Types of Problems—Review for Some of the QUAL problems

Notation: Let $A^T = A'$ be the transpose of $A$.

0) Covariance and Expected Value = Mean, and the Multivariate Normal (MVN) Distribution:

Notation: Unless told otherwise, assume expectations exist and that conformable matrices and vectors are used.

The population mean of a random $n \times 1$ vector $x = (x_1, ..., x_n)^T$ is $E(x) = \mu = (E(x_1), ..., E(x_n))^T$ and the $n \times n$ population covariance matrix $\text{Cov}(x) = \Sigma_x = E((x - E(x))(x - E(x))^T) = (\sigma_{i,j})$ where $\text{Cov}(x_i, x_j) = \sigma_{i,j}$. The population covariance matrix of $x$ with $y$ is

$$\text{Cov}(x, y) = \Sigma_{x,y} = E[(x - E(x))(y - E(y))^T].$$

If $X$ and $Y$ are $n \times 1$ random vectors, $a$ a conformable constant vector, and $A$ and $B$ are conformable constant matrices, then

$$E(X + Y) = E(X) + E(Y), \ E(a + Y) = a + E(Y), \ & E(AXB) = AEXB.$$ 

Also

$$\text{Cov}(a + AX) = \text{Cov}(AX) = A\text{Cov}(X)A^T.$$ 

Note that $E(AY) = AE(Y)$ and $\text{Cov}(AY) = A\text{Cov}(Y)A^T$.

If $X$ ($m \times 1$) and $Y$ ($n \times 1$) are random vectors, and $A$ and $B$ are conformable constant matrices, then

$$\text{Cov}(AX, BY) = A\text{Cov}(X,Y)B^T.$$ 

If $X \sim N_p(\mu, \Sigma)$, then $E(X) = \mu$, $\text{Cov}(X) = \Sigma$, and $m_X(t) = \exp(t^T\mu + \frac{1}{2}t^T\Sigma t)$.

If $X \sim N_p(\mu, \Sigma)$ and if $A$ is a $q \times p$ matrix, then $AX \sim N_q(A\mu, A\Sigma A^T)$. If $a$ ($p \times 1$) and $b$ ($q \times 1$) are constant vectors, then $X + a \sim N_p(\mu + a, \Sigma)$ and $AX + b \sim N_q(A\mu + b, A\Sigma A^T)$.

Let $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.

The conditional distribution of a MVN is MVN. If $X \sim N_p(\mu, \Sigma)$, then the conditional distribution of $X_1$ given that $X_2 = x_2$ is multivariate normal with mean $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. That is,

$$X_1|X_2 = x_2 \sim N_q(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Notation:

$$X_1|X_2 \sim N_q(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

1) Projection Matrices, Generalized Inverses, and the Column Space $C(X)$:

Let $A = [a_1 \ a_2 \ ... \ a_m]$ be an $n \times m$ matrix. The space spanned by the columns of $A$ = column space of $A = C(A)$. 

1
Let \( X = [v_1 \ v_2 \ldots \ v_p] \) be an \( n \times p \) matrix. Then
\[
C(X) = \{ y \in \mathbb{R}^n : y = X\beta \text{ for some } \beta \in \mathbb{R}^p \}.
\]
One way to show \( C(A) = C(B) \) is to show that i) \( Ax = By \in C(B) \) and ii) \( By = Ax \in C(A) \).

The null space of \( A = N(A) = \{ x : Ax = 0 \} \) is the orthogonal complement of \( V \).
\[
N(A^T) = [C(A)]^\perp, \text{ so } N(A) = [C(A^T)]^\perp.
\]

A generalized inverse of an \( m \times n \) matrix \( A \) is any \( n \times m \) matrix \( A^\dagger \) satisfying \( AA^\dagger A = A \). Other names are conditional inverse, pseudo inverse, g-inverse, and p-inverse. Usually a generalized inverse is not unique, but if \( A^{-1} \) exists, then \( A^\dagger = A^{-1} \) is unique. Notation: \( G := A^\dagger \) means \( G \) is a generalized inverse of \( A \).

Let \( V \) be a subspace of \( \mathbb{R}^k \). Then every \( y \in \mathbb{R}^k \) can be expressed uniquely as \( y = w + z \) where \( w \in V \) and \( z \in V^\perp \).

Let \( X = [v_1 \ v_2 \ldots \ v_p] \) be \( n \times p \), and let \( V = C(X) = \text{span}(v_1, \ldots, v_p) \). Then the \( n \times n \) matrix \( P_V = P_X \) is a projection matrix on \( C(X) \) if \( P_X y = w \forall y \in \mathbb{R}^n \). (Here \( y = w + z = wy + zy \), so \( w \) depends on \( y \).)

