250 Notes

Nicholas Marco

December 9, 2020

Contents

1 Linear Algebra

1.1 Vectorspaces

Definition 1 *A* **vectorspace** *satisfies the following:*

1. $X + Y = \{x_i + y_i\} \in \mathbb{R}^n$
2. $0 \in \mathbb{R}^n$
3. $\alpha \in \mathbb{R}^n \implies \alpha x \in \mathbb{R}^n$ *for any* $x \in \mathbb{R}^n$

$\mathbb{R}$ is called a **field**.

Definition 2 *Consider* $V \subseteq \mathbb{R}^n$*.* V *is a* **subspace** *if it satisfies the following:*

1. V *is non-empty*
2. $X, Y \in V \implies X + Y \in V$
3. $X \in V, \alpha \in \mathbb{R} \implies \alpha X \in V$

From this definition, we can see that the vector space also has the following properties:

1. $x \in V \implies (-1)x = -x \in V$
2. $x + (-x) = 0 \in V$

If S and T are subspaces of V, then $S \cap T = \{x \in S \text{ and } x \in T | x \in V\}$ is a subspace, and $S + T = \{x + y | x \in S, y \in T\}$ is also a subspace.

Definition 3 *A sum is a* **direct sum** *if* $w \in S \bigoplus T = u + v$ *for a unique* $u \in S$ *and* $v \in T$*.*

Consider $S = \{(x, 0, 0)|x \in \mathbb{R}\}$ and $T = \{(0, y, 0)|y \in \mathbb{R}\}$. Thus we can see that for any $w \in S + T = \{(x, y, 0)|x, y \in \mathbb{R}\}$, we have a unique representation in terms of $w = u + v$, where $u \in S$ and $v \in T$. Therefore, $S + T$ is a direct sum ($S \bigoplus T$).

Theorem 1 *Let* S *and* T *be subspaces of* V*.* $S + T$ *is direct iff* $0 = u + v \implies u = v = 0$

Proof: ($\implies$) Suppose $S \bigoplus T$. Notice that $0 = u + v = 0 + 0$ where $u \in S$ and $v \in T$. Since direct sums are unique, this means that $u = v = 0$.
($\impliedby$) Suppose $0 = u + v \implies u = v = 0$. Let $w \in S + T$. Thus, by definition, $w = u_1 + v_1$ for some $u_1 \in S$ and $v_1 \in T$. Suppose $w = u_2 + v_2$ for some $u_2 \in S$ and $v_2 \in T$. Thus, we have

$$0 = w - w = (u_1 - u_2) + (v_1 - v_2) \implies (u_1 - u_2) = (v_1 - v_2) = 0$$

by the hypothesis. Therefore, we have that $u_1 = u_2$ and $v_1 = v_2$. Therefore, there is a unique representation, and $S + T$ is a direct sum. □

Lemma 1 $S \bigoplus T$ *iff* $S \cap T = \{0\}$.

Proof: ($\Longrightarrow$) Suppose $S \cap T = \{0, a\}$ for some $a \neq 0$. Since $S \cap T$ is a subspace, we know that $-a \in S \cap T$. Therefore, $0 = a - a$. Thus we have a contradiction from the theorem above.
($\Longleftarrow$) Suppose $S \cap T = \{0\}$. Let $0 = u + v$ for some $u \in S$ and $v \in T$. Thus we can see that $u = -v$. Since $u \in S$, $-u = v \in T$, so $v \in S \cap T$. Thus since $S \cap T = \{0\}$, $u = 0$ and therefore $v = 0$. Thus we can see that $0 = u + v \implies u = v = 0$. Therefore, by the theorem above, $S + T$ is direct. □

Definition 4 *Let $V = \{v_1, \ldots, v_n\}$. V is a* ***linearly independent*** *set if $a_1v_1 + a_2v_2 + \cdots + a_nv_n = 0 \implies a_1 = \cdots = a_n = 0$. Otherwise V is a* ***linearly dependent*** *set.*

Lemma 2 *If $V = \{v_1, \ldots, v_n\}$ is a linearly dependent set with $v_1 \neq 0$, then $\exists j \in \{2, \ldots, n\}$ such that:*

1. *$v_j \in span\{v_1, \ldots, v_{j-1}\}$*
2. *$span(V \setminus \{v_j\}) = span(V)$*

Proof: Suppose $\{v_1, \ldots, v_n\}$ is a set of linearly dependent variables. Let $v_1 \neq 0$. Therefore, we have

$$a_1v_1 + a_2v_2 + \cdots + a_nv_n = 0$$

There exists a $a_j \neq 0$ for $2 \leq j \leq n$. Choose the largest j such that $a_j \neq 0$. Thus $a_jv_n = -a_1v_1 - \cdots - a_{j-1}v_{j-1}$. Therefore, by definition, $a_j \in span(\{v_1, \ldots, v_{j-1}\})$. Thus we have proved (1).
Let $A = \{v_1, \ldots, v_n\}$. Let $\tilde{A} = A \setminus \{v_j\}$ where $v_j \in span(\{v_1, \ldots, v_{j-1}\})$. Thus we have to prove that $span(\tilde{A}) = span(A)$. It is obvious that $span(\tilde{A}) \subseteq span(A)$. Thus, we have to prove that $span(A) \subseteq span(\tilde{A})$. Suppose that $x \in span(A)$. Thus

$$x = a_1v_1 + \cdots + a_nv_n$$

Since $v_j \in span(\{v_1, \ldots, v_{j-1}\})$, we know that $v_j = c_1v_1 + \ldots c_{j-1}v_{j-1}$.

$$x = a_1v_1 + \cdots + a_{j-1}v_{j-1} + a_jv_j + \cdots + a_nv_n$$

$$x = a_1v_1 + \cdots + a_{j-1}v_{j-1} + a_j(c_1v_1 + \cdots + c_{j-1}v_{j-1}) + \cdots + a_nv_n$$

$$x = (a_1 + a_jc_1)v_1 + (a_2 + a_jc_2)v_2 + \cdots + (a_{j-1} + a_jc_{j-1})v_{j-1} + a_{j+1}v_{j+1} + \cdots + a_nv_n$$

Therefore, we can see that $x \in span(\tilde{A})$. Therefore we can see that $span(A) \subseteq span(\tilde{A})$. Therefore we have that $span(A) = span(\tilde{A})$ (2). □

Lemma 3 *The* ***linear dependence lemma*** *uses the lemma above to find a linearly independent set of vectors. We can start by searching through the set of vectors to look for a $v_j \in span\{v_1, \ldots, v_{j-1}\}$. If we find one, then we can remove it from the set without changing the span of this set. We can continue in this iterative fashion until we do not find a v_j such that $v_j \in span\{v_1, \ldots, v_{j-1}\}$. Once this happens, we have a linearly independent set.*

Using the linear dependence lemma, we can explore the connection between linearly independent sets and spanning sets. We will prove that spanning sets must have at least as many elements as a linearly independent set.
Let V be a subspace. Let $U = \{u_1, \ldots, u_m\}$ be a linearly independent sets such that $U \subseteq V$. Let $W = \{w_1, \ldots, w_n\}$ be a spanning set such that $span(W) = V$.
Suppose that $m > n$.

Let $A_0 = W$. Thus we can see that $span(A_0) = V$. Let $A_1 = \{u_1\} \cup A_0$. Since $\{u_1\} \in span(A_0)$, by the linear dependence lemma, we can find a $w_{j1} \in \{w_1, \dots, w_n\}$ such that $span(B_1) = span(A_1) = V$, where $B_1 = A_1 \setminus w_{j1}$.
Let $A_2 = \{u_2\} \cup B_1 = \{u_2, u_1, w_1, \dots, w_{j-1}, w_{j+1}, \dots, w_n\}$. Since $u_2 \in span(B_1) = V$, we can find a $w_{j2} \in \{w_1, \dots, w_{j-1}, wj+1, \dots, w_n\}$ such that $span(B_2) = span(A_2) = span(B_1) = V$ where $B_2 = A_2 \setminus w_{j2}$.
We can continue in this fashion until we have $A_n = \{u_n\} \cup B_{n-1} = \{u_n, \dots, u_1, w_i\}$ for some $w_i \in W$. Since $span(B_{n-1}) = V$, $u_n \in spanV$, and since $u_n \perp u_i$ for all $i \neq n$, we can see that $span(B_n) = span(A_n) = V$, where $B_n = \{u_n, \dots, u_1\}$.
Consider $u_{n+1} \in V$. Therefore, $u_{n+1} \in span(\{u_n, \dots, u_1\})$. Therefore, $u_{n+1} = a_1u_1 + \dots a_nv_n$. Therefore $a_1u_1 + \dots + a_nv_n - u_{n+1} = 0$, so by definition, $\{u_1, \dots, u_n\}$ is not a linearly independent set (which is a contradiction). Therefore, we can see that spanning sets must have at least as many elements as a linearly independent set.

Definition 5 *A **basis** is a linearly independent spanning set*

Thus, from the definition we can see that every basis must contain the same number of elements. Let $B_1 = \{u_1, \dots, u_m\}$ and $B_2 = \{v_1, \dots, v_n\}$ both be bases for V. Since B_1 is a linearly independent set and B_2 is a spanning set, we have $m \leq n$. Since B_2 is a linearly independent set and B_1 is a spanning set, we have $n \leq m$. Therefore, we have $m = n$.

Definition 6 *The **dimension** of a subspace is the number of vectors in a basis for that subspace.*

Let V_1 be a subspace of V, where $dim(V) = n$. Let $B_1 = \{u_1, \dots, u_m\}$ be a basis for V_1. We can **extend** B_1 to a basis for V. All we have to do is find a $v_1 \in V$ such that $v_1 \notin span(V_1)$. Therefore, $\{u_1, \dots, u_m, v_1\}$ is a linearly independent set. If $n = m + 1$, then $\{u_1, \dots, u_m, v_1\}$ is a basis for V. If not we can continue in this fashion until we have a basis for V. If $V \subseteq \mathbb{R}^n$, we can append the vectors $\{e_1, \dots, e_n\}$ to $\{u_1, \dots, u_m\}$, where e_i is a vector of zeros except for the i^{th} element which is 1. We can use the linear dependence lemma to find a linearly independent subset. The resulting set will be a basis for V.

Lemma 4 *Let V be a vector space and let S and T be subspaces of V. Thus we have $dim(S+T) = dim(S) + dim(T) - dim(S \cap T)$.*

Proof: Let $\{u_1, \dots, u_r\}$ be a basis for $S \cap T$. Extend the basis so that $\{u_1, \dots, u_r, w_1, \dots, w_m\}$ is a basis for S and $\{u_1, \dots, u_r, v_1, \dots, v_n\}$ is a basis for T. Therefore we have $dim(S) = m + r$ and $dim(T) = n + r$.
We need to prove that $\{u_1, \dots u_r, w_1, \dots, w_m, v_1, \dots v_n\}$ is a basis for $S + T$. By construction we can see that it is a linearly independent set. Thus have to prove that it spans $S + T$. Let $x \in S + T$. Thus by definition, $x = w + v$ for some $w \in S$ and $v \in T$. Since $w \in span(\{u_1, \dots, u_r, w_1, \dots, w_m\})$ and $v \in span(\{u_1, \dots, u_r, v_1, \dots, v_n\})$, we have that $x \in span(\{u_1, \dots u_r, w_1, \dots, w_m, v_1, \dots v_n\})$. Therefore, $\{u_1, \dots u_r, w_1, \dots, w_m, v_1, \dots v_n\}$ is a basis for $S+T$ and $m+n+r = dim(S+T) = dim(S)+dim(T)-dim(S \cap T) = m+r+n+r-r = m+n+r$
□

Definition 7 *Two vector spaces V_1 and V_2 over the same field $\mathcal{F}$ are **isomorphic** if there is a map ψ from V_1 to V_2 such that:*

1. *$\psi(x)$ is linear, meaning that $\psi(x + y) = \psi(x) + \psi(y)$ and $\psi(\alpha x) = \alpha\psi(x)$ for all $x, y \in V_1$ and $\alpha \in \mathcal{F}$*

2. $\psi(x)$ is a one-to-one and onto function

*The map $\psi(x)$ is called an **isomorphism**.*

Theorem 2 *Two vector spaces V_1 and V_2 over the same field $\mathcal{F}$ are isomorphic if and only if they have the same dimension.*

Proof: ($\Longleftarrow$) Let $\{x_1, \dots, x_n\}$ be a basis for V_1 and $\{y_1, \dots, y_n\}$ be a basis for V_2. Consider $x \in V_1$. Thus $x = a_1x_1 + \dots a_nx_n$ for a unique set of $a_1, \dots, a_n$. Define the map

$$\psi(x) = a_1y_1 + \cdots + a_ny_n$$

Clearly, we can see that $\psi(x) : V_1 \to V_2$. We can also see that $\psi(x)$ is linear. Since each element in V_1 can we expressed in such a way (with unique $a_1, \dots, a_n$), we can see that $\psi(x)$ is a both one-to-one and onto function.
($\Longrightarrow$) Now suppose that V_1 and V_2 are isomorphic. Let $\phi : V_1 \to V_2$ be an isomorphism. Let $\{x_1, \dots, x_n\}$ be a basis for V_1.
Claim: $\{\psi(x_1), \dots, \psi(x_n)\}$ is a linearly independent set.
Note that $\psi(x) = 0 \iff x = 0$ (If not, suppose $\psi(x) = y$. Then $\psi(x + 0) = \psi(x) + \psi(0) \neq y$ which contradicts that $\psi(x)$ is linear.) Therefore, we have:

$$= 0\alpha_1\psi(x_1) + \dots \alpha_n\psi(x_n) = \psi(\alpha_1x_1 + \cdots + \alpha_nx_n)$$

Thus we have,

$$\alpha_1x_1 + \cdots + \alpha_nx_n = 0 \implies \alpha_1 = \cdots = \alpha_n = 0$$

Therefore we can see that $\{\psi(x_1), \dots, \psi(x_n)\}$ is a linearly independent set. Therefore we know that $dim(V_2) \geq dim(V_1)$.
We can do the same by defining an isomorphism, $\psi_2 : V_2 \to V_1$. Thus we will get $dim(V_1) \leq dim(V_2)$. Therefore $dim(V_1) = dim(V_2)$.

□

Consider $\mathbb{R}^n$ and $\mathcal{P}^n = \{p(x) = a_0 + a_1x^2 + \cdots + a_{n-1}x^{n-1} | a_i \in \mathbb{R}\}$. $\mathbb{R}^n$ and $\mathcal{P}^n$ are an example of two vector spaces that are isomorphic.

1.2 Inner Product Spaces

Definition 8 *An **inner product**, $\langle .,. \rangle : V \times V \to \mathcal{F}$ satisfies the following:*

1. $\langle x, x \rangle \geq 0 \ \forall x$
2. $\langle x, x \rangle = 0 \iff x = 0$
3. *Bilinear:* $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ *and* $\langle z, \alpha x + \beta y \rangle = \overline{\alpha} \langle z, x \rangle + \overline{\beta} \langle z, y \rangle$ *where* $\overline{\gamma}$ *is the complex conjugate of* γ.
4. $\langle x, y \rangle = \overline{\langle y, x \rangle}$

Thus we can say that $\langle x, y \rangle = \sum \overline{y_i} x_i = y^*x$ where y^* is the adjoint of y (defined later). For $x, y \in \mathbb{R}^n$, $y^*x = y'x = x'y$.

Definition 9 *The **norm** (length) of a vector is* $\sqrt{\langle x, x \rangle} = ||x|| = \sqrt{x^*x}$.

We can introduce the notion of an angle between two vectors in a vector space.

$$\langle x, y \rangle = x^*y = ||x||||y||cos(\theta) \implies cos(\theta) = \frac{\langle x, y \rangle}{||x||y||}$$

Definition 10 *We call two vectors X and Y* ***orthogonal*** *($X \perp Y$) iff $\langle x, y\rangle = 0$*

Thus we can see that two vectors are orthogonal if $cos(\theta) = 0$ or $\theta = \pi/2$.

Theorem 3 ***(Cauchy- Schwarz)*** *Let u and v be any two vectors in a vector space V. Then*

$$|\langle u, v\rangle| \leq ||u||||v||$$

Proof: Consider $||u - \frac{\langle u,v\rangle}{||v||^2}v||^2$. By axioms of norms, we know that $||u - \frac{\langle u,v\rangle}{||v||^2}v||^2 \geq 0$. Thus we have

$$||u - \frac{\langle u, v\rangle}{||v||^2}v||^2 = \langle u - \frac{\langle u, v\rangle}{||v||^2}v, u - \frac{\langle u, v\rangle}{||v||^2}v\rangle$$

$$= \langle u, u\rangle - 2\langle u, \frac{\langle u, v\rangle}{||v||^2}v\rangle + \langle \frac{\langle u, v\rangle}{||v||^2}v, \frac{\langle u, v\rangle}{||v||^2}v\rangle$$

$$= ||u||^2 - 2\frac{\langle u, v\rangle^2}{||v||^2} + \frac{\langle u, v\rangle^2}{||v||^2}$$

$$= ||u||^2 - \frac{\langle u, v\rangle^2}{||v||^2}$$

Since $||.|| \geq 0$ (axiom of norms), we have

$$||u||^2 - \frac{\langle u, v\rangle^2}{||v||^2} \geq 0$$

Thus we have

$$||u||^2||v||^2 \geq \langle u, v\rangle^2 \implies ||u||||v|| \geq |\langle u, v\rangle|$$

since $||u||$ and $||v||$ are non-negative □

Consider the statistical model

$$y = \beta x$$

where β is a scalar. We know that

$$||y - \beta x||^2 \geq 0$$

by the axiom of norms. Consider the estimated β, $\hat{\beta}$. Thus we have

$$||y - \beta x||^2 = ||y - \hat{\beta}x + \hat{\beta}x - \beta x||^2 = ||y - \hat{\beta}x||^2 + 2\langle y - \hat{\beta}x, \hat{\beta}x - \beta x\rangle + (\hat{\beta} - \beta)||x||^2$$

Thus, we can se that

$$||y - \hat{\beta}x||^2 = ||y - \beta x||^2 - 2\langle y - \hat{\beta}x, \hat{\beta}x - \beta x\rangle - (\hat{\beta} - \beta)||x||^2$$

If set $\hat{\beta}$ such that $\langle y - \hat{\beta}x, \hat{\beta}x - \beta x\rangle = 0$, then we have that

$$||y - \hat{\beta}x||^2 = ||y - \beta x||^2 - 2\langle y - (\hat{\beta} - \beta)||x||^2 \implies ||y - \hat{\beta}x||^2 \leq ||y - \beta x||^2 \; \forall\beta$$

$$\langle y - \hat{\beta}x, \hat{\beta}x - \beta x\rangle = 0 \implies (\hat{\beta} - \beta)\langle y, x\rangle = (\hat{\beta} - \beta)\hat{\beta}||x||^2$$

Thus we can see that $\hat{\beta} = \frac{\langle y,x\rangle}{||x||^2}$ is the optimal solution such that the residuals are minimized $(||y - \hat{\beta}x||^2 \leq ||y - \beta x||^2 \; \forall\beta)$.

