{\beta} + \left(X'X\right)^{-1}A'\left[A\left(X'X\right)^{-1}A'\right]^{-1}(c - A\hat{\beta})\right] \\ &= PY - \underbrace{X\left(X'X\right)^{-1}A'\left[A\left(X'X\right)^{-1}A'\right]^{-1}A\left(X'X\right)^{-1}X'Y}_{P_1} \\ &= \underbrace{(P - P_1)}_{P_H}Y\end{aligned}$$

here P_H is symmetric since we know P_1 is symmetric and idempotent. We can also show $P_1P = PP_1 = P_1$. Further, P_H is idempotent

$$\begin{aligned}P_1^2 = \left(P - P_1\right)\left(P - P_1\right) &= P^2 - P_1P - PP_1 + P_1^2 \\ &= P - 2P_1 + P_1 = P - P_1 = P_H\end{aligned}$$

also $PP_H = P_H$ and

$$\begin{aligned}RSS_H &= \left\|Y - X\hat{\beta}_H\right\|^2 = \left\|Y - P_HY\right\|^2 = Y'\left(I - P_H\right)Y \\ \text{RSS } &= Y'(I - P)Y\end{aligned}$$

6 Some comments about ANOVA

Consider one-way ANOVA

		simple mean
trt 1	$Y_{11}, Y_{12}, \ldots, Y_{1,J_1}$	$\bar{y}_1$
trt 2	$Y_{21}, Y_{22}, \ldots, Y_{2,J_2}$	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$
trt I	$Y_{I1}, Y_{I2}, \ldots, Y_{I,J_I}$	$\bar{y}_I$

the model is $Y_{ij} = \mu_i + \varepsilon_{ij}\quad,\quad (i = 1, 2, \ldots, I, j = 1, \ldots, J_i)$, $\beta = \begin{pmatrix}\mu_1 \\ \mu_2 \\ \vdots \\ \mu_I\end{pmatrix}$, assume $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$, then

we can write as $\begin{vmatrix} y_{11} \\ \vdots \\ y_{1,J_1} \\ y_{21} \\ \vdots \\ y_{2,J_2} \\ \vdots \\ y_{I1} \\ \vdots \\ y_{I,J_I}\end{vmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ & & & \vdots & \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ & & & \vdots & \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & 1\end{bmatrix}\begin{bmatrix}\mu_1 \\ \vdots \\ \mu_2 \\ \mu_I\end{bmatrix}, \Rightarrow X = \begin{bmatrix}\mathbb{1}_{J_1} & \mathbb{0}_{J_1} & \cdots & \mathbb{0}_{J_1} \\ \mathbb{0}_{J_2} & \mathbb{1}_{J_2} & \cdots & \mathbb{0}_{J_2} \\ & & \vdots & \\ \mathbb{0}_{J_I} & \mathbb{0}_{J_I} & \cdots & \mathbb{1}_{J_I}\end{bmatrix}$. To test hypothesis, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ (testable since X is full rank) $A\beta = 0 \quad \Leftrightarrow \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ 0 & 0 & 1 & -1 & \cdots \\ 0 & & & & \\ & & \vdots & & \\ & & & & \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix} \beta = 0.$

Then proceed with usual F-test $F = \frac{(\mathrm{RSS_H - RSS})/(\mathrm{I}-1)}{\mathrm{RSS}/(\mathrm{n-I})}$. **ALTERNATIVELY**, we can write $S = \sum_i \sum_j (y_{ij} - \mu_i)^2$, where S is the error sum of squared ϵ_{ij}^2, $\frac{\partial S}{\partial \mu_i} = 0 \Rightarrow \sum_j 2(y_{ij} - \mu_i) = 0 \Rightarrow \mu_i = \frac{\sum_j y_{ij}}{J_i} = \overline{y_i}$, RSS $= \sum_i \sum_j (y_{ij} - \hat{\mu}_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ and under H_0 $S = \sum_i \sum_j (y_{ij} - \mu)^2 -$ error sum of squared , $\frac{\partial S}{\partial \mu} = 0 \Rightarrow \sum_i \sum_j (y_{ij} - \mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum_i \sum_j (y_{ij})}{\sum_i J_i} = \bar{Y} \leftarrow$ overall mean . Thus $RSS_H = \sum_i \sum_j (y_{ij} - \hat{\mu})^2 = \sum_i \sum_j \left(y_{ij} - \bar{Y}\right)^2$. Then $RSS_H - RSS = \sum_i \sum_j (\hat{y}_{ij} - \hat{y}_H)^2 = \sum_i \sum_j \left(\bar{y}_i - \bar{Y}\right)^2 = \sum_i J_i \left(\bar{y}_i - \bar{Y}\right)^2$. Therefore, using the F-test $F = \frac{\sum_i J_i \left(\bar{y}_i - \bar{Y}\right)/(I-1)}{\sum_i \sum_j (y_i - \bar{y}_i)^2/(n-I)} \quad, p = I, q = I - 1$