**Projection Matrix Theorem:**

a) \( P_X \) is unique.
b) \( P_X = X(X^TX)^+X^T \) where \( (X^TX)^+ \) is any generalized inverse of \( X^TX \).
c) \( A \) is a projection matrix on \( C(A) \) iff \( A \) is symmetric and idempotent. Hence \( P_X \) is a projection matrix on \( C(P_X) = C(X) \).
d) \( I_n - P_X \) is the projection matrix on \( [C(X)]^\perp \).
e) \( A = P_X \) iff i) \( y \in C(X) \) implies \( Ay = y \) and ii) \( y \perp C(X) \) implies \( Ay = 0 \).
f) \( P_X X = X \) and \( P_X W = W \) if each column of \( W \in C(X) \).
g) \( P_X v_i = v_i \).
h) If \( C(X_R) \) is a subspace of \( C(X) \), then \( P_X P_X_R = P_X_R P_X = P_X_R \).
i) \( \text{rank}(P_X) = \text{tr}(P_X) = \text{rank}(X) \).

Note that \( P \) is a projection matrix iff \( P \) is symmetric and idempotent. Partition \( X \) as \( X = [X_1 \ X_2] \), let \( P \) be the projection matrix for \( C(X) \) and let \( P_1 \) be the projection matrix for \( C(X_1) \). Since \( C(P_1) = C(X_1) \subseteq C(X) \), \( PP_1 = P_1 \). Hence \( P_1P = (PP_1)' = P_1' = P_1 \).

1a): Given small \( X \), be able to find the projection matrix \( P \) for \( C(X) \).

1b): Given small \( X \), be able to find \( \text{rank}(X) \), a basis for \( C(X) \), and \( [C(X)]^\perp = \text{nullspace of } X^T \).

1c): Be able to show that \( G := A^{-1} \).

2) **Quadratic Forms** \( Y'AY \) and terms like \( AY \):

The matrix \( A \) in a quadratic form \( x^TAx \) is symmetric. \( A \) is positive definite \( (A > 0) \) if \( x^TAx > 0 \forall x \neq 0 \). \( A \) is positive semidefinite \( (A \geq 0) \) if \( x^TAx \geq 0 \forall x \).

Let \( A \) be symmetric. If \( A \geq 0 \) then the eigenvalues of \( A \) are real and nonnegative. If \( A \geq 0, \) let \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0 \). If \( A > 0, \) then \( \lambda_n > 0 \).

**Theorem 2.5 (Seber and Lee Th. 1.5):** expected value of a quadratic form:

Let \( X \) be a random vector with \( E(X) = \mu \) and \( \text{Cov}(X) = \Sigma \). Then
\[
E(X^TAX) = tr(AS) + [E(X)]^TAE(X) = tr(AS) + \mu^TA\mu.
\]
Theorems 2.6 and 2.7: If $AY \parallel BY$, then $f(AY) \parallel g(BY)$ where $f$ and $g$ are functions (such that $f(AY)$ only depends on $A$ and $AY$ and $g(BY)$ only depends on $B$ and $BY$). Note that $Y'AY = Y'A^rA^r Y = f(AY)$ (for a quadratic form $A$ is symmetric). $Y'(I - P)Y = \|(I - P)Y\|^2$, and $Y'PY = \|PY\|^2$ where the squared Euclidean norm $\|Z\|^2 = Z'Z$.

Theorem 2.8. Let $Y \sim N_n(\mu, \Sigma)$. a) Let $u = AY$ and $w = BY$. Then $AY \parallel BY$ iff $\text{Cov}(u, w) = A\Sigma B^T = 0$ iff $B\Sigma A^T = 0$. Note that if $\Sigma = \sigma^2 I_n$, then $AY \parallel BY$ if $AB^T = 0$ if $BA^T = 0$.

b) If $A$ is a symmetric $n \times n$ matrix, and $B$ is an $m \times n$ matrix, then $Y^TAY \parallel BY$ iff $A\Sigma B^T = 0$ iff $B\Sigma A = 0$.

Craig’s Theorem: Let $Y \sim N_n(\mu, \Sigma)$.

a) If $\Sigma > 0$, then $Y^TAY \parallel Y^TBY$ iff $A\Sigma B = 0$ iff $B\Sigma A = 0$.

b) If $\Sigma \geq 0$, then $Y^TAY \parallel Y^TBY$ if $A\Sigma B = 0$ (or if $B\Sigma A = 0$).

c) If $\Sigma \geq 0$, then $Y^TAY \parallel Y^TBY$ iff

(*) \[ A\Sigma B \Sigma = 0, \Sigma A\Sigma B \mu = 0, \Sigma B\Sigma A \mu = 0, \text{ and } \mu^T A\Sigma B \mu = 0. \]

Note that if $A\Sigma B = 0$, then $(*)$ holds.