Lemma 5 *An orthonormal set is also a linearly independent set.*

Proof: Let $\{u_1, \ldots, u_n\}$ be an orthonormal set. Therefore $u_i \perp u_j$ for all $i \neq j$. Let $a_1u_1 + \cdots + a_nu_n = 0$. Consider

$$\langle u_i, a_1u_1 + \cdots + a_nu_n \rangle = \langle u_i, 0 \rangle = 0$$

Since $u_i \perp u_j$ for all $i \neq j$, we have

$$\langle u_i, a_1u_1 + \cdots + a_nu_n \rangle = \langle u_i, a_iu_i \rangle = 0$$

Therefore, since $u_i \neq 0$, we know that $a_i = 0$. Since this holds for $1 \leq i \leq n$, we can see that $a_1 = \cdots = a_n = 0$. Therefore, by definition, $\{u_1, \ldots, u_n\}$ is an linearly independent set. □

Lemma 6 *If $V \perp \{u_1, \ldots, u_m\}$, then $V \perp span(\{u_1, \ldots, u_m\})$.*

Proof: Let $u \in span(\{u_1, \ldots, u_m\})$. Thus, $u = a_1u_1 + \cdots + a_mu_m$.

$$\langle v, u \rangle = a_1\langle v, u_1 \rangle + \cdots + a_m\langle v, u_m \rangle = a_1(0) + \cdots + a_m(0)$$

Thus $V \perp u$. Therefore, $V \perp span(\{u_1, \ldots, u_m\})$. □

Let $\{u_1, \ldots, u_m\}$ be an orthonormal set. Let $x \in span(\{u_1, \ldots, u_m\})$. Let $x = c_1u_1 + \cdots + c_nu_n$. we can see that $\langle u_i, x \rangle = \langle u_i, c_1u_1 + \cdots + c_nu_n \rangle = \langle u_i, c_i, u_i \rangle = c_i$. Therefore we can rewrite x as

$$x = \langle x, u_1 \rangle u_1 + \cdots + \langle x, u_m \rangle u_m$$

Thus

$$\begin{aligned}\langle x, x \rangle &= \langle \langle x, u_1 \rangle u_1 + \cdots + \langle x, u_m \rangle u_m, \langle x, u_1 \rangle u_1 + \cdots + \langle x, u_m \rangle u_m \rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \langle x, u_i \rangle \langle x, u_j \rangle \langle u_i, u_j \rangle \\ &= \sum_{i=1}^{m} \langle x, u_i \rangle \langle x, u_i \rangle \langle u_i, u_i \rangle = \sum_{i=1}^{m} \langle x, u_i \rangle^2\end{aligned}$$

Now we can look into the property of vectors orthogonal to a subspace. Let $x \in V$. Let $\{u_1, \ldots, u_m\} \subseteq V$ be an orthonormal set. Let $u = \langle x, u_1 \rangle u_1 + \cdots + \langle x, u_m \rangle u_m$ (note this is known as the projection (defined later) of x onto $span(\{u_1, \ldots, u_m\})$). Clearly $u \in span(\{u_1, \ldots, u_m\})$. Let $v = x - u$.

$$\langle u, u_i \rangle = \langle \langle x, u_1 \rangle u_1 + \cdots + \langle x, u_m \rangle u_m, u_i \rangle = \sum j = 1^m \langle x, u_j \rangle \langle u_j, u_i \rangle = \langle x, u_i \rangle$$

$$\langle v, u_i \rangle = \langle x - u, u_i \rangle = \langle x, u_i \rangle - \langle u, u_i \rangle = \langle x, u_i \rangle - \langle x, u_i \rangle = 0$$

Therefore we can see that $v \perp span(\{u_1, \ldots, u_m\})$. We say that $v \in U^{\perp}$. Is $U^{\perp}$ a subspace?

1. Notice that $\langle 0, u \rangle = 0$, so $0 \in U^{\perp}$.
2. Let $x, y \in U^{\perp}$. Therefore, $\langle x, u \rangle = 0$ and $\langle y, u \rangle = 0$. Therefore by the linearity of inner products, $\langle x + y, u \rangle = 0$, so $x + y \in U^{\perp}$.
3. Let $\alpha \in \mathcal{F}$. Therefore, $\langle \alpha x, u \rangle = \alpha \langle x, u \rangle = \alpha(0) = 0$. Therefore $\alpha x \in U^{\perp}$.

Therefore, we can see that $U^{\perp}$ is a subspace.
Let $\{v_1, \ldots, v_n\}$ be a basis for V. Can we find an orthonormal basis for V?

Theorem 4 ***(Gram-Schmidt)*** *Let $\{v_1, \ldots, v_n\}$ be a basis for V.
Define $u_1 = \frac{v_1}{||v_1||}$, $u_i = \frac{w_i}{||w_i||}$, where $w_i = v_i - \sum_{j=1}^{i-1} \langle v_i, u_j \rangle u_j$ for $2 \leq i \leq n$.
Then $\{u_1, \ldots, u_n\}$ is an orthonormal basis for V.*

Proof: We need to show

1. $\{u_1, \dots, u_n\}$ is an orthonormal set
2. $span(\{v_1, \dots, v_n\}) = span(\{u_1, \dots, u_n\})$

(1) We can immediately see that $||u_i|| = 1$, so we are left to prove that $u_i \perp u_j$ for $i \neq j$. Consider

$$\langle u_1, w_2 \rangle = \langle u_1, v_2 - \langle v_2, u_1 \rangle u_1 \rangle = \langle u_1, v_2 \rangle - \overline{\langle v_2, u_1 \rangle} \langle u_1, u_1 \rangle = 0$$

Therefore, we can see that $u_1 \perp w_2 \implies u_1 \perp u_2$.
Now suppose that $u_{i-1} \perp \{u_1, \dots, u_{i-2}\}$. Thus we need to prove that $w_i \perp \{u_1, \dots, u_{i-1}\}$. Let $1 \leq j \leq i-1$

$$\langle w_i, u_j \rangle = \langle v_i - \sum_{k=1}^{i-1} \langle v_i, u_k \rangle u_k, u_j \rangle = \langle v_i, u_j \rangle - \langle v_i, u_j \rangle \langle u_j, u_j \rangle = \langle v_i, u_j \rangle - \langle v_i, u_j \rangle = 0$$

Thus we can see that $w_i \perp \{u_1, \dots, u_{i-1}\} \implies u_i \perp \{u_1, \dots, u_{i-1}\}$. Therefore, we can see that $\{u_1, \dots, u_n\}$ is an orthonormal set
(2) We can see that $v_1 = r_{11} u_1$ where $r_{11} = ||u_1||$.
Continuing in this fashion, we have $v_2 = r_{12} u_1 + r_{22} u_2$ where $r_{12} = \langle v_2, u_1 \rangle$ and $r_{22} = ||w_{22}||$.
For the i^{th} step, we have $v_i = r_{ii} u_i + \sum_{j=1}^{i-1} r_{ji} u_j$ where $r_{ji} = \langle v_i, u_j \rangle$ and $r_{ii} = ||w_{ii}||$. Therefore, we have $v_i \in span(\{u_1, \dots, u_i\})$.
Therefore, we have $span(\{v_1, \dots, v_n\}) = span(\{u_1, \dots, u_n\})$. □

Lemma 7 *Using the Gram-Schmidt process, we can derive the* ***QR Decomposition*** $X = QR$ *where Q is an orthogonal matrix and R is an upper triangular matrix. Let $X = [v_1, \dots, v_m] \in \mathbb{R}^{n \times m}$ and $Q = [u_1, \dots, u_m] \in \mathbb{R}^{n \times m}$. Thus we have*

$$X = [u_1, \dots, u_m] \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ 0 & r_{22} & \dots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{mm} \end{bmatrix}$$

Let $U = \{u_1, \dots, u_k\} \subseteq V$ be an orthonormal set. We can extend this orthonormal set to an orthonormal basis for V, by the following:

1. extend U to a basis for V
2. Apply Gram-Schmidt to the expanded basis

Let V be a vector space with $dim(V) = n$. Let $S \subseteq V$ be a subspace of V. Let $\{u_1, \dots, u_m\}$ be an orthonormal basis of S. We can extend the basis of S to a basis of V such that $\{u_1, \dots, u_m, v_{m+1}, \dots, v_n\}$ is an orthonormal basis of V.

Lemma 8 *$\{v_{m+1}, \dots, v_n\}$ forms an orthonormal basis for $S^\perp$.*

Proof: Let $v \in S^\perp \subseteq V$. Thus we have

$$v = c_1 u_1 + \dots + c_m u_m + c_{m+1} v_{m+1} + \dots + c_n v_n$$

Since $v \perp \{u_1, \dots, u_m\}$, we know that $0 = \langle v, u_i \rangle = c_i$ for $1 \leq i \leq m$. Therefore, we have

$$v = c_{m+1} v_{m+1} + \dots + c_n v_n$$

Therefore, we can see that $v \in span(\{v_{m+1}, \dots, v_n\})$. Therefore $S^{\perp} \subseteq span(\{v_{m+1}, \dots, v_n\})$
Now suppose that $v \in span(\{v_{m+1}, \dots, v_n\})$. Thus, by definition, we have

$$v = a_{m+1}v_{m+1} + \cdots + a_n v_n$$

$$\langle v, u_i \rangle = a_{m+1}\langle v_{m+1}, u_i \rangle + \cdots + a_n \langle v_n, a_n \rangle = 0$$

for $1 \leq i \leq m$ since $u_i \perp v_j$ $(m+1 \leq j \leq n)$. Therefore, we can see that $v \perp span(u_1, \dots, u_m)$, so by definition $v \in S^{\perp}$. Therefore $span(\{v_{m+1}, \dots, v_n\}) \subseteq S^{\perp}$.
Therefore we can see that $\{v_{m+1}, \dots, v_n\}$ spans $S^{\perp}$, and since it is part of an orthonormal basis (of V), we know that it is orthonormal itself. Therefore $\{v_{m+1}, \dots, v_n\}$ is an orthonormal basis for $S^{\perp}$ □

From this proof, we can see that $dim(S^{\perp}) = n - m$ and $dim(S) = m$.

Lemma 9 *Let $V = S + T$. If $dim(V) = dim(S) + dim(V)$, then $V = S \bigoplus T$.*

Proof: Let $\{u_1, \dots, u_r\}$ be a basis for S and let $\{v_1, \dots, v_k\}$ be a basis for T. Thus by the hypothesis, we know that $dim(V) = r+k$. Consider $x \in V$. Thus we know that $x = u+v$, where $u \in span(u_1, \dots, u_r)$ and $v \in span(\{v_1, \dots, v_k\})$. Thus we know that $\{u_1, \dots, u_r, v_1, \dots, v_k\}$ spans V. Since $dim(V) = k + r$, we know that $\{u_1, \dots, u_r, v_1, \dots, v_k\}$ is linearly independent (if not we could find a smaller spanning set), and therefore is a basis.
Now suppose $0 = u + v = a_1u_1 + \cdots + a_ru_r + b_1v_1 + \cdots + b_kv_k$. Since $\{u_1, \dots, u_r, v_1, \dots, v_k\}$ is a linearly independent set, we know that

$$0 = a_1u_1 + \cdots + a_ru_r + b_1v_1 + \cdots + b_kv_k \implies a_1 = \cdots = a_r = b_1 = \cdots = b_k = 0$$

Therefore, $0 = u+v \implies u = 0$ and $v = 0$. Therefore, by theorem, we know that $V = S \bigoplus T$. □

Therefore, from the lemma above, we know that if $V = S + S^{\perp}$, then $V = S \bigoplus S^{\perp}$.
If $V = S \bigoplus S^{\perp}$, then for any $x \in V$, we know that $x = u + v$ for a unique $u \in S$ and $v \in S^{\perp}$. If $\{u_1, \dots, u_m\}$ is an orthonormal basis for S (with respect to an inner product) then we know how to construct u. Namely,

$$u = \langle x, u_1 \rangle u_1 + \langle x, u_2 \rangle u_2 + \cdots + \langle x, u_m \rangle u_m$$

Note this is called the projection of x onto S. We can also see that $v = x - u$.

Definition 11 *Let $x \in V$ and U be a subspace of V. Thus we can see that $V = U \bigoplus U^{\perp}$. Thus we can see that for any $x \in V$, $x = u + v$ for a unique $u \in U$ and $v \in U^{\perp}$. We can define the unique mapping $P : x \in V \to u \in U$ as an **orthogonal projector**. $u = P(x)$ is known as the **orthogonal projection**.*

What does P look like? Suppose $x \in U \subseteq V$. Then we have that $P(x) = x$ since $x = x + 0$ and $0 \perp x$. Suppose that $x \in V$. By definition, we know $P(x) \in U$, so $P(P(x)) = P(x)$. Therefore, we have that $P \circ P = P^2 = P$ (idempotent). Let $x, y \in V$, what does $P(x + y)$ look like? We can decompose x and y such that $x = u_1 + v_1$ and $y = u_2 + v_2$ where $u_1, u_2 \in U$ and $v_1, v_2 \in U^{\perp}$. Therefore, we can see that $P(x + y) = u_1 + u_2 = P(x) + P(y)$. We can also see that $\alpha x = \alpha u_1 + \alpha v_1$. Therefore, we can see that $P(\alpha x) = \alpha u_1 = \alpha P(x)$. Therefore we can see that P is a linear map or linear transformation.
For any $x \in V$, we can see that $x = P(x) + v = P(x) + (I - P)(x)$. Thus we can see that $I - P : x \in V \to v \in U^{\perp}$.

Definition 12 *If $V = S \bigoplus T$, then $x = P_s(x) + (I - P_s)(x)$ is called an **oblique projection**. If $T = S^{\perp}$, then it is considered an orthogonal projection.*

Theorem 5 ***(Pythagoras)*** *Let $x \in V$ and let U be a subspace of V. Then $x = u + v$ where $u \in U$ and $v \in U^{\perp}$.*

$$||x||^2 = ||u||^2 + ||v||^2$$

Proof: Let $x \in V$ and let U be a subspace of V. Then $x = u + v$ where $u \in U$ and $v \in U^{\perp}$. Then

$$|x||^2 = ||u+v||^2 = \langle u+v, u+v\rangle = \langle u, u\rangle + 2\langle u, v\rangle + \langle v, v\rangle = ||u||^2 + 2\langle u, v\rangle + ||v||^2$$

since $u \perp v$, we have

$$||x||^2 = ||u||^2 + ||v||^2$$

□

Theorem 6 ***(Approximation Theorem)*** *Let V be a vector space. Let U be a subspace of V. Let $x \in V$. Then*

$$||x - P_U(x)|| \leq ||x - u|| \ \ \forall u \in U$$

where $P_U(x)$ is the orthogonal projection of x onto U.

Proof:

$$||x-u||^2 = ||x-P_U(x)+P_U(x)-u||^2 = ||x-P_U(x)||^2 + 2\langle x-P_U(x), P_U(x)-u\rangle + ||P_U(x)-u||^2$$

Notice that $x - P_U(x) \in U^{\perp}$ and $P_U(x) - u \in U$. Therefore by the Pythagoras theorem, we know that

$$||x-u||^2 = ||x-P_U(x)||^2 + ||P_U(x)-u||^2 \implies ||x-P_U(x)||^2 = ||x-u||^2 - ||P_U(x)-u||^2$$

Therefore, by the positivity of norms, we know that $||x - P_U(x)|| \leq ||x - u|| \ \ \forall u \in U$ with equality when $P_U(x) = u$. □

Using the approximation theorem, we can construct the projection matrix used in linear regression.
Let $V = \mathbb{R}^n$ and $\{u_1, \ldots, u_p\}$ be an orthonormal basis for U.

$$P_U(x) = \langle x, u_1\rangle u_1 + \langle x, u_2\rangle u_2 + \cdots + \langle x, u_p\rangle u_p$$
$$= u_1' x u_1 + u_2' x u_2 + \cdots + u_p' x u_p = u_1 u_1' x + u_2 u_2' x + \cdots + u_p u_p' x$$
$$= (u_1 u_1' + u_2 u_2' + \cdots + u_p u_p')x = QQ'x$$

where $Q = [u_1, \ldots, u_p]$. Suppose you are given a matrix of predictors $X \in \mathbb{R}^{n \times p}$, where the columns of X are linearly independent and $span(U) = colspace(X)$.
We can start with a QR decomposition of $X = QR$, where Q is orthogonal and R is upper triangular.

$$X = QR \implies Q'X = Q'QR = R$$

Since the columns of X are linearly dependent, we know that the diagonal elements of R are non-zero. This means that $Rx = 0 \implies x = 0 \implies \exists R^{-1}$ Since we know the projection matrix is QQ', we have

$$QQ' = X(R^{-1}R^{-T})X' = X(R'R)^{-1}X' = X(X'QQ'X)^{-1}X'$$

Notice that $QQ'X$ is the projection of X onto itself, so $QQ'X = X$. Therefore, we have

$$QQ' = X(X'X)^{-1}X'$$

which is the projection matrix found in classical linear algebra books.

1.3 Linear Transformations and Matrices

Definition 13 *Let V and W be vector spaces. Let $dim(V) = n$ and $dim(W) = m$. $T : V \to W$ is defined as a **linear transformation** and has the following properties*

1. $T(x + y) = T(x) + T(y) \;\; \forall x, y \in V$
2. $T(\alpha x) = \alpha T(x) \;\; \forall x \in V, \alpha \in \mathcal{F}$

From this definition, the we can see that $T(0) = 0$ ($T(x) = T(x+0) = T(x)+T(0) \implies T(0) = 0$).
Suppose $T : V \to W$ ($T \in \mathcal{L}(V, W)$). Let $\{v_1, \dots v_n\}$ be a basis for V. and $\{w_1, \dots, w_m\}$ be a basis for W.
Thus consider $x \in V$. Thus we have $x_1v_1 + \cdots + x_nv_n$. Therefore we have

$$T(x) = x_1T(v_1) + \cdots + x_nT(v_n)$$

Since $T(v_j) \in W$, we can say that $T(v_j) = t_{1j}w_1 + \dots t_{mj}w_m$. Thus we can see that

$$T(x) = x_1(t_{11}w_1 + \dots t_{m1}w_m) + \cdots + x_n(t_{1n}w_1 + \dots t_{mn}w_m)$$

Thus we can see that t_{ij} is the coordinates of $T(x)$ with respect to the basis vector v_i and basis vector w_j. We can use this idea to formulate a matrix.