Theorem 2.13. If $Y \sim N_n(\mu, \Sigma)$ where $\Sigma > 0$, then $Y^TAY \sim \chi^2(\text{rank}(A), \mu^T A \mu / 2)$ iff $A\Sigma$ is idempotent.

Remark 1: If the theorem is for $Y \sim N_n(\mu, I)$ and $Z \sim N_n(E(Z), \sigma^2 I)$, then use $Y = Z / \sigma \sim N_n(\mu = E(Z) / \sigma, I)$.

Theorem 2.14. Let $A = A^T$ be symmetric.

a) If $Y \sim N_n(0, \Sigma)$ where $\Sigma$ is a projection matrix, then $Y^TAY \sim \chi^2(\text{rank}(\Sigma))$ where $\text{rank}(\Sigma) = tr(\Sigma)$.

b) If $Y \sim N_n(0, I)$, then $Y^TAY \sim \chi_r^2$ if $A$ is idempotent with $\text{rank}(A) = tr(A) = r$.

c) Let $Y \sim N_n(0, \sigma^2 I)$. Then

\[ \frac{Y^TAY}{\sigma^2} \sim \chi_r^2 \text{ or } Y^TAY \sim \sigma^2 \chi_r^2 \]

iff $A$ is idempotent of rank $r$.

d) If $Y \sim N_n(0, \Sigma)$ where $\Sigma > 0$, then $Y^TAY \sim \chi_r^2$ if $A\Sigma$ is idempotent with $\text{rank}(A) = r = \text{rank}(A\Sigma)$.

e) If $Y \sim N_n(0, \sigma^2 I)$ then \[ \frac{Y^TY}{\sigma^2} \sim \chi^2 \left( n, \frac{\mu^T \mu}{2 \sigma^2} \right). \]

f) If $Y \sim N_n(\mu, I)$ then $Y^TAY \sim \chi^2(r, \mu^T A \mu / 2)$ iff $A$ is idempotent with $\text{rank}(A) = tr(A) = r$.

g) If $Y \sim N_n(\mu, \sigma^2 I)$ then \[ \frac{Y^TAY}{\sigma^2} \sim \chi^2 \left( r, \frac{\mu^T A \mu}{2 \sigma^2} \right) \]

iff $A$ is idempotent with $\text{rank}(A) = tr(A) = r$.

3) MLE: The following problem is typical. It is assumed than $\sigma > 0$ and $\beta \in \mathbb{R}^p$.

Suppose $Y_i = x_i^T \beta + \epsilon_i$ with $Q(\beta) \geq 0$. Let $c_n$ be a constant that does not depend on $\beta$ or $\sigma^2$. Suppose the likelihood function is

\[ L(\beta, \sigma^2) = c_n \frac{1}{\sigma^n} \exp \left( \frac{-1}{2 \sigma^2} Q(\beta) \right). \]
a) Suppose that $\hat{\beta}_Q$ minimizes $Q(\beta)$. Show that $\hat{\beta}_Q$ is the MLE of $\beta$.

b) Then find the MLE $\hat{\sigma}^2$ of $\sigma^2$.

Solution: a) For fixed $\sigma > 0$, $L(\beta, \sigma^2)$ is maximized by minimizing $Q(\beta) \geq 0$. So $\hat{\beta}_Q$ maximizes $L(\beta, \sigma^2)$ regardless of the value of $\sigma^2 > 0$. So $\hat{\beta}_Q$ is the MLE.

b) Let $Q = Q(\hat{\beta}_Q)$. Then the MLE $\hat{\sigma}^2$ is found by maximizing the profile likelihood,

$$L_p(\sigma^2) = L(\hat{\beta}_Q, \sigma^2) = c_n \frac{1}{\sigma^n} \exp \left( -\frac{1}{2\sigma^2} Q \right).$$

Let $\tau = \sigma^2$. The $L_p(\tau) = c_n \frac{1}{\tau^{n/2}} \exp \left( -\frac{1}{2\tau} Q \right)$, and the log profile likelihood $\log L_p(\tau) = d - \frac{n}{2} \log(\tau) - \frac{Q}{2\tau}$. Thus

$$\frac{d}{d\tau} \log L_p(\tau) = -\frac{n}{2\tau} + \frac{Q}{2\tau^2} = 0$$

or $-n\tau + Q = 0$ or $\hat{\tau} = \hat{\sigma}^2 = Q/n$, unique. Then

$$\frac{d^2}{d\tau^2} \log L_p(\tau) = \frac{n}{2\tau^2} - \frac{2Q}{2\tau^3} \Bigg|_{\hat{\tau}} = \frac{n}{2\tau^2} - \frac{2n\hat{\tau}}{2\hat{\tau}^3} = \frac{n}{2\hat{\tau}^2} < 0$$

which proves that $\hat{\sigma}^2$ is the MLE of $\sigma^2$.

Note: A negative second derivative shows that $\hat{\sigma}^2$ is a local max. The result that $\hat{\sigma}^2$ was the unique solution to setting the first derivative of the profile likelihood equal to zero makes $\hat{\sigma}^2$ the global max.

Common errors: Students use $Q(\beta)$ instead of $Q(\hat{\beta})$ in the profile likelihood. Students forget to write the word “unique.”

Variant: $Q(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)^T(Y - X\beta)$ is the least squares criterion. Recognize that $Q(\beta)$ is minimized by $\hat{\beta} = \beta_{OLS}$, and proceed as in the above problem.

Note: If the $e_i$ are iid $N(0, \sigma^2)$ and least squares is used, then the MLE of $\beta$ is the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the MLE of $\sigma^2$ is

$$\hat{\sigma}^2_M = \frac{n-p}{n} \text{MSE} = \frac{1}{n} \sum_{i=1}^n r_i^2.$$  

4) LS Estimators for $p \leq 2$:

Given a least squares model with $p \leq 2$, derive or find the least squares estimator $\hat{\beta}$.

**Tip:** If the LS model is $Y_i = x_i^T \beta + e_i$ for $i = 1, ..., n$, then the LS criterion is $Q(\beta) = \sum_{i=1}^n (Y_i - x_i^T \beta)^2 = \sum_{i=1}^n r_i^2(\beta)$.

To derive the LS estimator, let $Q(\beta_1, \beta_2) = \sum_{i=1}^n (Y_i - \beta_1 - \beta_2 x_i)^2$ be the residual sum of squares where $\beta_i$ vary on $\mathbb{R}$. Take the partial derivatives, set them to 0, and solve for the least squares estimators. If $p = 2$, we will assume 2nd derivatives do not need to be taken. If $p = 1$, show the solution is unique and show that the second derivative evaluated at $\hat{\beta}$ is positive. The $\beta_i$ could be replaced by other symbols such as $\eta_i$.

**Location model:** $Y_i = \beta + e_i$ or $Y = 1\beta + e$. The parameter $\beta$ could be replaced with $\mu$ or $\theta$. The LS criterion $Q(\beta) = \sum_{i=1}^n (Y_i - \beta)^2$, and $\hat{\beta} = \bar{Y}$, the sample mean.

Proof: $\frac{dQ(\beta)}{d\beta} = -2 \sum_{i=1}^n (Y_i - \beta)$. 

4
Setting the derivative equal to 0 and calling the unique solution $\hat{\beta}$ gives $\sum_{i=1}^{n} Y_i = n\hat{\beta}$ or $\hat{\beta} = \bar{Y}$. The second derivative

$$\frac{d^2Q(\beta)}{d\beta^2} = 2n > 0,$$

hence $\hat{\beta}$ is the global minimizer.

**Simple linear regression (SLR):** $Y_i = \beta_1 + x_i\beta_2 + e_i$ or $Y = X\beta + e$ where $X = [1 \ x]$ and $\beta = (\beta_1 \ \beta_2)^T$. The LS criterion $Q(\beta_1, \beta_2) = \sum_{i=1}^{n}(Y_i - \beta_1 - x_i\beta_2)^2$.

The least squares (OLS) line is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where the slope

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n}k_iY_i$$

with

$$k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{X_i - \bar{X}}{(n-1)S_X^2},$$

and the intercept $\hat{\beta}_1 \equiv \hat{\alpha} = \bar{Y} - \beta_2\bar{X}$.

By the chain rule,

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_1 - \beta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \beta_1^2} = 2n.$$

Similarly,

$$\frac{\partial Q}{\partial \beta_2} = -2\sum_{i=1}^{n}X_i(Y_i - \beta_1 - \beta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \beta_2^2} = 2\sum_{i=1}^{n}X_i^2.$$

Setting the first partial derivatives to zero and calling the solutions $\hat{\beta}_1$ and $\hat{\beta}_2$ shows that the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy the **normal equations**:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \quad \text{and} \quad $$

$$\sum_{i=1}^{n} X_iY_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2.$$

The first equation gives $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$.