Definition 14 *A **matrix**, $A = M(T(x), \{v_1, \dots, v_n\}, \{w_1, \dots, w_m\})$, are the coordinates of $T(X) \in \mathcal{L}(V, W)$ with respect to the bases $\{v_1, \dots, v_n\} \in V$ and $\{w_1, \dots, w_m\} \in W$.*

$$A = \begin{bmatrix} t_{11} & t_{21} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \dots & t_{mn} \end{bmatrix}$$

Thus form this definition, we can see that $T(x) = Ax = x_1(t_{11}w_1 + \dots t_{m1}w_m) + \cdots + x_n(t_{1n}w_1 + \dots t_{mn}w_m)$ or

$$Ax = \begin{bmatrix} t_{11} & t_{21} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \dots & t_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Lemma 10 $M(S \circ T) = M(S)M(T)$

Proof: Let $T : V \to W$ and $S : W \to U$. Let $\{v_1, \dots v_n\}$ be a basis for V, $\{w_1, \dots, w_m\}$ be a basis for W, and $\{u_1, \dots, u_p\}$ be a basis for U.

$$S \circ T(v_j) = S(T(v_j)) = S(t_{1j}w_1 + \cdots + t_{mj}w_m) = t_{1j}S(w_1) + \cdots + t_{mj}S(w_m)$$

We can see that $S(w_k) = s_{1k}u_1 + \cdots + s_{pk}u_p$. Therefore, we have

$$S(T(v_j)) = t_{1j}(s_{11}u_1 + \dots s_{p1}u_p) + \cdots + t_{mj}(s_{1m}u_1 + \dots s_{pm}u_p)$$

$$= (t_{1j}s_{11} + t_{2j}s_{12} + \cdots + t_{mj}s_{1m})u_1 + \cdots + (t_{1j}s_{p1} + t_{2j}s_{p2} + \cdots + t_{mj}s_{pm})u_p$$

thus we can see that the $(k, j)^{th}$ is $\sum_i t_{ij}s_{ki}$ for $1 \leq k \leq p$ and $1 \leq j \leq n$. Thus we can see that $\sum_i t_{ij}s_{ki}$ is the $(k, j)^{th}$ element of $M(S)M(T)$. Thus we have $M(S \circ T) = M(S)M(T)$. □

Lemma 11 $M(S + T) = M(S) + M(T)$

Proof: Let $\{v_1 \dots v_m\}$ be a basis for V. Let $\{w_1, \dots, w_n\}$ be a basis for W. Thus, we can express $S(v_j)$ and $T(v_j)$ as:

$$S(v_j) = s_{1j}w_1 + \cdots + s_{nj}w_n$$

$$T(v_j) = t_{1j}w_1 + \cdots + t_{nj}w_n$$

$$(S+T)(v_j) = S(v_j) + T(v_j) = (s_{1j} + t_{1j})w_1 + \dots (s_{nj} + t_{nj})w_n$$

Thus the $(i, j)^{th}$ element of $M(S+T)$ is $(s_{ij}+t_{ij})$ for $1 \leq i \leq n$ and $1 \leq j \leq m$. We can see that this is equal to the $(i, j)^{th}$ element of $M(S+T)$. Therefore, we have $M(S+T) = M(S)+M(T)$. □

Note that for any $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, we have that

$$AB = [Ab_1 : Ab_2 : \cdots : Ab_p]$$

Thus we can see that

$$[Ax]^T = [x_1a_{*1} + \cdots + x_na_{*n}]^T$$

$$= x_1a_{*1}^T + \cdots + x_na_{*n}^T$$

$$= [x_1, \dots, x_n] \begin{bmatrix} a_{*1}^T \\ \vdots \\ a_{*n}^T \end{bmatrix} = x^TA^T$$

Similarly we have

$$[AB]^T = [Ab_1 : Ab_2 : \cdots : Ab_p]^T$$

$$= \begin{bmatrix} (Ab_1)^T \\ \vdots \\ (Ab_p)^T \end{bmatrix} = \begin{bmatrix} b_1^TA^T \\ \vdots \\ b_p^TA^T \end{bmatrix} = \begin{bmatrix} b_1^T \\ \vdots \\ b_p^T \end{bmatrix} A^T = B^TA^T$$

Lemma 12 *Let $C : V \to W$, $B : W \to U$, and $A : U \to T$. Then we have $A \circ (B \circ C) = (A \circ B) \circ C$ (Association of linear transformations).*

Proof: Let $x \in V$. Thus $C(x) = y$ for a unique $y \in W$. Similarly, we have $B(y) = z$ for a unique $z \in U$ and $A(z) = m$ for a unique $m \in T$.
Let $D = A \circ B$. Thus, by construction, $D(y) = m$.

$$(A \circ (B \circ C))(x) = A \circ B(C(x)) = A \circ B(y) = A(B(y)) = A(z) = m$$

$$((A \circ B) \circ C)(x) = (D \circ C)(x) = D(C(x)) = D(y) = m$$

Therefore we can see that $A \circ (B \circ C) = (A \circ B) \circ C$. □

1.4 Adjoint (Transposes) on an Inner Product Space

Theorem 7 ***(Riesz Representation Theorem)*** *Let $\phi : V \to \mathcal{F}$ ($\mathcal{F}$ is the field). Then there exists a unique $v \in V$ such that $\phi(u) = \langle u, v\rangle$.*

Proof: Let $\{u_1, \dots u_n\}$ be an orthonormal basis for V. Consider $u \in V$. Therefore, we have $u = \langle u, u_1\rangle u_1 + \cdots + \langle u, u_m\rangle u_m$

$$\phi(u) = \phi(\langle u, u_1\rangle u_1 + \cdots + \langle u, u_m\rangle u_m) = \langle u, u_1\rangle \phi(u_1) + \cdots + \langle u, u_m\rangle \phi(u_m)$$

$$= \langle u, \overline{\phi(u_1)} u_1 + \cdots + \overline{\phi(u_m)} u_m\rangle$$

Therefore if we let $v = \overline{\phi(u_1)} u_1 + \cdots + \overline{\phi(u_m)} u_m$, we have $\phi(u) = \langle u, v\rangle$.
Thus all we have to prove is the uniqueness of v. Suppose $v_1, v_2 \in V$ such that $\phi(u) = \langle u, v_1\rangle = \langle u, v_2\rangle \ \forall u \in V$. Thus we have

$$\langle u, v_1 - v_2\rangle = 0 \ \forall u \in V$$

Letting $u = v_1 - v_2 \in V$, we have that

$$\langle v_1 - v_2, v_1 - v_2\rangle = 0 \implies v_1 = v_2$$

Therefore, we proved that it is unique. □

Definition 15 *Let $T \in \mathcal{L}(V, W)$. Choose any $w \in W$. Let $\phi : V \to \mathcal{F}$ such that $\phi(v) = \langle Tv, w\rangle$. Let w^* we the unique vector such that $\phi(v) = \langle v, w^*\rangle \ \forall v \in V$ (Riesz-Representation Theorem). Let T^* be the map such that $T^*(w) = w^* \ \forall w \in W$. Therefore, we have*

$$\langle Tv, w\rangle = \langle v, w^*\rangle = \langle v, T^* w\rangle \ \forall v \in V, w \in W$$

*The mapping T^* is called the **adjoint**.*

Lets explore the properties of the adjoint. Is $T*$ linear?
Let $T^* \in L(W, V)$ thus we have

$$\langle v, T^*(w_1 + w_2)\rangle = \langle Tv, w_1 + w_2\rangle = \langle Tv, w_1\rangle + \langle Tv, w_2\rangle$$

$$= \langle v, T^* w_1\rangle + \langle v, T^* w_2\rangle = \langle v, T^* w_1 + T^* w_2\rangle$$

Now consider $\langle v, T^*(\alpha w)\rangle$.

$$\langle v, T^*(\alpha w)\rangle = \langle Tv, \alpha w\rangle = \overline{\alpha}\langle Tv, w\rangle = \overline{\alpha}\langle v, T^* w\rangle = \langle v, \alpha T^* w\rangle$$

Therefore we can see that $T*$ is linear.

Definition 16 *We call A **self-adjoint** (or **symmetric** in real vector spaces) if $A^* = A$.*

1.5 The Four Fundamental Subspaces

Definition 17 *Let $A \in \mathcal{L}(V, W)$. The **range** (or **column space**) of A is range$(A) = \mathcal{C}(A) = \{Ax | x \in V\} \subseteq W$*

Definition 18 *Let $A \in \mathcal{L}(V, W)$. The **null space** of A is $\mathcal{N}(A) = \{x \in V | Ax = 0\} \subseteq V$*

The four fundamental subspaces are:

1. $\mathcal{C}(A)$
2. $\mathcal{C}(A)^\perp$
3. $\mathcal{N}(A)$
4. $\mathcal{N}(A)^\perp$

Lemma 13 *The range(A) is a subspace of W and nullspace(A) is a subspace of V.*

Proof: Let $A \in \mathcal{L}(V, W)$. By definition, range$(A) = \{Ax|x \in V\}$. Therefore, $Ax = A(x) \in W$ by the construction of A. Therefore, range$(A) \subseteq W$.
Notice that $0 \in V$ and that $A(0) = 0$ (If $Ax = y$, then $Ax = A(x + 0) = A(x) + A(0) = y \implies A(0) = 0$). Therefore, we have that $0 \in$ range(A).
Let $y \in$ range(A). Thus for $A(x) = y$ for some $x \in V$. Let $\alpha \in \mathbb{R}$. Since V is a subspace, $\alpha x \in V$. Thus by the properties of linear transformations, $A(\alpha x) = \alpha A(x) = \alpha y$. Therefore, we have that $y \in$ range$(A) \implies \alpha y \in$ range(A).
Let $y \in$ range(A) and $z \in$ range(A). Therefore we have $Ax_1 = y$ and $Ax_2 = z$ for some $x_1, x_2 \in V$. Since V is a subspace, we know that $(x_1 + x_2) \in V$. From the properties of linear transformations $A(x_1 + x_2) = Ax_1 + Ax_2 = y + z$. Therefore, if $y \in$ range(A) and $z \in$ range(A), then $(y + z) \in$ range(A).
Thus from the properties above, we know that range(A) is a subspace of W.
By definition, nullspace$(A) = \{x \in V|Ax = 0\}$. Therefore, by construction of the nullspace, nullspace$(A) \subseteq V$.
Since $A0 = A(0) = 0$, we have that $0 \in$ nullspace(A).
Suppose that $y \in$ nullspace(A). Let $\alpha \in \mathbb{R}$. Thus by the properties of linear transformations, $A\alpha y = A(\alpha y) = \alpha A(y) = \alpha(0) = 0$. Therefore, we can see that if $y \in$ nullspace(A), then $\alpha y \in$ nullspace(A).
Let $y \in$ nullspace(A) and $z \in$ nullspace(A). By the properties of linear transformations, we have $A(y + z) = A(y) + A(z) = 0 + 0 = 0$. Therefore, if $y \in$ nullspace(A) and $z \in$ nullspace(A), then $(y + z) \in$ nullspace(A).
Therefore, from the properties above, we know that nullspace(A) is a subspace of V. □

Theorem 8 ***(Rank-Nullity Theorem)*** *Let $A \in \mathcal{L}(V, W)$. Thus we have $dim(\mathcal{C}(A))+dim(\mathcal{N}(A)) = dim(V)$.*

Proof: Let $\{x_1, \dots, x_k\}$ be a basis for the null space of A. Therefore, $k = dim(\mathcal{N}(A))$ Since $\mathcal{N}(A) \subseteq V$, we can extend this basis to a basis for V. Let $\{x_1, \dots, x_k, v_1, \dots, v_r\}$ be a basis for V.
Claim: $\{Av_1, \dots, Av_r\}$ is a basis for the column space of A.
Let $x \in \mathcal{C}(A)$. Therefore, $x = Ay$ for some $y \in V$. Thus we can write x as

$$x = A(y) = A(a_1x_1 + \dots + a_kx_k + a_{k+1}v_1 + \dots a_{k+r}v_r)$$

$$= a_1A(x_1) + \dots + a_kA(x_k) + a_{k+1}A(v_1) + \dots a_{k+r}A(v_r)$$

Notice that $A(x_i) = 0$ since $x_i \in \mathcal{N}(A)$ $(1 \leq i \leq k)$ Therefore we can see that

$$x = a_{k+1}A(v_1) + \dots a_{k+r}A(v_r) = a_{k+1}Av_1 + \dots + a_{k+r}Av_r$$

Therefore, we can see that $x \in span(\{Av_1, \dots, Av_r\})$. Therefore, $\mathcal{C}(\mathcal{A}) = span\{Av_1, \dots, Av_r\}$.
All that we have to do is prove that $\{Av_1, \dots, Av_r\}$ is linearly independent.
Suppose that $a_1Av_1 + \dots + a_rAv_r = 0$. Thus we have

$$A(a_1v_1 + \dots + a_rv_r) = 0$$

Thus we can see that $(a_1v_1 + \dots + a_rv_r) \in \mathcal{N}(A)$. Therefore we can say that

$$a_1v_1 + \dots + a_rv_r = b_1x_1 + \dots + b_kx_k \implies a_1v_1 + \dots + a_rv_r - b_1x_1 - \dots - b_kx_k = 0$$

Since $\{x_1, \dots, x_k, v_1, \dots, v_r\}$ is linearly independent, we know that $a_1 = \dots = a_r = b_1 = \dots = b_k = 0$. Therefore, $a_1Av_1 + \dots + a_rAv_r = 0 \implies a_1 = \dots = a_r = 0$, so $\{Av_1, \dots, Av_r\}$ is linearly independent. Therefore, $\{Av_1, \dots, Av_r\}$ is a basis for the column space of A.

Thus we have $dim(\mathcal{C}(A)) + dim(\mathcal{N}(A)) = dim(V)$. □

Let $A \in \mathcal{L}(V, W)$. Let $dim(V) = n$. From this, we have that:

$$x \in \mathcal{N}(A) \iff Ax = 0 \iff \langle y, Ax\rangle = 0 \;\; \forall y \in W \iff \langle A^* y, x\rangle = 0 \;\; \forall y \in W$$

$$\langle A^* y, x\rangle = 0 \;\; \forall y \in W \iff \langle A' y, x\rangle = 0 \;\; \forall y \in W \iff x \perp \mathcal{C}(A) \iff x \in \mathcal{C}(A)^{\perp}$$

Therefore, we can see that $\mathcal{N}(A) = \mathcal{C}(A)^{\perp}$. Thus, we have

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

We also have from the rank nullity theorem that

$$dim(\mathcal{N}(A)) + dim(\mathcal{C}(A)) = n$$

Therefore we can see that

$$dim(\mathcal{C}(A')) = dim(\mathcal{C}(A))$$

Definition 19 *The **rank** of A is defined as $rank(A) = dim(\mathcal{C}(A')) = dim(\mathcal{C}(A))$*

Definition 20 *The **nullity** of A is defined as $nullity(A) = dim(\mathcal{N}(A))$*

Lemma 14 *For any two matrices A and B, we have that $rank(AB) \leq \min\{rank(A), rank(B)\}$.*

Proof: Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Let $x \in \mathcal{C}(AB)$. Thus $x = ABy$ for some $y \in \mathbb{R}^n$. Thus we can see that $x = A(By)$, so $x \in \mathcal{C}(A)$. Therefore, we can say that $\mathcal{C}(AB) \subseteq \mathcal{C}(A)$. Therefore, we can see that $rank(AB) \leq rank(A)$.
Since $rank(A) = rank(A')$, we know that

$$rank(AB) = rank((AB)') = rank(B'A') \leq rank(B') = rank(B)$$

Therefore, we know that $rank(AB) \leq \min\{rank(A), rank(B)\}$.

Theorem 9 *(**Rank Factorization**)Suppose that $A \in \mathbb{R}^{m\times n}$. Let $rank(A) = r$. Let $C \in \mathbb{R}^{m\times r}$ be a matrix such that the columns of C form a basis for $\mathcal{C}(A)$. Then there exists a matrix $R \in \mathbb{R}^{r\times n}$ such that $A = CR$*

Using this theorem, we can prove (again) that $rank(A) = rank(A')$. From the rank factorization theorem, we have that $A' = R'C'$ we know that

$$rank(A') = rank(R'C') \leq rank(R') \leq \# \text{ of columns of } R = r = rank(A)$$

$$\implies rank(A') \leq rank(A)$$

Similarly, using $A = (A')'$, we have

$$rank((A')') \leq rank(A')$$

Therefore, we have that $rank(A) = rank(A')$.

□

Definition 21 *Let A be a square matrix. Let $\mathcal{N}(A) = \{0\}$ (A is full rank). Then there exists a matrix X, called the **inverse** such that $AX = I_n$.*

Lemma 15 *Let A be a square, full rank matrix. Thus there exists an inverse matrix X such that $AX = I_n$. X is also a matrix such that $XA = I_n$.*

Proof: Let $A \in \mathbb{R}^{n\times n}$ such that $\mathrm{N}(A) = \{0\}$. Thus there exists a unique $X \in \mathbb{R}^{n\times n}$ such that $AX = I_n$.

$$AX = I_n \implies (AX)A = (I_n)A$$

$$AXA = A \implies A(XA - I_n) = 0$$

Therefore, we can see that $(XA - I_n) \in \mathrm{N}(A)$. By the construction of A, $(XA - I_n) = 0$. Therefore, we have that $XA = I_n$. □

1.6 Eigenvectors and Eigenvalues

Let $A \in \mathbb{R}^{n\times n}$.Note that powers of a matrix exist $(A, A^2, A^3, \dots, A^n)$. Consider the following polynomial of matrices:

$$P(A) = c_0 I + c_1 A + c_2 A^2 + \dots c_n A^n$$

Similar to how we can decompose an ordinary polynomial, we have

$$P(A) = c_n(A - \lambda_1 I)(A - \lambda_2 I)\dots(A - \lambda_n I)$$

Let $x \neq 0$. Thus we know that $\{x, Ax, A^2x, \dots, A^nx\} \subseteq \mathbb{R}^n$ is a linearly dependent set $(n+1$ vectors).

By the linear dependence lemma, we know that there exists a j $(2 \leq j \leq n)$ such that $A^jx \in span(\{x, Ax, \dots, A^{j-1}x\})$ Let k be the largest such integer.

Thus there exists $c_0, \dots, c_k$ such that $c_0x + c_1Ax + \dots + c_kA^kx = 0$ where $c_k \neq 0$. Therefore we can rewrite this polynomial as

$$c_k(A - \lambda_1 I)\dots(A - \lambda_k I)x = 0$$

Let $u = (A - \lambda_k I)x$. If $u = 0$, then $x \in \mathcal{N}(A - \lambda_k I)$. If $u \neq 0$, then let $w = (A - \lambda_{k-1}I)u$. If $w = 0$, then we can see that $u \in \mathcal{N}(A - \lambda_{k-1}I)$.

If $w \neq 0$, we can continue in this fashion until we find a vector in the null space of $(A - \lambda_j I)$. If after k steps, the algorithm does not terminate, then we have

$$c_k(A - \lambda_1 I)y = 0$$

Since $c_k \neq 0$ and $y \neq 0$, we have $y \in \mathcal{N}(A - \lambda_1 I)$. Therefore any square matrix, there exists a scalar λ and a non-zero vector x such that $x(A - \lambda I) = 0$.

Definition 22 *Let $x(A - \lambda I) = 0$. We call the vector x an **eigenvector** corresponding to λ and the scalar λ an **eigenvalue** . Together we call $\mathcal{E}(A) = \{(\lambda, x) | x \in \mathcal{N}(A - \lambda I), x \neq 0\}$ the set of **eigenpairs**.*

Suppose that $\{\lambda_1, \dots, \lambda_k\}$ are distinct eigenvalues of A $(\lambda_i \neq \lambda_j,\ i \neq j)$. Let (λ_1, x_1), (λ_2, x_2), $\dots, (\lambda_k, x_k)$ be the eigenpairs of A. Then $\{x_1, \dots, x_k\}$ is linearly independent.

Proof: Suppose that $\{x_1, \dots, x_k\}$ is linearly dependent. Thus, $\exists j, 2 \leq j \leq k$ such that $x_j \in span(\{x_1, \dots, x_{j-1}\})$.

Chose the smallest such j. Thus we have that

$$(1)\ \ x_j = c_1x_1 + \dots + c_{j-1}x_{j-1}$$

Multiplying (1) by λ_j

$$\lambda_j x_j = c_1\lambda_j x_1 + \dots + c_{j-1}\lambda_j x_{j-1}$$

Multiplying (1) by A, we get

$$Ax_j = \lambda_j x_j = c_1 A x_1 + \cdots + c_{j-1} A x_{j-1}$$

$$\lambda_j x_j = c_1 \lambda_1 x_1 + \cdots + c_{j-1} \lambda_{j-1} x_{j-1}$$

Thus we have

$$0 = c_1(\lambda_j - \lambda_1)x_1 + \cdots + c_{j-1}(\lambda_j - \lambda_{j-1})x_{j-1}$$

Since $\{x_1, \ldots, x_{j-1}\}$ are linearly independent, we know that $c_i(\lambda_j - \lambda_i)$ must be zero. But since the $\lambda's$ are distinct, we know that $c_1 = \cdots = c_{j-1}$ must be zero. However, by (1), we know that $x_j = 0$. Since x_j is an eigenvector, we know that it cannot be zero (→←).

□

Definition 23 *We say that A is* ***similar*** *to B if there exists a non-singular matrix P such that $A = PBP^{-1}$ or $B = P^{-1}AP$.*

Definition 24 *We say that P is an* ***orthogonal matrix*** *if it is square and the columns of P are orthonormal. Thus $P'P = I_n \implies P' = P^{-1}$. Thus we have $PP' = I_n$ as well.*

Theorem 10 (Schur's Theorem) *Let $A \in \mathbb{R}^{n \times n}$ Then $\exists$ an orthogonal matrix P such that $P'AP = T$, where T is upper triangular. The diagonals of T are precisely the eigenvalues of A.*

Proof: We will start by proving that there exists an orthogonal matrix P such that $P'AP = T$ where T is a upper triangular matrix.
It is clear that the result holds for $n = 1$. Thus assume the result holds for $(n-1) \times (n-1)$ matrices. Thus we have that

$$Av_1 = \lambda_1 v_1 \quad v_1 \neq 0$$

We can extend v_1 to an orthonormal basis for $\mathbb{R}^n$ ($\{v_1, v_2, \ldots, v_n\} = V = [v_1 : \Gamma]$). Thus we have

$$AV = A[v_1 : \Gamma] = [Av_1 : A\Gamma]$$

$$V'AV = \begin{bmatrix} v_1' \\ \Gamma' \end{bmatrix} A \begin{bmatrix} v_1 & \Gamma \end{bmatrix} = \begin{bmatrix} v_1'Av_1 & v_1'A\Gamma \\ \Gamma'Av_1 & \Gamma'A\Gamma \end{bmatrix} = \begin{bmatrix} \lambda_1 & * \\ \mathbf{0} & B \end{bmatrix}$$

Since B is a $(n-1) \times (n-1)$ matrix, by the induction hypothesis, we know that $B = U\tilde{T}U'$, where U is an orthogonal matrix. Therefore, we have

$$V'AV = \begin{bmatrix} \lambda_1 & v_1'A\Gamma \\ \mathbf{0} & U\tilde{T}U' \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ 0 & U \end{bmatrix} \begin{bmatrix} \lambda_1 & * \\ \mathbf{0} & \tilde{T} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & U' \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_1 & * \\ \mathbf{0} & U\tilde{T} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & U' \end{bmatrix} = \begin{bmatrix} \lambda_1 & *U' \\ \mathbf{0} & U\tilde{T}U' \end{bmatrix}$$

Therefore, we know that $* = v_1'A\Gamma U$. Thus we have that

$$V'AV = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & U \end{bmatrix} \begin{bmatrix} \lambda_1 & v_1'A\Gamma U \\ \mathbf{0} & \tilde{T} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & U' \end{bmatrix} = \tilde{U}T\tilde{U}'$$

Thus we can see that $\tilde{U}$ is orthogonal and T is upper triangular.
Now we will prove that the diagonal elements of T are eigenvalues of T. Suppose that $\lambda \notin diag(T) = \{t_{11}, \ldots, t_{nn}\}$. Since λ is not on the diagonal of T, we know that $(T - \lambda I)x = 0 \implies x = 0$ since all diagonal elements are non-zero. Thus $\mathcal{N}(T - \lambda I) = \{0\}$, which would imply that the eigenvector would be $\mathbf{0}$, which is a contradiction (→←).

Now suppose that $\lambda = t_{jj} \implies T - \lambda I$ has at least one diagonal element equal to zero. Thus, there is a free variable, which implies that there exist $x \neq 0$ such that $(T - \lambda I)x = 0$. Therefore we know that (λ, x) is an eigenpair of T. Using lemma 16, since T and A are similar matrices, we know that they have the same eigenvalues.

□

Lemma 16 *Let A and B be $n \times n$ matrices such that $P^{-1}AP = B$ (A similar to B). Prove that $dim(\mathcal{N}(A - \lambda I)) > 0 \iff dim(\mathcal{N}(B - \lambda I)) > 0$.*

Proof: ($\implies$) Let $P^{-1}AP = B$. Suppose $dim(\mathcal{N}(A - \lambda I)) > 0$. Let $x \in \mathcal{N}(A - \lambda I)$ $(x \neq 0)$. Thus $(A - \lambda I)x = 0$. Thus we have

$$P^{-1}(A - \lambda I)x = (P^{-1}A - \lambda P^{-1})x = 0$$

Since $x = PP^{-1}x$, we have

$$(P^{-1}A - \lambda P^{-1})x = (P^{-1}A - \lambda P^{-1})PP^{-1}x = (P^{-1}AP - \lambda P^{-1}P)P^{-1}x = (B - \lambda I)P^{-1}x = 0$$

Therefore $P^{-1}x \in \mathcal{N}(B - \lambda I)$. Note that $P^{-1}x \neq 0$ (If it was, then $PP^{-1}x = x = 0$, which by construction is not possible). Therefore, we know that $dim(\mathcal{N}(B - \lambda I)) > 0$.
($\impliedby$) Now suppose $dim(\mathcal{N}(B - \lambda I)) > 0$. Let $x \in \mathcal{N}(B - \lambda I)$ $(x \neq 0)$. Thus we have $(B - \lambda I)x = (P^{-1}AP - \lambda I)x = 0$.

$$(P^{-1}AP - \lambda I)x = 0 \implies P(P^{-1}AP - \lambda I)x = (AP - \lambda P)x = 0$$

We can see that $x = P^{-1}Px$. Thus we have

$$(AP - \lambda P)x = (AP - \lambda P)P^{-1}Px = (A - \lambda I)Px = 0$$

Therefore, we can see that $Px \in \mathcal{N}(A - \lambda I)$, where $Px \neq 0$. Therefore we see that $dim(\mathcal{N}(B - \lambda I)) > 0$.
Therefore $dim(\mathcal{N}(A - \lambda I)) > 0 \iff dim(\mathcal{N}(B - \lambda I)) > 0$. □

Theorem 11 (Spectral Theorem) *Let A be a square and symmetric matrix. There exists an orthogonal P such that $P'AP = \Lambda$, where Λ is diagonal.*

Proof: From Schur's Theorem, we have that

$$P'AP = T$$

where T is an upper triangular matrix. However, notice that

$$(P'AP)' = P'A'P = P'AP = T'$$

since A is symmetric. Thus we have that

$$P'AP = T' = T = \Lambda$$

Since T is upper triangular and T' is lower triangular, we know that Λ must be diagonal. □

Definition 25 *We call a matrix **positive definite** if A is real, symmetric, and all its eigenvalues are positive.*

Note an alternate definition may be that a positive definite matrix is defined as any matrix A such that $x'Ax > 0 \ \forall x \in \mathbb{R}^n \setminus \{0\}$. We can see that this is an equivalent definition since we have

$$x'Ax = x'P'\Lambda Px = y'\Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2$$

where $y = Px$. Thus we can see that this will be greater than zero for all $x \neq \mathbf{0}$.

Why must covariance matrices be positive definite (positive semi-definite)?

Let z be a random vector in $\mathbb{R}^n$. We know that $cov(\alpha' z) = \alpha'\Sigma_z\alpha$. Since $cov(\alpha' z) > 0$, we have that $\alpha'\Sigma_z\alpha > 0$. Therefore, we can see that Σ_z must be positive definite.

Theorem 12 (Cholesky Decomposition) *Let A be a positive semi-definite matrix. Then there exists a decomposition of A such that $A = LL'$ where L is a lower triangular matrix.*

Proof: We will first start with a matrix A that is positive semi-definite. Consider the Spectral Decomposition of A

$$A = P'\Lambda P = P'\Lambda^{1/2}\Lambda^{1/2}P = BB'$$

where $B = P'\Lambda^{1/2}$. Note that we can take the square root of Λ since it is a diagonal matrix with positive elements. We can then use a QR-Decomposition on B'

$$B' = QR \implies BB' = R'Q'QR = R'R = LL'$$

Since R is upper triangular, we know that L is lower triangular. □

1.7 Singular Value Decomposition, Generalized Inverses, and PCA

Consider the structure of $A \in \mathbb{R}^{m\times n}$. Let $\{v_1, \dots, v_n\}$ be an orthonormal basis for $\mathbb{R}^n$. Suppose we wish to form vectors in the column space of A, where A acts on the V_j's. We wish to define

$$Av_i = \sigma_i u_i \ \text{ where } \sigma_i = ||Av_i||$$

Thus we can see that $u_i = \frac{Av_i}{\sigma_i}$ (if well defined). When is u_i not well defined? We can see that it is not well defined when $\sigma_i = 0$ or $Av_i = 0$.
Suppose that we are able to find $Av_i = \sigma_i u_i$ for $i = 1, \dots, r$ $(r \le n)$ such that $\sigma_i \neq 0 \ \ i = 1, \dots, r$. Can we choose orthonormal v_i's such that $\{u_1, \dots, u_r\}$ is also an orthonormal set? Thus we have that

$$\langle u_i, u_j\rangle = \langle \frac{1}{\sigma_i}Av_i, \frac{1}{\sigma_j}Av_j\rangle = \frac{1}{\sigma_i\sigma_j}\langle Av_i, Av_j\rangle = \frac{1}{\sigma_i\sigma_j}v_j'A'Av_i = 0$$

Note that $A'A$ is real and symmetric, thus by the Spectral Theorem we have

$$A'AP = P\Lambda$$

Let v_j be the j^{th} column of P. Thus we have

$$A'Av_j = \lambda_j v_j \ \ j = 1, \dots, n$$

Thus we have

$$v_i'A'Av_j = v_i'\lambda_j v_j = \lambda_j\langle v_i, v_j\rangle = 0$$

Therefore, by letting v_j be the j^{th} column of P, we have that $\{u_1, \ldots, u_r\}$ is also an orthonormal set. Thus putting it all together, we have

$$A\begin{bmatrix} v_1 & \ldots & v_m \end{bmatrix} = \begin{bmatrix} u_1 & \ldots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_r \end{bmatrix}$$

Theorem 13 (Singular Value Decomposition) *Let $A \in \mathbb{R}^{m\times n}$ that has rank r. Thus we can decompose A as $A = UDV'$ Where U and V are orthogonal matrices and D is a diagonal matrix. We can write it as*

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix}$$

Where $U_1 \in \mathbb{R}^{m\times r}$ are the eigen vectors of AA' corresponding to the non-zero eigenvalues. $U_2 \in \mathbb{R}^{m\times m-r}$ are the eigen vectors of AA' corresponding to the zero eigenvalues (also can be thought of as a basis extension). $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing the square root of the non-zero eigenvalues of $A'A$ or AA'. $V_1 \in \mathbb{R}^{n\times r}$ are the eigen vectors of $A'A$ corresponding to the non-zero eigenvalues. $V_1 \in \mathbb{R}^{n\times n-r}$ are the eigen vectors of $A'A$ corresponding to the zero eigenvalues.

Proposition: $\mathcal{C}(U_1) = \mathcal{C}(A)$ and $\mathcal{C}(V_2) = \mathcal{N}(A)$

Proof: We can start with the fact that

$$A = U_1\Sigma V_1' \implies U_1 = AV_1\Sigma^{-1}$$

Let $w \in \mathcal{C}(A)$. Thus $w = Av$ for some $v \in \mathbb{R}^n$. Since $V = [V_1 : V_2]$ is a basis for $\mathbb{R}^n$, we have that $v = \sum_{i=1}^n \alpha_i v_i$ thus we have

$$Av = \sum_{i=1}^n \alpha_i Av_i = \sum_{i=1}^r \alpha_i Av_i$$

since $Av_j = 0$ for $j > r$. Thus we have that

$$AV_1\alpha = U_1\Sigma\alpha \in \mathcal{C}(U_1)$$

Therefore we have that $\mathcal{C}(A) \subseteq \mathcal{C}(U_1)$. It is apparent that $\mathcal{C}(U_1) \subseteq \mathcal{C}(A)$ since $A = U_1\Sigma V_1' \implies U_1 = AV_1\Sigma^{-1}$. Thus we have that $\mathcal{C}(U_1) = \mathcal{C}(A)$.

Now suppose that $w \in \mathcal{N}(A)$. Therefore, we know that $Aw = 0$ ($w \in \mathbb{R}^n$). Thus we have that $w = \sum_{i=1}^n \alpha_i v_i$. Thus,

$$Aw = 0 \implies \sum_{i=1}^n \alpha_i Av_i = \sum_{i=1}^r \alpha_i Av_i = 0$$

Thus we have that

$$\sum_{i=1}^r \alpha_i\sigma_i u_i = 0 \implies \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

Therefore,

$$w = \sum_{i=1}^n \alpha_i v_i = \sum_{i=r+1}^n \alpha_i v_i \in \mathcal{C}(V_2)$$

Therefore, we have that $\mathcal{N}(A) \subseteq \mathcal{C}(V_2)$. From the original SVD, we have that $[AV_1 : AV_2] = [U_1\Sigma : \mathbf{0}]$. Thus we can see that $\mathcal{C}(V_2) \subseteq \mathcal{N}(A)$. Therefore, $\mathcal{C}(V_2) = \mathcal{N}(A)$. □

From SVD, we have that $A = UDV'$ which means that $A' = VD'U'$. Thus by analogy, we have that $\mathcal{C}(A') = \mathcal{C}(V_1)$ and $\mathcal{N}(A') = \mathcal{C}(U_2)$.

Thus we have the following:

1. Columns of U_1 create an orthonormal basis for $\mathcal{C}(A)$
2. Columns of U_2 create an orthonormal basis for $\mathcal{C}(A)^\perp = \mathcal{N}(A')$
3. Columns of V_1 create an orthonormal basis for $\mathcal{C}(A')$
4. Columns of V_2 create an orthonormal basis for $\mathcal{C}(A')^\perp = \mathcal{N}(A)$

Using these properties, we can easily find the following orthogonal projectors:

1. $P_A = U_1U_1'$
2. $P_{A'} = V_1V_1'$
3. $P_{\mathcal{N}(A)} = V_2V_2'$
4. $P_{\mathcal{N}(A')} = U_2U_2'$

Definition 26 *Let $A \in \mathbb{R}^{n\times m}$. A^- is the* ***generalized inverse*** *of A if $AA^-A = A$.*

There are an infinite number of generalized inverses in general, and many ways to construct them. One way would be to consider the full-rank factorization of A. Thus we have

$$A = CR$$

Thus we can see that

$$CRR'(RR')^{-1}(C'C)^{-1}C'CR = CR$$

Thus we can see that one possible generalized inverse is $A^- = R'(RR')^{-1}(C'C)^{-1}C'$.

Definition 27 *Let $A \in \mathbb{R}^{n\times m}$. A^+ is the* ***Moore-Penrose inverse*** *of A if the following conditions hold:*

1. $AA^+A = A$

2. $A^+AA^+ = A^+$

3. $(AA^+)^* = (AA^+)$

4. $(A^+A)^* = A^+A$

One property of the Moore-Penrose (MP) inverse is that it always exists and is unique. One way to get the MP inverse is to use the SVD of a matrix. Consider a matrix $A \in \mathbb{R}^{n\times m}$. From SVD, we have

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix} = UDV'$$

Consider setting $A^+ = VD^+U'$, where $D^+ = \begin{bmatrix} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$. Thus, we can see that

$$AA^+A = UDV'VD^+U'UDV' = UDD^+DV$$

Since

$$DD^+D = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = D,$$

we have that $AA^+A = UDV' = A$. One can check that the other 3 properties hold as well.

Suppose we have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ of rank r, where n is the number of samples and p are the number of variables. Suppose that we wish to reduce the dimension of the problem (to $k < p$). Suppose that $\mathbf{X}$ is mean-centered. Then we can see that the covariance matrix is $\mathbf{C} = \mathbf{X}'\mathbf{X}/(n-1)$. Consider the SVD of X. Thus we have

$$X = UDV' = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix} \text{ where } \Sigma = diag(\sigma_1, \dots, \sigma_r), \;\; \sigma_1 > \dots > \sigma_r$$

. Consider trying to find a linear combination of the columns of X such that the variance of that variable is maximized. We can see that it is equivalent to the following maximization problem:

$$max_v ||Xv|| \;\; s.t. \;\; ||v||^2 = 1$$

Notice that V creates a basis for $\mathbb{R}^n$. Thus we can see that $v \in span(V)$, and

$$v = c_1v_1 + c_2v_2 + \dots + c_pv_p$$
$$||v||^2 = 1 \implies c_1^2 + c_2^2 + \dots + c_p^2 = 1$$

Thus we have

$$Av = c_1Xv_1 + c_2Xv_2 + \dots + c_nXv_n = c_1\sigma_1u_1 + \dots + c_r\sigma_ru_r$$
$$||Xv||^2 = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \dots + c_r^2\sigma_r^2 \le (c_1^2 + \dots + c_r^2)\sigma_1^2 \le \sigma)_1^2$$

We can see that we achieve this upper bound when $c_1 = 1$ and $c_2 = \dots c_r = 0$. Thus we have that v_1 maximizes this problem. We call Xv_1 the first principal component, and v_1 the loading vector. We can see that maximizing $||Xv||^2$ will maximize the sample covariance of this principal component. We can see that this principal component will roughly account for $\sigma_1^2 / \sum_{i=1}^n \sigma_i^2$ proportion of the total variance. When talking about the total variation of the data X, we will be referring to the $trace(X)$ (the sum of the variances of each column). Since we know that the trace is the sum of the diagonal elements, or the sum of the eigenvalues of $X'X/(n-1)$, we know that $V'XX'V/(n-1)$ will have the same eigenvalues.

$$V'XX'V = V'UD'V'VDU'V = (V'U)D'D(V'U)'$$

Thus we can see that the corresponding variance explained by the j^{th} PC is $\sigma_j^2 / \sum_{i=1}^r \sigma_i^2$. We can find the second principle component by solving the following maximization problem:

$$max_v ||Xv||^2 \;\; s.t. \;\; v \perp v_1, ||v||^2 = 1$$

The details are worked out in problem 38.

In general, we can just use the SVD of X to calculate the principal components. The matrix V are the loadings, and XV are the principal components. If we wish to approximate a matrix X by a rank z matrix, then we can use the SVD decomposition. Notice that the SVD decomposition can be broken down into a series of rank-one updates

$$X = \sum_{i=1}^r \sigma_i u_i' v_i$$

where u_i' is the i^{th} row of U and v_i is the i^{th} column of V. Thus we can use the first z sums to have a rank z approximation of X. This method will preserve $\sum_{i=1}^z \sigma_i^2 / \sum_{i=1}^r \sigma_i^2$ proportion of the variance.

1.8 Determinants, Partition Matrices and Other Useful Lemmas

Definition 28 *The **Determinant** of a matrix* $\mathbf{A} \in \mathbb{R}^{p \times p}$ *is the product of the eigenvalues.*

The following are some basic properties of determinants

1. $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are square
2. $|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$

Theorem 14 (Simultaneous Diagonalization Theorem) *Let* $\mathbf{A}$ *be positive definite and* $\mathbf{B}$ *be positive semi-definite. Then there exists a non-singular matrix* $\mathbf{U}$ *such that* $\mathbf{U}'\mathbf{AU} = \mathbf{I}$ *and* $\mathbf{U}'\mathbf{BU} = \mathbf{D}$ *where* $\mathbf{D}$ *are the eigenvalues of* $\mathbf{BA}^{-1}$.