There are several equivalent formulas for the slope $\hat{\beta}_2$.

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}X_iY_i - \frac{1}{n}(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{\sum_{i=1}^{n}X_i^2 - \frac{1}{n}(\sum_{i=1}^{n}X_i)^2}$$
\[
= \frac{\sum_{i=1}^{n} (X_i - \bar{X})Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_iY_i - nX\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \hat{\rho} s_Y / s_X.
\]

Here the sample correlation \(\hat{\rho} \equiv \hat{\rho}(X,Y) = \text{corr}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_X s_Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2}\)

where the sample standard deviation

\[
s_W = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2}
\]

for \(W = X, Y\). Notice that the term \(n - 1\) that occurs in the denominator of \(\hat{\rho}, s_Y^2\), and \(s_X^2\) can be replaced by \(n\) as long as \(n\) is used in all 3 quantities.

**SLR through the origin**: \(Y_i = x_i \beta + e_i\) or \(Y = x \beta + e\). The LS criterion

\(Q(\beta) = \sum_{i=1}^{n} (Y_i - x_i \beta)^2\), and \(\beta = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2\).

**Known intercept**: \(Y_i = a + x_i \beta + e_i\) where the intercept \(a\) is known.

\(Q(\beta) = \sum_{i=1}^{n} (Y_i - a - x_i \beta)^2\).

**Known slope**: \(Y_i = \beta + x_i b + e_i\) where the slope \(b\) is known.

\(Q(\beta) = \sum_{i=1}^{n} (Y_i - \beta - x_i b)^2\). Here, \(\beta\) may be replaced by \(a\).

5) WLS:

For the WLS model \(Y | x = x^T \beta + e\) where the \(e_i\) are independent with \(E(e_i) = 0\) and \(V(e_i) = \sigma_i^2\). Hence \(Y = Y | X = X \beta + e\) where \(E(e) = 0\) and \(\text{Cov}(e) = \text{diag}(\sigma_i^2)\).

An alternative model is \(Y | x^T \beta = x^T \beta + u\) where the \(u_i\) are independent with \(E(u_i) = 0\) and \(V(u_i) = \tau_i^2\). Hence \(Y = Y | X \beta = X \beta + u\) where \(E(u) = 0\) and \(\text{Cov}(u) = \text{diag}(\tau_i^2)\).

6) Non-full rank linear models:

The nonfull rank linear model is \(Y = X \beta + e\) where \(X\) has rank \(r < p \leq n\), and \(X\) is an \(n \times p\) matrix.

**Theorem 3.1.** i) \(P = X (X^T X)^{-} X^T\) is the unique projection matrix on \(C(X)\) and does not depend on the generalized inverse \((X^T X)^{-}\).

ii) \(\hat{\beta} = (X^T X)^{-} X^T Y\) does depend on \((X^T X)^{-}\) and is not unique.

iii) \(\hat{Y} = X \hat{\beta} = PY\), \(r = Y - \hat{Y} = Y - X \hat{\beta} = (I - P)Y\) and \(\text{RSS} = r^T r\) are unique and so do not depend on \((X^T X)^{-}\).

iv) \(\hat{\beta}\) is a solution to the normal equations: \(X^T X \hat{\beta} = X^T Y\).

v) \(\text{Rank}(P) = r\) and \(\text{rank}(I-P) = n-r\).

vi) If \(\text{Cov}(Y) = \text{Cov}(e) = \sigma^2 I\), then \(\text{MSE} = \frac{\text{RSS}}{n-r} = \frac{r^T r}{n-r}\) is an unbiased estimator of \(\sigma^2\).

vii) Let the columns of \(X_1\) form a basis for \(C(X)\). For example, take \(r\) linearly independent columns of \(X\) to form \(X_1\). Then \(P = X_1 (X_1^T X_1)^{-} X_1^T\).

7) Estimability and the Gauss Markov Theorem:
Let \( a \) and \( b \) be constant vectors. Then \( a^T \beta \) is estimable if there exists a linear unbiased estimator \( b^T Y \) so \( E(b^T Y) = a^T \beta \). Also, \( a^T \beta \) is estimable iff \( a^T = b^T X \) iff \( a = X^T b \) iff \( a \in C(X^T) \).

The linear estimator \( a^T Y \) of \( c^T \theta \) is the best linear unbiased estimator (BLUE) of \( c^T \theta \) if \( E(a^T Y) = c^T \theta \), and if for any other unbiased linear estimator \( b^T Y \) of \( c^T \theta \), \( V(a^T Y) \leq V(b^T Y) \). Note that \( E(b^T Y) = c^T \theta \).