Proof: Consider the Cholesky decomposition of $\mathbf{A}$, $\mathbf{A} = \mathbf{LL}^T$, where $\mathbf{L}$ is lower triangular and is also positive definite (since $\mathbf{A}$ is positive definite).
Therefore, $\mathbf{L}^{-1}$ exists, and $\mathbf{L}^{-1}\mathbf{A}\left(\mathbf{L}^{-1}\right)^T = \mathbf{I}$. Consider the spectral decomposition of $\mathbf{L}^{-1}\mathbf{B}\left(\mathbf{L}^{-1}\right)^T$. Therefore, we have $\mathbf{TDT}^T = \mathbf{L}^{-1}\mathbf{B}\left(\mathbf{L}^{-1}\right)^T$, where $\mathbf{TT}^T = \mathbf{I}$, and $\mathbf{D}$ is diagonal. Let $\mathbf{U} = \left(\mathbf{L}^{-1}\right)^T \mathbf{T}$. Therefore, we have:

$$\mathbf{U}^T\mathbf{BU} = \mathbf{D}$$

$$\mathbf{U}^T\mathbf{AU} = \mathbf{T}^T\mathbf{L}^{-1}\mathbf{LL}^T\left(\mathbf{L}^T\right)^{-1}\mathbf{T} = \mathbf{T}^T\mathbf{T} = \mathbf{I}$$

Therefore, $\mathbf{U} = \left(\mathbf{L}^{-1}\right)^T \mathbf{T}$.
Claim: The non-zero eigenvalues of $\mathbf{BA}$ are the same as the non-zero eigenvalues of $\mathbf{AB}$.

$$|\mathbf{BA} - \mathbf{I}\lambda| = (-\lambda)^n \left|\mathbf{I} - \frac{1}{\lambda}\mathbf{BA}\right| = (-\lambda)^n \left|\mathbf{I} - \frac{1}{\lambda}\mathbf{AB}\right| = |\mathbf{AB} - \mathbf{I}\lambda|$$

Therefore, we can see that $\mathbf{D}$ are the eigenvalues of $\mathbf{L}^{-1}\mathbf{B}\left(\mathbf{L}^{-1}\right)^T$, which are the eigenvalues of $\mathbf{B}\left(\mathbf{L}^{-1}\right)^T\mathbf{L}^{-1} = \mathbf{BA}^{-1}$. Since $\mathbf{L}^{-1}\mathbf{B}\left(\mathbf{L}^{-1}\right)^T$ and $\mathbf{BA}^{-1}$, are the same size, we know that $\mathbf{D}$ are the eigenvalues of $\mathbf{BA}^{-1}$. □

Theorem 15 (Sherman–Morrison–Woodbury formula) *For all conformable matrices* $\mathbf{A}, \mathbf{U}, \mathbf{C}, \mathbf{V}$, *we have that*

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$$

Lemma 17 *Consider the following block matrix:*

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$$

We can invert this matrix in one of the following ways

1.

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1} & (\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1} \end{bmatrix}$$

2.

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{BD}T^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1} \end{bmatrix}$$

2 Multivariate, Quadratic Forms, and Other Distributions

2.1 Multivariate Normal

Definition 29 *We say that* $\mathbf{x}$ *has a multivariate normal distribution if it has the following pdf:*

$$f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

What is the MGF of $\mathbf{y} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$?

$$\Psi_{\mathbf{y}}(\mathbf{t}) = \mathbb{E}(e^{\mathbf{t}'\mathbf{y}})$$

$$= \int e^{\mathbf{t}'\mathbf{y}} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\right\} d\mathbf{y} = \int \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - 2\mathbf{t}'\mathbf{y})\right\} d\mathbf{y}$$

$$= exp\left\{\frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right\} \int \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{t}\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{t}\boldsymbol{\Sigma})\right\} d\mathbf{y}$$

Therefore, we have that

$$\Psi_{\mathbf{y}}(\mathbf{t}) = exp\left\{\frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right\}$$

If $\boldsymbol{\mu} \neq \mathbf{0}$, let $\mathbf{x} = \mathbf{y} - \boldsymbol{\mu}$. Then $\mathbb{E}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{x} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$.

$$\Psi_{\mathbf{y}}(\mathbf{t}) = \mathbb{E}(e^{\mathbf{t}'\mathbf{y}}) = \mathbb{E}(e^{\mathbf{t}'(\mathbf{x}+\boldsymbol{\mu})}) = \mathbb{E}(e^{\mathbf{t}'\mathbf{x}}) + e^{\mathbf{t}'\boldsymbol{\mu}}$$

Thus we have

$$\Psi_{\mathbf{x}}(\mathbf{t}) = exp\left\{\frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t} + \mathbf{t}'\boldsymbol{\mu}\right\}$$

Suppose $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. What is the distribution of $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$?

$$\Psi_{\mathbf{y}}(\mathbf{t}) = \mathbb{E}(e^{\mathbf{t}'\mathbf{y}}) = \mathbb{E}(e^{\mathbf{t}'(\mathbf{A}\mathbf{x}+\mathbf{b})}) = \mathbb{E}(e^{\mathbf{t}'\mathbf{A}\mathbf{x}})e^{\mathbf{t}'\mathbf{b}} = \Psi_{\mathbf{y}}(\mathbf{A}'\mathbf{t})e^{\mathbf{t}'\mathbf{b}}$$

$$e^{\boldsymbol{\mu}'\mathbf{A}'\mathbf{t}+\frac{1}{2}\mathbf{t}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{t}}e^{\mathbf{t}'\mathbf{b}} = e^{(\mathbf{A}\boldsymbol{\mu}+\mathbf{b})'\mathbf{t}+\frac{1}{2}\mathbf{t}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{t}}$$

Thus we can see that $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. How do we guarantee that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A} > 0$? We want to show that $\mathbf{x}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{x} > 0 \ \forall \mathbf{x} \neq \mathbf{0} \implies \mathbf{y}'\boldsymbol{\Sigma}\mathbf{y} > 0 \ \forall \mathbf{y} \ s.t. \ \mathbf{x} \neq \mathbf{0}$. Thus we need to argue that $\mathbf{x} \neq \mathbf{0} \iff \mathbf{y} \neq \mathbf{0}$. We can see that if $\mathbf{x} = \mathbf{0}$, then $\mathbf{y} = \mathbf{0}$. Suppose that $\mathbf{x} \neq \mathbf{0} \implies \mathbf{y} \neq \mathbf{0}$.

Thus we have $\mathbf{y} = \mathbf{A}'\mathbf{x} = \mathbf{0}$. How can we arrive at a contradiction? Consider $(\mathbf{A}\mathbf{A}')^{-1}\mathbf{A}\mathbf{A}'\mathbf{x} = \mathbf{x} = \mathbf{0}$. Thus we arrive at a contradiction, and we know that $\mathbf{A}$ must be full row-rank. Therefore, we have that $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ if $\mathbf{A}$ is full row rank (If not, then we have a rank deficient distribution, which means that the pdf does not exist, but the distribution is still valid).

Theorem 16 (Cramer Wald's device) $\mathbf{x}$ *is multivariate normal iff* $\mathbf{a}'\mathbf{x} \sim \mathcal{N}(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ *for all* $\mathbf{a} \neq \mathbf{0}$.

Normal distributions have the following properties:

1. If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mathbf{T}$ is an orthogonal matrix, then we have $\mathbf{Tx} \sim \mathcal{N}(\mathbf{T}\boldsymbol{\mu}, \mathbf{TIT}' = \mathbf{I})$
2. Subsets of $\mathbf{x}$ are also multivariate normal
3. Uncorrelated $\implies$ Independence

Note: The reverse of number 2 is not true. If all subsets of $\mathbf{x}$ are multivariate normal, that does not imply that $\mathbf{x}$ is multivariate normal.

Now we will look at the conditional distribution of a multivariate random vector.

Lemma 18 $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$ *if and only if* $y_1|y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mu_2 - y_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ *and* $y_2 \sim N(\mu_2, \Sigma_{22})$

Proof: $(\implies)$
Suppose $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$. Consider the following transformation

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Thus we have

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} y_1 - \Sigma_{12}\Sigma_{22}^{-1} y_2 \\ y_{22} \end{bmatrix}$$

Using basic properties of the normal distribution, we know that

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix}'\right)$$

Simplifying, we have that

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right)$$

Since $cov(z_1, z_2) = 0$ we know that $z_1 \perp\!\!\!\perp z_2$ (property of Normal distribution). Consider

$$z_1|z_2 = \frac{f_{z_1,z_2}(z_1, z_2)}{f_{z_2}(z_2)} = \frac{f_{z_1}(z_1)f_{z_2}(z_2)}{f_{z_2}(z_2)} = f_{z_1}(z_1)$$

Thus

$$y_1 - \Sigma_{12}\Sigma_{22}^{-1}y_2|y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Thus we can see that

$$y_1|y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mu_2 - y_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Since $\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ has a normal distribution, we know that $z_2 = y_2 \sim N(\mu_2, \Sigma_{22})$. Therefore, we have if $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$, then $y_1|y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mu_2 - y_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ and $y_2 \sim N(\mu_2, \Sigma_{22})$.
$(\impliedby)$

Suppose we have that $y_1|y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mu_2 - y_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ and $y_2 \sim N(\mu_2, \Sigma_{22})$. We know that $f_{y_1,y_2}(y_1, y_2) = f_{y_1|y_2}(y_1|y_2)f_{y_2}(y_2)$. Thus we have that

$$(1) f_{y_1,y_2}(y_1, y_2) \propto exp\{-\frac{1}{2}(y_1-\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\mu_2-y_2))'(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(y_1-\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\mu_2-y_2))-\frac{1}{2}y_2'\Sigma_{22}^{-1}y_2$$

Note that we can invert the $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ by using the fact that

$$A^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

Thus we have that

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}$$

Let $\psi = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}$ and $\gamma = \Sigma_{12}\Sigma_{22}^{-1}$. Thus we have that

$$\Sigma^{-1} = \begin{bmatrix} \psi & -\psi\gamma \\ \gamma'\psi & \Sigma_{22}^{-1} + \gamma'\psi\gamma \end{bmatrix}$$

Consider $Q = \begin{bmatrix} (y_1 - \mu_1)' & (y_2 - \mu_2)' \end{bmatrix} \Sigma^{-1} \begin{bmatrix} (y_1 - \mu_1) \\ (y_2 - \mu_2) \end{bmatrix}$

$$= (y_1 - \mu_1)'\psi(y_1 - \mu_1) - 2(y_1 - \mu_1)'\psi\gamma(y_2 - \mu_2) + (y_2 - \mu_2)'\Sigma_{22}^{-1}(y_2 - \mu_2) + (y_2 - \mu_2)'\gamma'\psi\gamma(y_2 - \mu_2)$$

$$((y_1 - \mu_1) - \gamma(y_2 - \mu_2))'\psi((y_1 - \mu_1) - \gamma(y_2 - \mu_2)) + (y_2 - \mu_2)'\Sigma_{22}^{-1}(y_2 - \mu_2)$$

Notice from (1), that we have

$$f_{y_1,y_2}(y_1, y_2) \propto exp\{-\frac{1}{2}Q\}$$

Thus we have

$$f_{y_1,y_2}(y_1, y_2) \propto exp\left\{-\frac{1}{2}\begin{bmatrix} (y_1 - \mu_1)' & (y_2 - \mu_2)' \end{bmatrix} \Sigma^{-1} \begin{bmatrix} (y_1 - \mu_1) \\ (y_2 - \mu_2) \end{bmatrix}\right\}$$

We can see that this the kernel of a normal distribution, thus we know that

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

Therefore, $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$ if and only if $y_1|y_2 \sim N(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mu_2 - y_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ and $y_2 \sim N(\mu_2, \Sigma_{22})$. □

Now we will explore the independence of normal random variables.

Lemma 19 *Suppose that* $\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$. *Then we have that* $\mathbf{y}_1 \perp\!\!\!\perp \mathbf{y}_2$ *if and only if* $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Proof: We have that

$$\Psi_{\mathbf{y}}(\mathbf{t}) = exp\left\{\mathbf{t}_1'\boldsymbol{\mu}_1 + \mathbf{t}_2'\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_1'\boldsymbol{\Sigma}_{11}\mathbf{t}_1 + \frac{1}{2}\mathbf{t}_2'\boldsymbol{\Sigma}_{22}\mathbf{t}_2 + \mathbf{t}_1'\boldsymbol{\Sigma}_{12}\mathbf{t}_2\right\}$$

We can see that this factors into two normal mgfs if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$

$$\Psi_{\mathbf{y}_1}(\mathbf{t}_1)\Psi_{\mathbf{y}_2}(\mathbf{t}_2) = exp\left\{\mathbf{t}_1'\boldsymbol{\mu}_1 + \frac{1}{2}\mathbf{t}_1'\boldsymbol{\Sigma}_{11}\mathbf{t}_1\right\} + exp\left\{\mathbf{t}_2'\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_2'\boldsymbol{\Sigma}_{22}\mathbf{t}_2\right\}$$

□

Lemma 20 *Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and define $\mathbf{U} = \mathbf{AY}$, $\mathbf{V} = \mathbf{BY}$. Then $\mathbf{U}$ and $\mathbf{V}$ are independent if and only if $cov(\mathbf{U}, \mathbf{V}) = \mathbf{A\Sigma B}' = \mathbf{0}$.*

Proof: Let $\mathbf{W} = \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\mathbf{y}$. Thus we have that

$$cov(\mathbf{W}) = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\boldsymbol{\Sigma}\begin{bmatrix}\mathbf{A}' & \mathbf{B}'\end{bmatrix} = \begin{bmatrix}\mathbf{A\Sigma A}' & \mathbf{A\Sigma B}'\\ \mathbf{B\Sigma A}' & \mathbf{B\Sigma B}'\end{bmatrix}$$

From lemma 19, we know that $\mathbf{A}$ and $\mathbf{B}$ are independent if and only if $\mathbf{A\Sigma B}' = \mathbf{0}$. □

2.2 Quadratic Forms

Definition 30 *Let $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{p\times p}$. We say that $Q = \mathbf{x}'\mathbf{Ax}$ is a **quadratic form** in $\mathbf{A}$.*

Suppose $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Lets take a look at the expectation of Q.

$$\mathbb{E}(Q) = \mathbb{E}(\mathbf{x}'\mathbf{Ax}) = \mathbb{E}((\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu})'\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu})) = \mathbb{E}\left\{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + 2(\mathbf{x}-\boldsymbol{\mu})\mathbf{A}\boldsymbol{\mu}\right\}$$

$$= \mathbb{E}\left\{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right\} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \mathbb{E}\left\{tr((\mathbf{x}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}))\right\} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

$$\mathbb{E}\left\{tr(\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})')\right\} + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = tr(\mathbf{A\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$

Therefore, we have if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbb{E}(\mathbf{x}'\mathbf{Ax}) = tr(\mathbf{A\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

Lemma 21 *If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then we have that $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_p$.*

Theorem 17 (Fundamental Theorem of Quadratic Forms) *Let $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $Q = \mathbf{y}'\mathbf{Ay}$. Then $Q \sim \chi^2_r \iff \mathbf{A}^2 = \mathbf{A}$ and $rank(\mathbf{A}) = r$.*

Proof: ($\Longrightarrow$) Suppose that $Q \sim \chi^2_r$. Then we have that

$$\Psi_Q(t) = \frac{1}{(1-2t)^{r/2}} = \mathbb{E}(e^{tQ})$$

$$= \int e^{t\mathbf{y}'\mathbf{Ay}}\frac{e^{-\frac{1}{2}\mathbf{y}'\mathbf{y}}}{(\sqrt{2\pi})^p}d\mathbf{y}$$

$$= \int \frac{e^{-\frac{1}{2}\mathbf{y}'(\mathbf{I}-2t\mathbf{A})\mathbf{y}}}{(\sqrt{2\pi})^p}d\mathbf{y}$$

$$= \int \frac{e^{-\frac{1}{2}\mathbf{y}'((\mathbf{I}-2t\mathbf{A})^{-1})^{-1}\mathbf{y}}}{(\sqrt{2\pi})^p|\mathbf{I}-2t\mathbf{A}|^{-1/2}}d\mathbf{y}|\mathbf{I}-2t\mathbf{A}|^{-1/2}$$

$$= |\mathbf{I} - 2t\mathbf{A}|^{-1/2}$$

Thus we have that

$$(1-2t)^r = |\mathbf{I} - 2t\mathbf{A}| = |\mathbf{T}'\mathbf{T} - 2\mathbf{t}\mathbf{T}\mathbf{D}\mathbf{T}'| = |\mathbf{T}\mathbf{T}'|^2|\mathbf{I} - 2t\mathbf{D}|$$

Thus we have

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

This equation holds for all $t \in N_\epsilon(0)$ and by the fundamental theorem of algebra it has p roots. We can see that the only way that can happen is if $\mathbf{A}$ has r eigenvalues equal to 1, and the rest equal to zero. This means that $\mathbf{A}^2 = \mathbf{A}$ and $rank(\mathbf{A}) = r$.
$() \Longleftarrow)$ Suppose that $\mathbf{A}^2 = \mathbf{A}$ and $rank(\mathbf{A}) = r$. Then we have that

$$Q = \mathbf{y}'\mathbf{A}\mathbf{y} = \mathbf{y}'\mathbf{T}\mathbf{D}\mathbf{T}'\mathbf{y} = \mathbf{z}\mathbf{D}\mathbf{z}$$

Notice that $\mathbf{z} = \mathbf{T}'\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus we know that the $\mathbf{z}_i$'s are independent and have a standard normal distribution. Therefore, we have

$$\mathbf{z}'\mathbf{D}\mathbf{z} = \sum_{i=1}^{p} z_i^2\lambda_i = \sum_{i=1}^{r} z_i^2\lambda_i = \sum_{i=1}^{r} z_i^2$$

Therefore, we know that $Q \sim \chi^2_r$. □

What if $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$? Let $\mathbf{z} = \mathbf{y}\mathbf{\Sigma}^{-1/2} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then we have

$$Q = \mathbf{y}'\mathbf{A}\mathbf{y} = \mathbf{z}'\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}\mathbf{z}$$

From the Fundamental theorem of Quadratic Forms, we know that $Q \sim \chi^2_r \iff \mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$ or in other words, $\mathbf{A}\mathbf{\Sigma}\mathbf{A} = \mathbf{A}$.

Lemma 22 *If* $\mathbf{U}$ *and* $\mathbf{V}$ *are two independent normal random variables, then* $\mathbf{U} \perp\!\!\!\perp \mathbf{V}'\mathbf{V}$ *and* $\mathbf{U}'\mathbf{U} \perp\!\!\!\perp \mathbf{V}'\mathbf{V}$.

Example: Lets prove that $\bar{\mathbf{y}} \perp\!\!\!\perp \mathbf{S}^2$, where $\mathbf{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})$ and $y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that $\bar{y} = \frac{1}{n}\mathbf{1}'\mathbf{y}$ and $S^2 = \mathbf{y}'(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{y}$. Let $\mathbf{U} = \frac{1}{n}\mathbf{1}'\mathbf{y}$ and $\mathbf{V} = (\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{y}$. Thus we have

$$cov(\mathbf{U}, \mathbf{V}) = cov(\frac{1}{n}\mathbf{1}'\mathbf{y}, (\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{y})$$

$$= \frac{1}{n}\mathbf{1}'cov(\mathbf{y})(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}) = \frac{1}{n}\mathbf{1}'(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})$$

$$= \frac{1}{n}\mathbf{1}' - \frac{\mathbf{1}'\mathbf{1}\mathbf{1}'}{n^2} = \frac{1}{n}\mathbf{1}' - \frac{1}{n}\mathbf{1}' = 0$$

From lemma 22, we know that $\bar{\mathbf{y}} = \mathbf{U} \perp\!\!\!\perp \mathbf{V}'\mathbf{V} = S^2$.

Theorem 18 *Let* $Q_i = \mathbf{y}'\mathbf{A}_i\mathbf{y}$ *where* $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. *If* $Q_i \sim \chi^2_{r_i}$, *then* $Q_1 \perp\!\!\!\perp Q_2 \iff \mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$.

Proof: ($\Longrightarrow$) Suppose that $\mathbf{Q}_1 \perp\!\!\!\perp \mathbf{Q}_2$. Since $Q_1 \sim \chi^2_{r_1}$ and $Q_2 \sim \chi^2_{r_2}$, we know that $Q_1 + Q_2 \sim \chi^2_{r_1+r_2}$. We know that $Q_1 + Q_2 = \mathbf{y}'(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{y}$. From the fundamental theorem of quadratic forms, we know that $(\mathbf{A}_1 + \mathbf{A}_2) = (\mathbf{A}_1 + \mathbf{A}_2)(\mathbf{A}_1 + \mathbf{A}_2)$. Thus we have

$$(\mathbf{A}_1 + \mathbf{A}_2)(\mathbf{A}_1 + \mathbf{A}_2) = \mathbf{A}_1^2 + \mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_2\mathbf{A}_1 + \mathbf{A}_2^2 = \mathbf{A}_1 + \mathbf{A}_2$$

$$\Longrightarrow (*)\mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_2\mathbf{A}_1 = \mathbf{0}$$

$$\mathbf{A}_1\mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_1\mathbf{A}_2\mathbf{A}_1 = \mathbf{A}_1\mathbf{A}_2 + \mathbf{A}_1\mathbf{A}_2\mathbf{A}_1 = \mathbf{0}$$

$$\mathbf{A}_1\mathbf{A}_2\mathbf{A}_1 + \mathbf{A}_2\mathbf{A}_1\mathbf{A}_1 = \mathbf{A}_1\mathbf{A}_2\mathbf{A}_1 + \mathbf{A}_2\mathbf{A}_1 = \mathbf{0}$$

From this, we have

$$\mathbf{A}_1\mathbf{A}_2 = \mathbf{A}_2\mathbf{A}_1$$

From $(*)$, we have that

$$\mathbf{A}_1\mathbf{A}_2 = \mathbf{A}_2\mathbf{A}_1 = \mathbf{0}$$

($\Longleftarrow$) If $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$, $cov(\mathbf{A}_1\mathbf{y}, \mathbf{A}_2\mathbf{y}) = \mathbf{A}_1\mathbf{I}\mathbf{A}_2' = \mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$. Therefore, by lemma 22, we know that since $\mathbf{A}_1\mathbf{y} \perp\!\!\!\perp \mathbf{A}_2\mathbf{y}$, we have $Q_1 = \mathbf{y}'\mathbf{A}_1\mathbf{y} \perp\!\!\!\perp \mathbf{y}'\mathbf{A}_2\mathbf{y} = Q_2$.

□

Lemma 23 *Let $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ Suppose $Q_i = \mathbf{y}'\mathbf{A}_i\mathbf{y}$ ($Q_i \sim \chi^2_{r_i}$) for $i = 1, 2$ and $Q_1 - Q_2 \geq 0$, then $Q_1 - Q_2 \sim \chi^2_{r_1-r_2}$ and $Q_1 - Q_2 \perp\!\!\!\perp Q_2$,*

Proof: Suppose that $Q_1 - Q_2 \geq 0$, Thus we know that $0 \leq \mathbf{y}'(\mathbf{A}_1 - \mathbf{A}_2)\mathbf{y} \;\; \forall\mathbf{y}$. In particular, consider $\mathbf{y} \in \mathcal{N}(\mathbf{A}_1)$. Thus we have

$$0 \leq 0 - \mathbf{y}'\mathbf{A}_2\mathbf{y}$$

Since $\mathbf{y}'\mathbf{A}_2\mathbf{y} \geq 0$ we know that $\mathbf{y}'\mathbf{A}_2\mathbf{y}$. From the idempotency of $\mathbf{A}_2$, we have that

$$\mathbf{y}'\mathbf{A}_2\mathbf{A}_2\mathbf{y} = 0 \implies \mathbf{A}_2 \in \mathcal{N}(\mathbf{A}_2)$$

Therefore, we have $\mathcal{N}(\mathbf{A}_1) \subseteq \mathcal{N}(\mathbf{A}_2)$. For all $\mathbf{y} \in \mathbb{R}^n$, we have

$$(\mathbf{I} - \mathbf{A}_1)\mathbf{y} \in \mathcal{N}(\mathbf{A}_1) \subseteq \mathcal{N}(\mathbf{A}_2)$$

Since $\mathbf{A}_2(\mathbf{I} - \mathbf{A}_1)\mathbf{y} = \mathbf{0}\forall\mathbf{y}$, we have that $\mathbf{A}_2(\mathbf{I} - \mathbf{A}_1) = \mathbf{0}$ or $\mathbf{A}_2 = \mathbf{A}_1\mathbf{A}_2$ (we also have that $\mathbf{A}_2' = \mathbf{A}_2 = \mathbf{A}_2'\mathbf{A}_1' = \mathbf{A}_2\mathbf{A}_1$).
Thus we have

$$(\mathbf{A}_1 - \mathbf{A}_2)^2 = \mathbf{A}_1^2 - \mathbf{A}_1\mathbf{A}_2 - \mathbf{A}_2\mathbf{A}_1 + \mathbf{A}_2^2 = \mathbf{A}_1 - \mathbf{A}_2 - \mathbf{A}_2 + \mathbf{A}_2 = \mathbf{A}_1 - \mathbf{A}_2$$

Therefore, by the fundamental theorem of quadratic forms, we know that $Q_1 - Q_2 \sim \chi^2_{r_1-r_2}$, where $r_1 = tr(\mathbf{A}_1)$ and $r_2 = tr(\mathbf{A}_2)$.
We can also see that $(\mathbf{A}_1 - \mathbf{A}_2)\mathbf{A}_2 = \mathbf{A}_2 - \mathbf{A}_2 = \mathbf{0}$, so by theorem 18, we have that $Q_1 - Q_2 \perp\!\!\!\perp Q_2$.
□

2.3 Non-Central Distributions

We know from previous classes that if $\mathbf{y} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$, that $\mathbf{y}'\mathbf{y} \sim \chi^2_n$. What happens if $y_i \sim \mathcal{N}(\theta_i, 1)$ where $y_i \perp\!\!\!\perp y_j$ for $i \neq j$?

Definition 31 *Let $y_i \sim \mathcal{N}(\theta_i, 1)$ where $y_i \perp\!\!\!\perp y_j$ for $i \neq j$. Then $z = \mathbf{y}'\mathbf{y} = \sum_{i=1}^n y_i^2 \sim \chi^2_n(\delta)$, where $\delta = \sum_{i=1}^n \theta_i^2$. We call this distribution the* ***Non-Central Chi-Squared Distribution****, and δ is known as the non-centrality parameter.*

The Non-Central Chi-Squared distribution has the following properties:

1. $\mathbb{E}(z) = \sum_{i=1}^n \mathbb{E}y_i^2 = \sum_{i=1}^n \mathbb{E}(y_i - \theta_i + \theta_i)^2 = \sum_{i=1}^n \mathbb{E}(y_i - \theta_i)^2 + \theta_i^2 + \theta_i \mathbb{E}(y_i - \theta_i) = n + \sum_{i=1}^n \theta_i^2 = n + \delta$
2. $var(z) = \sum_{i=1}^n \mathbb{E}(y_i)^4 - \mathbb{E}(y_i)^2 = 2n + 4\delta$
3. If $z_i \sim \chi^2(\delta_i)$ and $z_1 \perp\!\!\!\perp z_2$, then we have $z_1 + z_2 \sim \chi^2(\delta_1 + \delta_2)$

We know from previous classes that if we have $y \sim \mathcal{N}(0,1)$ and $z \sim \chi^2_n$ where $y \perp\!\!\!\perp z$, then $\frac{y}{\sqrt{z/n}} \sim t_n$. We can extend this to a non-central distribution.

Definition 32 *If $y \sim \mathcal{N}(0,1)$ and $z \sim \chi^2_n(\delta)$ where $y \perp\!\!\!\perp z$, then $\frac{y}{\sqrt{z/n}} \sim t_n(\delta)$. We call this distribution the* ***Non-Central t Distribution****.*

We know that if $x \sim \chi^2_n$ and $y \sim \chi^2_r$ where $x \perp\!\!\!\perp y$, then $\frac{x/n}{y/r} \sim \mathcal{F}_{n,r}$. We can similarly extend this to a non-central distribution.

Definition 33 *If $x \sim \chi^2_n(\delta)$ and $y \sim \chi^2_r$ where $x \perp\!\!\!\perp y$, then $\frac{x/n}{y/r} \sim \mathcal{F}_{n,r}(\delta)$. We call this distribution the* ***Non-Central F Distribution****.*

If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then what is the distribution of $Q = \mathbf{y}'\mathbf{A}\mathbf{y}$? We can rewrite Q in the following form:

$$Q = \mathbf{y}'\boldsymbol{\Sigma}^{-1/2}\mathbf{T}\mathbf{T}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{T}\mathbf{T}'\boldsymbol{\Sigma}^{-1/2}\mathbf{y}$$

Let $\mathbf{T}$ be the orthogonal matrix obtained by the spectral decomposition of $\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}$. Thus we know that $\mathbf{T}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{T} = \mathbf{D}$ is diagonal. Let $\mathbf{z} = \mathbf{T}'\boldsymbol{\Sigma}^{-1/2}\mathbf{y}$. Thus we know that $\mathbf{z} \sim \mathcal{N}(\mathbf{T}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \mathbf{I})$. Therefore, we have

$$Q = \mathbf{z}'\mathbf{D}\mathbf{z} = \sum_{i=1}^n \lambda_i z_i^2$$

where λ_i are the eigenvalues of $\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}$. Therefore, we can see that Q is a weighted linear combination of independent $\chi^2(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1/2}t_i t_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu})$ random variables, and the weights are the eigenvalues of $\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}$.
Consider the following special cases:

1. $\boldsymbol{\Sigma} = \mathbf{I}$, $\mathbf{A}^2 = \mathbf{A}$ and has rank r.

   Thus we have that $\mathbf{D} = diag(1, 1, \ldots, 1, 0, \ldots 0)$ with r 1's (since idempotent matrices have eigenvalues of either 1 or 0). Therefore, we have

   $$Q = \mathbf{z}'\mathbf{D}\mathbf{z} = \sum_{i=1}^r z_i^2$$

   Since we know $\mathbf{z} \sim \mathcal{N}(\mathbf{T}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu} = \mathbf{T}'\boldsymbol{\mu}, \mathbf{I})$, we have that

   $$Q = \sum_{i=1}^r \chi^2_1(\boldsymbol{\mu}'\mathbf{t}_i\mathbf{t}_i'\boldsymbol{\mu})$$

$$= \chi^2_r \left(\boldsymbol{\mu}' \left(\sum_{i=1}^{r} \mathbf{t}_i \mathbf{t}_i' \right) \boldsymbol{\mu} \right)$$

However, notice that $(\sum_{i=1}^{r} \mathbf{t}_i \mathbf{t}_i') = \mathbf{T}'\mathbf{D}\mathbf{T} = \mathbf{A}$. Thus we have that

$$Q \sim \chi^2_r(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu})$$

2. $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$

   Thus we have that $\mathbf{D} = \mathbf{T}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{T} = \mathbf{I}$. We also have that $\mathbf{z} \sim \mathcal{N}(\mathbf{T}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \mathbf{I})$. Thus we have that

$$Q = \mathbf{z}'\mathbf{D}\mathbf{z} = \mathbf{z}'\mathbf{z}$$

$$Q = \sum_{i=1}^{n} \chi^2_1(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1/2}\mathbf{t}_i\mathbf{t}_i'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu})$$

$$Q = \chi^2_n \left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1/2} \left(\sum_{i=1}^{n} \mathbf{t}_i \mathbf{t}_i' \right) \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu} \right)$$

   Since $\sum_{i=1}^{n} \mathbf{t}_i \mathbf{t}_i' = \mathbf{T}'\mathbf{D}\mathbf{T} = \boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2} = \mathbf{I}$, we have that

$$Q \sim \chi^2_n(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$$

Let $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $Q = \mathbf{y}'\mathbf{A}\mathbf{y}$. What is the MGF of $\mathbf{Q}$?

$$\Psi_Q(t) = \mathbb{E}(e^{tQ}) = \int e^{t\mathbf{y}'\mathbf{A}\mathbf{y}} \frac{e^{-\frac{1}{2}\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} d\mathbf{y}$$

$$= \int \frac{e^{-\frac{1}{2}\mathbf{y}'(\boldsymbol{\Sigma}^{-1} - 2t\mathbf{A})\mathbf{y}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} d\mathbf{y}$$

$$= \int \frac{e^{-\frac{1}{2}\mathbf{y}'((\boldsymbol{\Sigma}^{-1} - 2t\mathbf{A})^{-1})^{-1}\mathbf{y}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}|(\boldsymbol{\Sigma}^{-1} - 2t\mathbf{A})^{-1}|^{1/2}|\boldsymbol{\Sigma}^{-1} - 2t\mathbf{A}|^{1/2}} d\mathbf{y}$$

$$= \frac{1}{|\boldsymbol{\Sigma}|^{1/2}|\boldsymbol{\Sigma}^{-1} - 2t\mathbf{A}|^{1/2}}$$

For t sufficiently small enough. If we have that $\boldsymbol{\Sigma} = \mathbf{I}$, then we have

$$= \frac{1}{|\mathbf{I} - 2t\mathbf{A}|^{1/2}}$$

Using the spectral decomposition of $\mathbf{A}$, we have

$$= \frac{1}{|\mathbf{T}'\mathbf{T} - 2t\mathbf{T}'\mathbf{D}\mathbf{T}|^{1/2}} = \frac{1}{|\mathbf{I} - 2t\mathbf{D}|^{1/2}} = \frac{1}{(1-2t)^{r/2}} \quad t < 1/2$$

We can see that this is the MGF of a χ^2_r distribution.

Theorem 19 (Craig's Theorem) *Let* $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ *and* $Q_i = \mathbf{y}'\mathbf{A}_i\mathbf{y} \quad i = 1, 2$. *Then* $Q_1 \perp\!\!\!\perp Q_2$ *iff* $\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2 = \mathbf{0}$.

Proof:

$$\Psi_{Q_1,Q_2}(t_1,t_2) = \mathbb{E}(e^{t_1Q_1+t_2Q_2}) = \int e^{\mathbf{y}'(t_1\mathbf{A}_1+t_2\mathbf{A}_2)\mathbf{y}}\frac{e^{-\frac{1}{2}\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}d\mathbf{y}$$

$$= \int \frac{e^{-\frac{1}{2}\mathbf{y}'(\boldsymbol{\Sigma}^{-1}-(2t_1\mathbf{A}_1+2t_2\mathbf{A}_2))\mathbf{y}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}d\mathbf{y}$$

$$= \int \frac{e^{-\frac{1}{2}\mathbf{y}'((\boldsymbol{\Sigma}^{-1}-(2t_1\mathbf{A}_1+2t_2\mathbf{A}_2))^{-1})^{-1}\mathbf{y}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}|(\boldsymbol{\Sigma}^{-1}-2t_1\mathbf{A}_1-2t_2\mathbf{A}_2)^{-1}|^{1/2}|\boldsymbol{\Sigma}^{-1}-2t_1\mathbf{A}_1-2t_2\mathbf{A}_2|^{1/2}}d\mathbf{y}$$

$$= \frac{1}{|\boldsymbol{\Sigma}|^{1/2}|\boldsymbol{\Sigma}^{-1}-2t_1\mathbf{A}_1-2t_2\mathbf{A}_2|^{1/2}}$$

$$= \frac{1}{|\mathbf{I}-2t_1\boldsymbol{\Sigma}\mathbf{A}_1-2t_2\boldsymbol{\Sigma}\mathbf{A}_2|^{1/2}}$$

$$= \frac{1}{|\mathbf{I}-2t_1\boldsymbol{\Sigma}\mathbf{A}_1-2t_2\boldsymbol{\Sigma}\mathbf{A}_2+4t_1t_2\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2-4t_1t_2\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2|^{1/2}}$$

$$= \frac{1}{|(\mathbf{I}-2t_1\boldsymbol{\Sigma}\mathbf{A}_1)(\mathbf{I}-2t_2\boldsymbol{\Sigma}\mathbf{A}_2)-4t_1t_2\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2|^{1/2}} \quad (*)$$

From above, we know that the MGF of the joint independent Q_1 and Q_2 would be

$$\frac{1}{|\mathbf{I}-2t_1\boldsymbol{\Sigma}\mathbf{A}_1|^{1/2}|\mathbf{I}-2t_2\boldsymbol{\Sigma}\mathbf{A}_2|^{1/2}} \quad (**)$$

We can see that $(*)$ and $(**)$ equal each other iff and only if $4t_1t_2\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2$ for all t_1, t_2 in a sufficiently small neighborhood around 0. We can see this can only happen if $\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2 = \mathbf{0}$. Therefore, we have that $Q_1 \perp\!\!\!\perp Q_2$ iff $\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_2 = \mathbf{0}$. □

Theorem 20 (Loyne's Theorem) *Let* $\mathbf{M}^2 = \mathbf{M} = \mathbf{M}'$ *and* $\mathbf{P} \geq 0$. *If* $\mathbf{I}-\mathbf{M}-\mathbf{P} > 0$, *then* $\mathbf{MP} = \mathbf{PM} = \mathbf{0}$.