The next theorem shows that the least squares estimator of an estimable function \( a^T \beta \) is \( a^T \beta = b^T X \beta = b^T PY \). Note that \( b^T Y \) is also an unbiased estimator of \( a^T \beta \) since \( E(b^T Y) = b^T X \beta = a^T \beta \).

**Theorem 3.2** (see Seber and Lee Th 3.2) Let \( Y = X \beta + e \) where where \( X \) has rank \( r \leq p \leq n \), \( E(e) = 0 \), and \( \text{Cov}(e) = \sigma^2 I \).

a) The quantity \( a^T \beta \) is estimable iff \( a^T = b^T X \) iff \( a = X^T b \) (for some constant vector \( b \)) iff \( a \in C(X^T) \).

b) Let \( \theta = X \beta \) and \( \theta = X \beta \). Suppose there exists a constant vector \( c \) such that \( E(c^T \theta) = c^T \theta \). Then among the class of linear unbiased estimators of \( c^T \theta \), the least squares estimator \( c^T \theta \) is the unique BLUE.

c) **Gauss Markov Theorem:** If \( a^T \beta \) is estimable and a least squares estimator \( \hat{\beta} \) is any solution to the normal equations \( X^T X \beta = X^T Y \), then \( a^T \hat{\beta} \) is the unique BLUE of \( a^T \beta \).

**Proof:** a) If \( a^T \beta \) is estimable, then \( a^T \beta = E(b^T Y) = b^T X \beta \) for all \( \beta \in \mathbb{R}^p \). Thus \( a^T = b^T X \) or \( a = X^T b \). Hence \( a^T \beta \) is estimable iff \( a^T = b^T X \) iff \( a = X^T b \) iff \( a \in C(X^T) \).

b) Since \( \theta = X \beta = PY \), it follows that \( E(c^T \theta) = E(c^T PY) = c^T PX \beta = c^T X \beta = c^T \theta \). Thus \( c^T \theta = c^T PY = (Pc)^T Y \) is a linear unbiased estimator of \( c^T \theta \). Let \( d^T Y \) be any other linear unbiased estimator of \( c^T \theta \). Hence \( E(d^T Y) = d^T \theta = c^T \theta \) for all \( \theta \in C(X) \). So \( (c - d)^T \theta = 0 \) for all \( \theta \in C(X) \). Hence \( (c - d) \in [C(X)]^\perp \) and \( P(c - d) = 0 \), or \( Pc = Pd \). Thus \( V(c^T \theta) = V(c^T PY) = V(d^T PY) = \sigma^2 d^T P^T Pd = \sigma^2 d^T Pd \).

Then \( V(d^T Y) - V(c^T \theta) = V(d^T Y) - V(d^T PY) = \sigma^2 [d^T d - d^T Pd] = \sigma^2 d^T (I_n - P)d = \sigma^2 g^T g \geq 0 \) with equality iff \( g = (I_n - P)d = 0 \), or \( d = Pd = Pc \). Thus \( c^T \theta \) has minimum variance and is unique.

c) Since \( a^T \beta \) is estimable, \( a^T \hat{\beta} = b^T X \beta \). Then \( a^T \hat{\beta} = b^T \hat{\theta} \) is the unique BLUE of \( a^T \beta = b^T \theta \) by b).

**Gauss Markov Theorem-Full Rank Case:** Let \( Y = X \beta + e \) where \( X \) is full rank, \( E(e) = 0 \), and \( \text{Cov}(e) = \sigma^2 I \). Then \( a^T \hat{\beta} \) is the unique BLUE of \( a^T \beta \) for every constant \( p \times 1 \) vector \( a \).

Notation: \( \beta \) is “estimable” by \( \hat{\beta} \) for the full rank model, but not for the non-full rank model.

8) **Hypothesis Testing:**

**Theorem 2.16.** Let \( \theta = X \eta \in C(X) \) where \( Y_i = x_i^T \eta + r_i(\eta) \) and the residual \( r_i(\eta) \) depends on \( \eta \). The least squares estimator \( \hat{\beta} \) is the value of \( \eta \in \mathbb{R}^p \) that minimizes the least squares criterion 
\[ \sum_{i=1}^n r_i^2(\eta) = ||Y - X\eta||^2. \]

**LS CLT (Least Squares Central Limit Theorem):** Consider the MLR model \( Y_i = x_i^T \beta + e_i \) and assume that the zero mean errors are iid with \( E(e_i) = 0 \) and \( \text{VAR}(e_i) = \)
\[ \frac{X^TX}{n} \rightarrow W^{-1} \]

as \( n \rightarrow \infty \). Then the least squares (OLS) estimator \( \hat{\beta} \) satisfies

\[ \sqrt{n}(\beta - \beta) \overset{D}{\rightarrow} N_p(0, \sigma^2 W). \] (1)

**Partial F Test Theorem:** Suppose \( H_0 : L\beta = 0 \) is true for the partial F test where \( L \) is a full rank \( r \times p \) matrix. Under the OLS full rank model, a)

\[ F_R = \frac{1}{r\text{MSE}} (L\hat{\beta})^T[L(X^TX)^{-1}L^T]^{-1}(L\hat{\beta}). \]

b) If \( e \sim N_n(0, \sigma^2 I) \), then \( F_R \sim F_{r,n-p} \).

c) For a large class of zero mean error distributions \( rF \overset{D}{\rightarrow} \chi^2_r \).

d) The partial F test that rejects \( H_0 : L\beta = 0 \) if \( F_R > F_{r,n-p}(1-\delta) \) is a large sample right tail \( \delta \) test for the OLS model for a large class of zero mean error distributions.

Assume \( H_0 \) is true. By the OLS CLT, \( \sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}L\hat{\beta} \overset{D}{\rightarrow} N_r(0, \sigma^2 LWL^T) \). Thus \( \sqrt{n}(\hat{\beta})^T(\sigma^2 LWL^T)^{-1}(\beta) \overset{D}{\rightarrow} \chi^2_r \). Let \( \hat{\sigma}^2 = \text{MSE} \) and \( W = n(X^TX)^{-1} \). Then

\[ n(L\hat{\beta})^T[\text{MSE} Ln(X^TX)^{-1}L^T]^{-1}L\hat{\beta} \overset{D}{\rightarrow} \chi^2_r. \]

**Partial F test:** Let the full model \( Y = X\beta + e \) with a constant \( \beta_0 \) in the model: 1 is the 1st column of \( X \). Let the reduced model \( \hat{Y} = X_{R}\hat{\beta} + e \) also have a constant in the model where the columns of \( X_R \) are a subset of \( k \) of the columns of \( X \). Let \( P_R \) be the projection matrix on \( C(X_R) \) so \( PP_R = P_R \). Then \( F_R = \frac{\text{SSE}(R) - \text{SSE}(F)}{r\text{MSE}(F)} \) where \( r = df_R - df_F = p - k \) is number of predictors in the full model but not in the reduced model. \( \text{MSE} = \text{MSE}(F) = \text{SSE}(F)/(n-p) \) where \( \text{SSE} = \text{SSE}(F) = Y(I - P)Y \). \( \text{SSE}(R) - \text{SSE}(F) = Y^T(P - P_R)Y \) where \( \text{SSE}(R) = Y^T(I - P)Y \).

Now assume \( Y \sim N_n(X\beta, \sigma^2 I) \), and when \( H_0 \) is true, \( Y \sim N_n(X_{R}\beta, \sigma^2 I) \). Since \((I - P)(P - P_R) = 0, [\text{SSE}(R) - \text{SSE}(F)] \sim \text{MSE}(F) \) by Craig’s Theorem. When \( H_0 \) is true, \( \mu = X_{R}\beta_R \) and \( \mu^T A \mu = 0 \) where \( A = (I - P) \) or \( A = (P - P_R) \). Hence the noncentrality parameter is 0, and by Theorem 2.14 g), \( \text{SSE} \sim \sigma^2 \chi^2_{n-p} \) and \( \text{SSE}(R) - \text{SSE}(F) \sim \sigma^2 \chi^2_{p-k} \) since \( \text{rank}(P - P_R) = tr(P - P_R) = p - k \). Hence under \( H_0, F_R \sim F_{p-k,n-p} \).

An ANOVA table for the partial F test is shown below, where \( k = p_R \) is the number of predictors used by the reduced model, and \( r = p - p_R = p - k \) is the number of predictors in the full model that are not in the reduced model.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced</td>
<td>( n - p_R )</td>
<td>( \text{SSE}(R) = Y^T(I - P_R)Y )</td>
<td>( \text{MSE}(R) )</td>
<td>( F_R = \frac{\text{SSE}(R) - \text{SSE}(F)}{r\text{MSE}} )</td>
</tr>
<tr>
<td>Full</td>
<td>( n - p )</td>
<td>( \text{SSE} = Y^T(I - P)Y )</td>
<td>( \text{MSE} )</td>
<td>( \frac{Y^T(P - P_R)Y/r}{Y^T(I - P)Y/(n-p)} )</td>
</tr>
</tbody>
</table>
The ANOVA $F$ test is the special case where $k = 1$, $X_R = 1$, $P_R = P_1$, and $SSE(R) - SSE(F) = SSTO - SSE = SSR$.