Proof: Let $x \in \mathbb{R}^p$ and $\mathbf{y} = \mathbf{Mx}$. Then we have that

$$0 \leq \mathbf{y}'(\mathbf{I}-\mathbf{M}-\mathbf{P})\mathbf{y}$$

$$= \mathbf{x}'\mathbf{M}'(\mathbf{I}-\mathbf{M}-\mathbf{P})\mathbf{Mx}$$

$$= \mathbf{x}'(\mathbf{M}-\mathbf{M}-\mathbf{MP})\mathbf{Mx}$$

$$= -\mathbf{x}'\mathbf{MPMx}$$

$$= -\mathbf{y}'\mathbf{Py} \leq 0 \implies \mathbf{Py} = \mathbf{0}$$

Thus we have $\mathbf{PMx} = 0$ for all $x \in \mathbb{R}^p$. Therefore, we have $\mathbf{PM} = \mathbf{MP} = \mathbf{0}$. □

Lemma 24 (Graybill and Marsaglia's Lemma) *Let* $\mathbf{D}_i' = \mathbf{D}_i \quad i = 1,\ldots,k$ *and* $\mathbf{D} = \sum_{i=1}^{K}\mathbf{D}_{ii}$. *Then any of the two following statements imply the third:*

1. $\mathbf{D}_i^2 = \mathbf{D}_i$
2. $\mathbf{D}^2 = \mathbf{D}$
3. $\mathbf{D}_i\mathbf{D}_j = \mathbf{0}$

Proof: $(1 + 2 \implies 3)$
Since (2) holds, we know that $\mathbf{D}$ is idempotent. Therefore, we know that $\mathbf{I} - \mathbf{D}$ is idempotent. Thus we have

$$\mathbf{I} - \mathbf{D}_i - \mathbf{D}_j = \mathbf{I} - \mathbf{D} + \mathbf{D} - \mathbf{D}_i - \mathbf{D}_j$$

We know that $\mathbf{I} - \mathbf{D}$ is positive semi-definite, symmetric, and idempotent. We know that $\mathbf{D} - \mathbf{D}_i - \mathbf{D}_j = \sum k \neq i, j\mathbf{D}_k$. Since by 1, $\mathbf{D}_i$ is idempotent and therefore positive semi-definite, we have that $\sum k \neq i, j\mathbf{D}_k$ is positive semi-definite. Thus we have that

$$\mathbf{I} - \mathbf{D}_i - \mathbf{D}_j \geq 0$$

Thus by Lyone's Theorem, we have that $\mathbf{D}_i\mathbf{D}_j = \mathbf{0}\ \ \forall i \neq j$.
$(1 + 3 \implies 2)$

$$\mathbf{D}^2 = \left(\sum_{i=1}^{K} \mathbf{D}_i\right)^2$$

$$\sum_{i=1}^{K} \mathbf{D}_i^2 + \sum i \neq j\mathbf{D}_i\mathbf{D}_j$$

$$= \sum_{i=1}^{K} \mathbf{D}_i + \mathbf{0} = \mathbf{D}$$

$(2 + 3 \implies 1)$
Let λ be an eigenvalue of $\mathbf{D}_i$. Then there exists $\mathbf{x} \neq 0$ such that $\mathbf{D}_i\mathbf{x} = \lambda\mathbf{x}$. If $\lambda \neq 0$, then $\mathbf{x} = \frac{\mathbf{D}_i\mathbf{x}}{\lambda}$. Thus we have

$$\mathbf{D}\mathbf{x} = \frac{\mathbf{D}\mathbf{D}_i\mathbf{x}}{\lambda} = \frac{\left(\sum_{i=1}^{K} \mathbf{D}_i\right)\mathbf{D}_i\mathbf{x}}{\lambda} = \frac{\mathbf{D}_i^2\mathbf{x}}{\lambda} = \mathbf{D}_i\left(\frac{\mathbf{D}_i\mathbf{x}}{\lambda}\right) = \mathbf{D}_i\mathbf{x} = \lambda\mathbf{x}$$

Since by (2) we have that $\mathbf{D}^2 = \mathbf{D}$, we have that $\lambda = 1$. Thus the eigenvalues of $\mathbf{D}_i$ are 0 or 1. Therefore, we know that $\mathbf{D}_i$ is idempotent. □

Theorem 21 (Cochran's Theorem) *Let $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and suppose that $\mathbf{y}'\mathbf{y} = \sum_{i=1}^{K} Q_i$, where $\mathbf{Q}_i = \mathbf{y}'\mathbf{A}_i\mathbf{y}$ where $\mathbf{A}_i' = \mathbf{A}_i$ and has rank r_i for $i = 1, \ldots, K$. Then we have that the following are equivalent:*

1. $Q_i \perp\!\!\!\perp Q_j \;\; 1 \leq i \neq j \leq K$
2. $Q_i \sim \chi^2_{r_i} \;\; i = 1, \ldots, K$
3. $\sum_{i=1}^{K} r_i = p$

Proof: $(1 \implies 2)$
We know that $Q_i \perp\!\!\!\perp Q_j \;\; 1 \leq i \neq j \leq K$. Thus we have that

$$\mathbf{y}'\mathbf{A}_{‘}\mathbf{y} \perp\!\!\!\perp \mathbf{y}'(\mathbf{A}_2 + \cdots + \mathbf{A}_K)\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{A}_1)\mathbf{y}$$

By Craig's Theorem, we know this is true iff $\mathbf{A}_1(\mathbf{I} - \mathbf{A}_1) = \mathbf{0}$. Therefore, we have $\mathbf{A}_1 = \mathbf{A}_1^2$. Thus by the Fundamental Theorem of Quadratic Forms, we have that $\mathbf{y}'\mathbf{A}_1\mathbf{y} \sim \chi^2_{r_1}$. Similarly, we can show that $\mathbf{y}'\mathbf{A}_i\mathbf{y} \sim \chi^2_{r_i}$ for $i = 2, \ldots, K$. Thus we have $\mathbf{y}'\mathbf{A}_i\mathbf{y} \sim \chi^2_{r_i}$ for $i = 1, \ldots, K$.
$(2 \implies 3)$

By the setup, we know that $\sum_{i=1}^K \mathbf{A}_i = \mathbf{I}$. By the Fundamental Theorem of Quadratic Forms, we have that $\mathbf{A}_i^2 = \mathbf{A}_i$. Thus we know that $tr(\mathbf{A}_i) = rank(\mathbf{A}_i)$. Thus we have

$$\sum_{i=1}^K r_i = \sum_{i=1}^K tr(\mathbf{A}_i) = tr(\sum_{i=1}^K \mathbf{A}_i) = tr(\mathbf{I}) = p$$

$(3 \implies 1)$
We know that $\mathbf{I} = \sum_{i=1}^K \mathbf{A}_i$. Thus we can decompose it in the following way:

$$\mathbf{I} = \mathbf{A}_1 + (\mathbf{I} - \mathbf{A}_1)$$

Using the spectral decomposition of $\mathbf{A}_1$, we have

$$\mathbf{T}'\mathbf{I}\mathbf{T} = \mathbf{T}\mathbf{A}_1\mathbf{T} + \mathbf{T}'(\mathbf{I} - \mathbf{A})\mathbf{T}$$

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} + \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_p \end{bmatrix}$$

We know that λ_i's are eigenvalues of $\mathbf{A}_1$. Since $rank(\mathbf{A}_1) = r_1$, only r_1 of them are non-zero.

$$diag(1, \dots, 1) = diag(\lambda_1, \dots, \lambda_{r_1}, 0, \dots, 0) + diag(d_1, \dots, d_{r_1}, d_{r_1}, \dots, d_n)$$

This implies that $d_{r_1+1} = \dots = d_p = 1$. Since $rank(\mathbf{T}'(\mathbf{I} - \mathbf{A}_1)\mathbf{T}) = rank(\mathbf{I} - \mathbf{A}) = p - r_1$, we know that $d_i = 0$ for $i = 1, \dots, r_1$.
Since we know that $\mathbf{I} - \mathbf{A}_1 = \mathbf{A}_2 + \mathbf{A}_3 + \dots + \mathbf{A}_K$. Consider $\mathbf{T}'\mathbf{A}_j\mathbf{T}$. We know that the first r_1 diagonal elements cannot be non-zero (if they were, then $rank(\mathbf{A}_2 + \mathbf{A}_3 + \dots + \mathbf{A}_K) > p - r_1$.) Thus we know that

$$\mathbf{A}_1\mathbf{A}_j = diag(\lambda_1, \dots, \lambda_{r_1}, 0, \dots, 0) diag(0, \dots, 0, l_{r_1+1}, \dots l_p) = \mathbf{0}$$

Therefore, by Craig's Theorem, we know that $Q_1 \perp\!\!\!\perp Q_j$ for $j = 2, \dots, K$. Similarly, we can replace $\mathbf{A}_1$ with $\mathbf{A}_j$ $(j = 2, \dots, K)$ to get the result that $Q_i \perp\!\!\!\perp Q_j \;\; 1 \le i \ne j \le K$. □

3 Linear Regression

3.1 Least Squares Estimate

Consider the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is full column rank. We wish to find the parameter estimates of $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}}$) such that $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ is minimized. One way to do this would be to take the derivative and solve the normal equations to get the least squares estimate. Another way would be to use to use the Approximation Theorem (theorem 6). We know that $\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$. Thus we know that $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}_{\mathbf{X}}(\mathbf{y}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (proof of $\mathbf{P}_{\mathbf{X}}(\mathbf{y}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ can be found after the Approximation Theorem).

In order to perform inference on the model, we need to specify some properties of $\boldsymbol{\epsilon}$. Thus consider the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0} \quad var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

Thus we can establish some properties of $\hat{\boldsymbol{\beta}}$.

1. $\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
2. $var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'var(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Theorem 22 (Gauss-Markov Theorem) *Let $\hat{\boldsymbol{\theta}}$ be the least squares estimate of $\boldsymbol{\theta} = \mathbf{x}\boldsymbol{\beta}$, where $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X})$. Then among the class of linear unbiased estimates of $\mathbf{c}'\boldsymbol{\theta}$, $\mathbf{c}'\hat{\boldsymbol{\theta}}$ is the unique estimate with minimum variance.*

Proof: We know that $\mathbf{c}'\hat{\boldsymbol{\theta}} = \mathbf{c}'\mathbf{P}_{\mathbf{X}}\mathbf{y}$. Let $\mathbf{d}'\mathbf{y}$ be any other unbiased estimate of $\mathbf{c}'\boldsymbol{\theta}$. Thus we have

$$\mathbf{c}'\boldsymbol{\theta} = \mathbb{E}(\mathbf{d}'\mathbf{y}) = \mathbf{d}'\boldsymbol{\theta} \implies (\mathbf{c} - \mathbf{d})'\boldsymbol{\theta} = 0$$

Thus we have that $\mathbf{c} - \mathbf{d} \perp\!\!\!\perp \mathcal{C}(\mathbf{X})$. Therefore, we know that

$$\mathbf{P}_{\mathbf{X}}(\mathbf{c} - \mathbf{d}) = \mathbf{0} \implies \mathbf{P}_{\mathbf{X}}\mathbf{c} = \mathbf{P}_{\mathbf{X}}\mathbf{d}$$

Thus we have

$$var(\mathbf{c}'\hat{\boldsymbol{\theta}}) = var(\mathbf{c}\mathbf{P}_{\mathbf{X}}\mathbf{y}) = var((\mathbf{P}_{\mathbf{X}}\mathbf{d})'\mathbf{y}) = (\mathbf{P}_{\mathbf{X}}\mathbf{d})'\sigma^2\mathbf{I}(\mathbf{P}_{\mathbf{X}}\mathbf{d})$$
$$= \sigma^2\mathbf{d}'\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}}\mathbf{d} = \sigma^2\mathbf{d}'\mathbf{P}_{\mathbf{X}}\mathbf{d}$$

Thus consider

$$var(\mathbf{d}'\mathbf{y}) - var(\mathbf{c}'\hat{\boldsymbol{\theta}}) = \mathbf{d}'\sigma^2\mathbf{I}\mathbf{d} - \sigma^2\mathbf{d}'\mathbf{P}_{\mathbf{X}}\mathbf{d} = \sigma^2\mathbf{d}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{d}$$
$$= \sigma^2\mathbf{d}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{d} = \sigma^2\mathbf{d}_1\mathbf{d}_1 \geq 0$$

Note that $\sigma^2\mathbf{d}_1'\mathbf{d}_1 = 0 \iff \mathbf{d}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = \mathbf{0}$ or $\mathbf{d} = \mathbf{P}_{\mathbf{X}}\mathbf{d} = \mathbf{P}_{\mathbf{X}}\mathbf{c}$. Thus we have $\mathbf{d}'\mathbf{y} = \mathbf{c}'\mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{c}'\hat{\boldsymbol{\theta}}$. Therefore, $\mathbf{c}'\hat{\boldsymbol{\theta}}$ has the minimum variance, and is unique □

Lemma 25 *If $\mathbf{X}$ has full rank, then $\mathbf{a}'\hat{\boldsymbol{\beta}}$ is the BLUE (Best Linear Unbiased Estimate) of $\mathbf{a}'\boldsymbol{\beta}$ for every vector $\mathbf{a}$.*

Proof: We know that $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\theta}$. Thus we have that $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\theta}}$. Therefore we can let $\mathbf{c}' = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and use the Gauss-Markov Theorem to prove that $\mathbf{a}'\hat{\boldsymbol{\beta}}$ is the BLUE (Best Linear Unbiased Estimate) of $\mathbf{a}'\boldsymbol{\beta}$ for every vector $\mathbf{a}$. □

Earlier in this section, we figured out that $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, but how do we estimate σ^2?

Theorem 23 *If $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is an $n \times p$ of rank $r \leq p$, and $var(\mathbf{y}) = \sigma^2\mathbf{I}$, then*

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

is an unbiased estimate of σ^2.

Proof: We know that $\mathbf{y} - \hat{\boldsymbol{\theta}} = (\mathbf{I} - \mathbf{P_X})\mathbf{y}$. Thus we can rewrite S^2 as

$$S^2 = \frac{\mathbf{y}(\mathbf{I} - \mathbf{P_X})\mathbf{y}}{n-r}$$

Taking the expectation of S^2, we have

$$\mathbb{E}(S^2) = \frac{\mathbb{E}(\mathbf{y}(\mathbf{I} - \mathbf{P_X})\mathbf{y})}{n-r} = \frac{tr((\mathbf{I} - \mathbf{P_X})\sigma^2\mathbf{I}) + \mathbb{E}(\mathbf{y})'(\mathbf{I} - \mathbf{P_X})\mathbb{E}(\mathbf{y})}{n-r}$$

$$= \frac{(n-r)\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P_X})\mathbf{X}\boldsymbol{\beta}}{n-r} = \frac{(n-r)\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{n-r}$$

$$= \frac{\sigma^2(n-r)}{n-r} = \sigma^2$$

Therefore, we can see that $S^2 = \frac{(\mathbf{y}-\hat{\boldsymbol{\theta}})'(\mathbf{y}-\hat{\boldsymbol{\theta}})}{n-r} = \frac{RSS}{n-r}$ is an unbiased estimate of σ^2. □

Suppose we want to know the variance of S^2. We know that

$$\frac{(n-r)S^2}{\sigma^2} = \frac{\mathbf{y}'}{\sigma}(\mathbf{I} - \mathbf{P_X})\frac{\mathbf{y}}{\sigma} \sim \chi^2_{n-r}(\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P_X})\mathbf{X}\boldsymbol{\beta} = 0)$$

Thus we know that

$$var\left(\frac{(n-r)S^2}{\sigma^2}\right) = 2(n-r)$$

$$\implies var(S^2) = \frac{2\sigma^4}{n-r}$$

Theorem 24 *If $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, where $\mathbf{X}$ is $n \times p$ of rank p, then:*

1. $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

2. $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$

3. *$\hat{\boldsymbol{\beta}}$ is independent of S^2*

4. $RSS/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$

Proof: (1)
We know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and that $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. Thus we have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

(2)
We know that $\mathbf{z} = \frac{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}}{\sigma} \sim \mathcal{N}_p(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1})$. Thus consider $\mathbf{z}'(\mathbf{X}'\mathbf{X})\mathbf{z}$. We know that $\mathbf{z}'(\mathbf{X}'\mathbf{X})\mathbf{z} \sim \chi^2_p \iff (\mathbf{X}'\mathbf{X})\boldsymbol{\Sigma}(\mathbf{X}'\mathbf{X}) = (\mathbf{X}'\mathbf{X})$. Thus we have that

$$(\mathbf{X}'\mathbf{X})\boldsymbol{\Sigma}(\mathbf{X}'\mathbf{X}) = (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = (\mathbf{X}'\mathbf{X})$$

. Thus we know that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$
(3)
In order to prove that S^2 is independent of $\hat{\boldsymbol{\beta}}$, it is sufficient to show that $\mathbf{V} = \hat{\boldsymbol{\beta}} \perp\!\!\!\perp \mathbf{U} = (\mathbf{I} - \mathbf{P_X})\mathbf{y}$. We know that $\hat{\boldsymbol{\beta}} \perp= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Thus we can see that

$$\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} = \begin{bmatrix}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\\ \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\end{bmatrix}\mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta}\end{bmatrix}, \sigma^2\begin{bmatrix}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\\ \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\end{bmatrix}\begin{bmatrix}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\end{bmatrix}\right)$$

Thus we have that

$$\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta}\end{bmatrix}, \sigma^2\begin{bmatrix}(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\end{bmatrix}\right)$$

Thus, by the properties of normal distributions, we know that $\mathbf{V} \perp\!\!\!\perp \mathbf{U}$. Therefore, by lemma, we know that $\mathbf{V} \perp\!\!\!\perp \mathbf{U}'\mathbf{U}$ or $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp S^2$.
(4)
We know that

$$RSS = \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}$$

Notice that

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{P}_X)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} - 2\mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_X)'\mathbf{X}\boldsymbol{\beta}$$

Since $\mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{X}\boldsymbol{\beta} = 0$ and $\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_X)'\mathbf{X}\boldsymbol{\beta} = 0$, we have

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{P}_X)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} = \boldsymbol{\epsilon}'(\mathbf{I} - \mathbf{P_X})\boldsymbol{\epsilon}$$

Where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. Thus we know that $\frac{\boldsymbol{\epsilon}}{\sigma} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. Thus by the Fundamental Theorem of Quadratic Form, since $\mathbf{I} - \mathbf{P_X}$ is idempotent, we have that $\frac{RSS}{\sigma^2} = \frac{\boldsymbol{\epsilon}'}{\sigma}(\mathbf{I} - \mathbf{P_X})\frac{\boldsymbol{\epsilon}}{\sigma} \sim \chi^2_{n-p}$ □

3.2 Adding Further Explanatory Variables

3.2.1 Mutually Orthogonal Columns of Design Matrix

We will first consider the case where the columns of $\mathbf{X}$ are mutually orthogonal. Let

$$\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{p-1})$$

Thus we have that

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

$$= \begin{bmatrix}\mathbf{x}_0'\mathbf{x}_0 & \mathbf{0} & \ldots & \mathbf{0}\\ \mathbf{0} & \mathbf{x}_1'\mathbf{x}_1 & \ldots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{x}_{p-1}'\mathbf{x}_{p-1}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}_0'\mathbf{y}\\ \mathbf{x}_1'\mathbf{y}\\ \vdots\\ \mathbf{x}_{p-1}'\mathbf{y}\end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{x}_0'\mathbf{x}_0)^{-1}\mathbf{x}_0'\mathbf{y} \\ (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y} \\ \vdots \\ (\mathbf{x}_{p-1}'\mathbf{x}_{p-1})^{-1}\mathbf{x}_{p-1}'\mathbf{y} \end{bmatrix}$$

Thus we can see that $\hat{\beta}_j = \mathbf{x}_j'\mathbf{y}/(\mathbf{x}_j'\mathbf{x}_j)$. We can compute the residual sum of squares:

$$RSS = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

$$= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

$$= \mathbf{y}'\mathbf{y} - \sum_{i=0}^{p-1} \hat{\beta}_i \mathbf{x}_i'\mathbf{y}$$

We know that $\mathbf{y} = \hat{\beta}_j \mathbf{x}_j$ for any j. Thus we have

$$RSS = \mathbf{y}'\mathbf{y} - \sum_{i=0}^{p-1} \hat{\beta}_i^2 (\mathbf{x}_i'\mathbf{x}_i)$$

Thus if we remove the j^{th} predictor, we would have the following residual sum of squares:

$$RSS_{-j} = \mathbf{y}'\mathbf{y} - \sum_{i=0, i\neq j}^{p-1} \hat{\beta}_i^2 (\mathbf{x}_i'\mathbf{x}_i)$$

Thus we have

$$RSS - RSS_{-j} = \hat{\beta}_j^2 (\mathbf{x}_j'\mathbf{x}_j)$$

3.2.2 General Design Matrix

Suppose we start with the following model:

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad var(\mathbf{y}) = \sigma^2\mathbf{I}$$

where $\mathbf{X}$ is $n \times p$. Suppose we wish to add a set of new explanatory variables $\mathbf{Z} \in \mathbb{R}^{n\times t}$. Thus we have

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

$$= \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_G \\ \boldsymbol{\gamma}_G \end{bmatrix}$$

$$= \mathbf{W}\boldsymbol{\delta}$$

Thus we know that

$$\hat{\boldsymbol{\delta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} \quad var(\hat{\boldsymbol{\delta}}) = \sigma^2(\mathbf{W}'\mathbf{W})^{-1}$$

From the linear algebra sections, we know that we can represent $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{P}_\mathbf{X}\mathbf{Z} + (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Z}$. Thus we have

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}_G + (\mathbf{P}_\mathbf{X}\mathbf{Z} + (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Z})\boldsymbol{\gamma}_G$$

$$= \mathbf{X}(\boldsymbol{\beta}_G + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}_G) + (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Z}\boldsymbol{\gamma}_G$$

$$= \mathbf{X}\boldsymbol{\alpha} + (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Z}\boldsymbol{\gamma}_G$$

Notice that $\mathbf{X} \perp\!\!\!\perp (\mathbf{I} - \mathbf{P_X})\mathbf{Z}$. Thus we can use the orthogonal columns of the design matrix. Thus we have that

$$\hat{\boldsymbol{\delta}} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} \end{bmatrix}$$

$$\hat{\boldsymbol{\delta}} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ (\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} \end{bmatrix}$$

Thus we have that

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}_G + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\gamma}_G = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

and

$$\hat{\gamma}_G = (\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}$$

Thus we have that

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}$$

Let $\hat{\mathbf{y}}_G$ be the estimated outcomes from using the extended model.

$$SSE_G = (\mathbf{y} - \hat{\mathbf{y}}_G)'(\mathbf{y} - \hat{\mathbf{y}}_G) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G - \mathbf{Z}\hat{\gamma}_G)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G - \mathbf{Z}\hat{\gamma}_G)$$

Notice that

$$\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_G - \mathbf{Z}\hat{\gamma}_G = \mathbf{y} - (\mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\gamma}_G)) - \mathbf{Z}\hat{\gamma}_G$$

$$= (\mathbf{I} - \mathbf{P_X})\mathbf{y} + (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{I})\mathbf{Z}\hat{\gamma}_G$$

$$= (\mathbf{I} - \mathbf{P_X})(\mathbf{y} - \mathbf{Z}\hat{\gamma}_G)$$

Thus we have that

$$SSE_G = (\mathbf{y} - \mathbf{Z}\hat{\gamma}_G)'(\mathbf{I} - \mathbf{P_X})(\mathbf{y} - \mathbf{Z}\hat{\gamma}_G)$$

$$= \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} - 2\mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}\hat{\gamma}_G + \hat{\gamma}_G'\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}\hat{\gamma}_G$$

$$= \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} - \hat{\gamma}_G'(\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} - \mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}\hat{\gamma}_G) - \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}\hat{\gamma}_G$$

Notice that $\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}\hat{\gamma}_G = \mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}(\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} = \mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}$. Thus we have

$$SSE_G = \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} - \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}\hat{\gamma}_G$$

$$= SSE - \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}(\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}$$

Notice that $SSE_G \leq SSE$ since $\mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}(\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y} \geq 0$. Therefore, we have that SSE will go down as we add more predictors. But is there a price to adding more predictors? Consider

$$Cov(\hat{\boldsymbol{\beta}}_G) = cov(\hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\gamma})$$

$$= var(\hat{\boldsymbol{\beta}}) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}var(\hat{\gamma})\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - 2cov(\hat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\gamma})$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + 0$$

Thus we can see that the variance of $\hat{\boldsymbol{\beta}}_G$ is at least as large as $\hat{\boldsymbol{\beta}}$. Therefore, adding more predictors often increases the variance of your estimated regression coefficients.