ANOVA table: $Y = X\beta + e$ with a constant $\beta_1$ in the model: 1 is the 1st column of $X$. $MS = SS/df$.

$$SSSTO = Y^T(I - \frac{1}{n}11^T)Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad SSE = \sum_{i=1}^{n}r_i^2, \quad SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2,$$

$SSSTO = SSR + SSE$. SSR is the SSE (residual sum of squares) for the location model $Y = 1\beta_1 + e$ that contains a constant but no nontrivial predictors. The location model has projection matrix $P_1 = 1(1^T1)^{-1}1^T = \frac{1}{n}11^T$. Hence $PP_1 = P_1$ and $P1 = P_11 = 1$.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>p-1</td>
<td>$SSR = Y^T(P - \frac{1}{n}11^T)Y$</td>
<td>MSR</td>
<td>$F_0 = MSR/MSE$</td>
<td>for $H_0$:</td>
</tr>
<tr>
<td>Residual</td>
<td>n-p</td>
<td>$SSE = Y^T(I - P)Y$</td>
<td>MSE</td>
<td>$\beta_2 = \cdots = \beta_p = 0$</td>
<td></td>
</tr>
</tbody>
</table>

The matrices in the quadratic forms for SSR and SSE are symmetric and idempotent and their product is $0$. Hence if $e \sim N_n(0, \sigma^2I)$ so $Y \sim N_n(X\beta, \sigma^2I)$, then $SSE \parallel SSR$ by Craig’s Theorem. If $H_0$ is true under normality, then $Y \sim N_n(1\beta_1, \sigma^2I)$, and by Theorem 2.14 g), $SSE \sim \sigma^2\chi^2_{n-p}$ and $SSR \sim \sigma^2\chi^2_{p-1}$ since $\text{rank}(I - P) = \text{tr}(I - P) = n-p$ and $\text{rank}(P - \frac{1}{n}11^T) = \text{tr}(P - \frac{1}{n}11^T) = p-1$. Hence under normality, $F_0 \sim F_{p-1,n-p}$.

9 Expected Value, Covariance Matrix and Large Sample Theory for least squares quantities:

For the full rank model, $Y = X\beta + e$ where $E(Y) = X\beta$, $E(e) = 0$ and Cov($e$) = Cov($Y$) = $\sigma^2I$, $E(AY) = AX\beta$ and Cov($AY$) = $\sigma^2AA^T$.

$A = (X^TX)^{-1}X^T$ is used for $\hat{\beta} = AY$. $A = I - P = I - H$ is used for the residual vector $Y - \hat{Y} = AY$. $A = P = H$ is used for the vector of fitted values $\hat{Y}$.

For the full rank Gaussian linear model, $Y \sim N_n(X\beta, \sigma^2I)$, and if $A$ is $k \times n$ with rank $k$, then $AY \sim N_k(AX\beta, \sigma^2AA^T)$.

If $\sqrt{n}(\hat{\beta} - \beta) \overset{D}{\rightarrow} N_p(0, \sigma^2 W)$, and $A$ is $k \times p$ with rank $k$, then $\sqrt{n}(A\hat{\beta} - A\beta) \overset{D}{\rightarrow} N_k(0, \sigma^2 AW A^T)$.

The non-full rank model $Y = X\beta + e$ also has $E(Y) = X\beta$, $E(e) = 0$, Cov($e$) = Cov($Y$) = $\sigma^2I$, $E(AY) = AX\beta$ and Cov($AY$) = $\sigma^2AA^T$.

For the non-full rank model $A = (X^TX)^{-1}X^T$ is used for $\hat{\beta} = AY$ and $P = X(X^TX)^{-}X^T$.

You should be able to handle the linear model written in different ways. The residual bootstrap model $Y^* = X\hat{\beta} + e^*$ with $E(e^*) = 0$ and Cov($e^*$) = Cov($Y^*$) = $\sigma^2I$. The parametric bootstrap model $Y^* = X\hat{\beta} + e^*$ with $Y^* \sim N_n(X\hat{\beta}, MSE I)$. In numerical linear algebra, the least squares solution to “$Ax = b$” is of interest where the problem is actually the multiple linear regression model $b = Ax + e$ where $A$ has full rank $p$, and we will assume that $E(e) = 0$, and Cov($e$) = $\sigma^2I_n$.

References:


Olive, D.J. (2023), *Theory of Linear Models*, online course notes, see (http://parker.ad.siu.edu/Olive/linmodbk.htm).