3.3 Linear Regression with Linear Restrictions

Consider the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_H + \boldsymbol{\epsilon} \ \ s.t. \ \ \mathbf{A}\boldsymbol{\beta}_H = \mathbf{c}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ with rank p, $\mathbf{A} \in \mathbb{R}^{q\times p}$ of rank q and $\mathbf{c}$ is a known vector of length q. Suppose that $\boldsymbol{\beta}_0$ is any solution of $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$. Then we have

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}(\boldsymbol{\beta}_H - \boldsymbol{\beta}_0) + \boldsymbol{\epsilon}$$

Thus we can transform our problem into the following problem

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \ \text{ and } \ \mathbf{A}\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\beta}_H - \mathbf{A}\boldsymbol{\beta}_0 = \mathbf{0}$$

Thus we have the model where $\tilde{\mathbf{y}} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X})$. We know that

$$\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\theta} = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\gamma} = \mathbf{0}$$

Thus we can see that $\boldsymbol{\theta} \in \mathcal{N}(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$. Let $\mathbf{A}_1 = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\omega = \mathcal{N}(\mathbf{A}_1) \cap \mathcal{C}(\mathbf{X})$. Thus we can see that we wish to find $\hat{\boldsymbol{\theta}} \in \omega$. Thus we have

$$\mathbf{X}\hat{\boldsymbol{\gamma}} = \mathbf{P}_\omega \tilde{\mathbf{Y}} = \mathbf{P}_{\mathcal{C}(\mathbf{X})}\tilde{\mathbf{y}} - \mathbf{P}_{\omega^\perp\cap\mathcal{C}(\mathbf{X})}\tilde{\mathbf{y}}$$

We are just left to find what $\mathbf{P}_{\omega^\perp\cap\mathcal{C}(\mathbf{X})}\tilde{\mathbf{y}}$ is. From an identity, we know that $\omega^\perp \cap \mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{P}_{\mathcal{C}(\mathbf{X})}\mathbf{A}_1')$ where

$$\mathbf{P}_{\mathcal{C}(\mathbf{X})}\mathbf{A}_1' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}$$

Thus we have that

$$\mathbf{P}_{\omega^\perp\cap\mathcal{C}(\mathbf{X})} = (\mathbf{P}_{\mathcal{C}(\mathbf{X})}\mathbf{A}_1')(\mathbf{A}_1\mathbf{P}^2_{\mathcal{C}(\mathbf{X})}\mathbf{A}_1')^{-1}(\mathbf{A}_1\mathbf{P}_{\mathcal{C}(\mathbf{X})})$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Thus we have

$$\mathbf{X}\hat{\boldsymbol{\gamma}} = \mathbf{P}_\omega \tilde{\mathbf{Y}} = \mathbf{P}_{\mathcal{C}(\mathbf{X})}\tilde{\mathbf{Y}} - \mathbf{P}_{\omega^\perp\cap\mathcal{C}(\mathbf{X})}\tilde{\mathbf{Y}}$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta}_0)$$

Since $\mathbf{A}\boldsymbol{\beta}_0 = \mathbf{c}$, we have

$$\mathbf{X}\hat{\boldsymbol{\beta}}_H - \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

$$\implies \hat{\boldsymbol{\beta}}_H = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

Therefore, we have

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

where $\hat{\boldsymbol{\beta}}_H$ is the constrained estimate, and $\hat{\boldsymbol{\beta}}$ is the unconstrained estimate.

3.4 Design Matrix of Less Than Full Rank

Consider the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \quad var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is of rank $r < p$. Thus we can see that $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. What is the orthogonal projector onto the column space of $\mathbf{X}$? We can show that $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is the orthogonal projector onto the column space of $\mathbf{X}$. Note that there is no linear unbiased estimator of $\boldsymbol{\beta}$ when $rank(\mathbf{X}) < p$. If $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, then we desire a matrix $\mathbf{C}$ such that $\mathbb{E}(\mathbf{C}\mathbf{y}) = \boldsymbol{\beta}$. In order to have this be true, we would need

$$\mathbf{C}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

This implies that $\mathbf{C}\mathbf{X} = \mathbf{I}$, which is not possible since $\mathbf{X}$ is not full rank (but $\mathbf{I}$ is). Therefore, it is impossible for us to unbiasedly estimate every β_i . However, it is possible for us to unbiasedly estimate some linear functions of $\boldsymbol{\beta}$ using a linear function of $\mathbf{y}$.

Definition 34 *The parametric function* $\mathbf{a}'\boldsymbol{\beta}$ *is said to be* ***estimable*** *if it has a linear unbiased estimate* $\mathbf{b}'\mathbf{y}$*.*

What implications does this definition have? Let $\mathbf{a}'\mathbf{y}$ be an unbiased linear estimator of $\mathbf{c}'\boldsymbol{\beta}$. By definition, $\mathbf{c}'\boldsymbol{\beta}$ is estimable if

$$\mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbb{E}(\mathbf{a}'\mathbf{y}) = \mathbf{c}'\boldsymbol{\beta} \;\; \forall\boldsymbol{\beta}$$

$$\implies \mathbf{c} = \mathbf{X}'\mathbf{a}$$

Thus we can see that $\mathbf{c}$ has to be in the rowspace of $\mathbf{X}$.

Theorem 25 $\mathbf{c}'\boldsymbol{\beta}$ *is estimable* $\iff$ $\mathbf{c}' = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$.

Proof: $(\implies)$
If $\mathbf{c}'\boldsymbol{\beta}$ is estimable, then $\mathbf{c} = \mathbf{X}'\mathbf{a}$ for some $\mathbf{a}$. Thus we have

$$\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{a}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{a}'\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{a}'\mathbf{X} = \mathbf{c}'$$

Therefore, we have that $\mathbf{c}' = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$.
$(\impliedby)$

$$\mathbb{E}(\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}) = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta}$$

Therefore, $\mathbf{a}'\mathbf{y}$ with $\mathbf{a}' = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is a linearly unbiased estimator of $\mathbf{c}'\boldsymbol{\beta}$. □

Lets look at an example. Consider the following one-way ANOVA model:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad i = 1, \ldots, K \quad j = 1, \ldots, n_i$$

where $\mathbb{E}(\epsilon_{ij}) = 0$ and $var(\epsilon_{ij}) = \sigma^2$. Our goal is to estimate $\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_K \end{bmatrix}$.

In matrix form, our model looks like:

$$\begin{bmatrix} \mathbf{y}_{1*} \\ \mathbf{y}_{2*} \\ \vdots \\ \mathbf{y}_{K*} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \ldots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \ldots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_K} & \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \ldots & \mathbf{1}_{n_K} \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_K \end{bmatrix}$$

We can see that $\mathbf{X}'\mathbf{X}$ is

$$\begin{bmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \dots & \mathbf{1}'_{n_K} \\ \mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} & \dots & \mathbf{0}'_{n_K} \\ \mathbf{0}'_{n_1} & \mathbf{1}'_{n_2} & \dots & \mathbf{0}'_{n_K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \dots & \mathbf{1}_{n_K} \end{bmatrix} \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_K} & \mathbf{0}_{n_K} & \mathbf{0}_{n_K} & \dots & \mathbf{1}_{n_K} \end{bmatrix} = \begin{bmatrix} n & n_1 & n_2 & \dots & n_K \\ n_1 & n_1 & 0 & \dots & 0 \\ n_2 & 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & \dots & n_K \end{bmatrix}$$

where $n = \sum_{i=1}^{K} n_i$. We know that $(\mathbf{X}'\mathbf{X})^-$ takes the following form:

$$(\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_1} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{n_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_K} \end{bmatrix}$$

Thus, consider

$$\mathbf{c}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X} = \begin{bmatrix} c_0 & c_1 & c_2 & \dots & c_K \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_1} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{n_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_K} \end{bmatrix} \begin{bmatrix} n & n_1 & n_2 & \dots & n_K \\ n_1 & n_1 & 0 & \dots & 0 \\ n_2 & 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & \dots & n_K \end{bmatrix}$$

$$= \begin{bmatrix} c_0 & c_1 & c_2 & \dots & c_K \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{K} c_i & c_1 & c_2 & \dots & c_k \end{bmatrix}$$

Thus we have the following condition: $\mathbf{c}'\boldsymbol{\beta}$ is estimable if and only if

$$\begin{bmatrix} c_0 & c_1 & c_2 & \dots & c_K \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{K} c_i & c_1 & c_2 & \dots & c_k \end{bmatrix}$$

or equivalently if and only if $c_0 = \sum_{i=1}^{K} c_i$.

Are the following estimable?

1. $\tau_1 - \frac{\tau_2+\tau_3}{2}$: **yes**

$$\mathbf{c}'\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & -1/2 & -1/2 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \vdots \\ \tau_K \end{bmatrix}$$

Thus we can see that $0 = c_0 = \sum_{i=1}^{K} c_i = 1 - 1/2 - 1/2 = 0$.

2. $\mu + \tau_1$: **yes**

$$\mathbf{c}'\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_K \end{bmatrix}$$

Thus we can see that $1 = c_0 = \sum_{i=1}^{K} c_i = 1$.

3. $\mu - \tau_1$: **no**

$$\mathbf{c}'\boldsymbol{\beta} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_K \end{bmatrix}$$

Thus we can see that $1 = c_0 \neq \sum_{i=1}^{K} c_i = -1$.

3.5 Generalized Least Squares

Suppose we have the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0} \quad var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$$

where $\mathbf{V}$ is a $n \times n$ positive definite matrix.

We know that we can take the square root of $\mathbf{V} = \mathbf{V}^{1/2}\mathbf{V}^{1/2}$ since $\mathbf{V}$ is positive definite. We also know that the inverse exists, so we have $\mathbf{V}^{-1} = \mathbf{V}^{-1/2}\mathbf{V}^{-1/2}$Thus consider multiplying the right side and the left side by $\mathbf{V}^{-1/2}$. Thus we have

$$\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\boldsymbol{\epsilon}$$

$$\implies \mathbf{z} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\eta}$$

We can see that $\mathbf{E}(\boldsymbol{\eta}) = \mathbf{V}^{-1/2}\mathbf{0} = \mathbf{0}$ and that $var(\boldsymbol{\eta}) = \mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2} = \mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{V}^{1/2}\mathbf{V}^{-1/2} = \mathbf{I}$. Thus we are back to our standard linear regression. Thus we have that

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{z} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

We can derive the following properties of our estimate:

1. $\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
2. $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$
3. $\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = (\mathbf{Z} - \mathbf{B}\hat{\boldsymbol{\beta}})'(\mathbf{Z} - \mathbf{B}\hat{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'V^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$

4 Hypothesis Testing and Inference

4.1 Likelihood Ratio Test

Consider the following linear model

$$G : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

Suppose that we wish to test the hypothesis $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{A}$ is $q \times p$ with rank q. The likelihood function for G is

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

It can be shown that the MLE estimates for $\boldsymbol{\beta}$ and σ^2 are $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma}^2 = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2/n$. Thus we have that

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

$$= (2\pi\hat{\sigma}^2)^{-n/2} exp\left[-\frac{n}{2}\right]$$

The next step is to find $\hat{\boldsymbol{\beta}}_H$ and $\hat{\sigma}^2_H$ which are the MLE estimates subject to $\mathbf{A}\hat{\boldsymbol{\beta}}_H = \mathbf{c}$. We can use the Lagrange multiplier approach to solve this problem. If you use this method, we will get that

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

Notice that this is the same estimate as in section 3.3. We will also get that $\hat{\sigma}^2_H = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H||^2/n$. Thus we will get that

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_H, \hat{\sigma}^2_H) = (2\pi\hat{\sigma}^2_H)^{-n/2} exp\left[-\frac{n}{2}\right]$$

Thus we can calculate the likelihood ratio test, which is given by

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

We will reject H if Λ is too small.

4.2 F-test

As stated in section 4.1, we know the the likelihood ratio test does not account for the difference in precision of the elements of $\mathbf{A}\hat{\boldsymbol{\beta}}$. One way to do this is to define a distance measure which depends on the covariance matrix of $\mathbf{A}\hat{\boldsymbol{\beta}}$. Consider using the following quadratic form:

$$(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'(var(\mathbf{A}\hat{\boldsymbol{\beta}}))^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

where $var(\mathbf{A}\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$. Let us define

$$RSS = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 = (n-p)S^2$$

$$RSS_H = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H||^2$$

from section 3.3, we know that

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

Theorem 26

1. *$RSS_H - RSS = ||\hat{\mathbf{y}} - \hat{\mathbf{y}}_H||^2 = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$*
2. *$\mathbb{E}[RSS_H - RSS] = \sigma^2 q + (\mathbf{A}\boldsymbol{\beta} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{\beta} - \mathbf{c})$*
3. *When H is true,*

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{= (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})}{qS^2}$$

is distributed $F_{q,n-p}$.

4. *When $\mathbf{c} = \mathbf{0}$, F can be expressed in the form*

$$F = \frac{n-p}{q}\frac{\mathbf{Y}'(\mathbf{P}_\mathbf{X} - \mathbf{P}_H)\mathbf{Y}}{\mathbf{Y}'(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y}}$$

Where $\mathbf{P}_H = \mathbf{P}_\mathbf{X} - \mathbf{P}_{\mathbf{X}_1}$ where $\mathbf{P}_{\mathbf{X}_1}$ is the projection onto the variables that we are testing if they are equal to zero.

Proof:

1.

$$RSS_H - RSS = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H||^2 - ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 = ||\mathbf{y} - \hat{\mathbf{y}}_H||^2 - ||\mathbf{y} - \hat{\mathbf{y}}||^2$$

We know that

$$||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H||^2 - ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 = ||\mathbf{X}(\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}})||^2$$

Thus letting $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{y}}_H = \mathbf{X}\hat{\boldsymbol{\beta}}_H$, we have

$$||\mathbf{y} - \hat{\mathbf{y}}_H||^2 - ||\mathbf{y} - \hat{\mathbf{y}}||^2 = ||\hat{\mathbf{y}}_H - \hat{\mathbf{y}}||^2$$

Thus we have

$$RSS_H - RSS = ||\mathbf{y} - \hat{\mathbf{y}}_H||^2 - ||\mathbf{y} - \hat{\mathbf{y}}||^2$$
$$= ||\mathbf{X}(\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}})||^2 = (\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_H - \hat{\boldsymbol{\beta}})$$

Using the least squares estimate of $\boldsymbol{\beta}_H$, we have

$$(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

Thus we have

$$RSS_H - RSS = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

2. Consider $\mathbf{z} = \mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}$. We know that $\mathbf{z} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\beta} - \mathbf{c}, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')$. From chapter 2, we know that

$$\mathbb{E}(RSS_H - RSS) = tr(\sigma^2[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}') + \mathbb{E}(\mathbf{z})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbb{E}(\mathbf{z})$$
$$= tr(\sigma^2\mathbf{I}_q) + (\mathbf{A}\boldsymbol{\beta} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{\beta} - \mathbf{c})$$
$$= \sigma^2 q + (\mathbf{A}\boldsymbol{\beta} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{\beta} - \mathbf{c})$$

3. Under H, we know that $\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')$. Thus by theorem, we know that
$$\frac{RSS_H - RSS}{\sigma^2} = \frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})}{\sigma^2} \sim \chi^2_q$$
We know that
$$\frac{(RSS_H - RSS)/q}{RSS/(n-p)} \sim F_{q,n-p}$$
if $(RSS_H - RSS) \perp\!\!\!\perp RSS$. We know that $RSS_H \sim \chi^2_{n-p-q}$ and $RSS \sim \chi^2_{n-p}$. We also know that $RSS_H - RSS \geq 0$. Thus by lemma 23, we know that $(RSS_H - RSS) \perp\!\!\!\perp RSS$, so we have that
$$\frac{(RSS_H - RSS)/q}{RSS/(n-p)} \sim F_{q,n-p}$$

4. We can see that
$$\begin{aligned}\hat{\mathbf{y}}_H &= \mathbf{X}\hat{\boldsymbol{\beta}}_H \\ &= \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} \\ &= (\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y} = \mathbf{P}_H\mathbf{y}\end{aligned}$$
One can show that $\mathbf{P}_H$ is idempotent and symmetric. Thus we have that
$$RSS_H = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{P}_H)\mathbf{y}$$
Thus we also have that
$$RSS_H - RSS = \mathbf{y}'(\mathbf{I} - \mathbf{P}_H)\mathbf{y} - \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y} = \mathbf{y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_H)\mathbf{y}$$
Thus we can see that
$$F = \frac{n-p}{q}\frac{\mathbf{Y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_H)\mathbf{Y}}{\mathbf{Y}'(\mathbf{I} - \mathbf{PX})\mathbf{Y}}$$

□

Consider the following example. Let $U_1, \dots, U_{n_1}$ be sampled independently from $\mathcal{N}(\mu_1, \sigma^2)$, and let $V_1, \dots, V_{n_2}$ be sampled independently from $\mathcal{N}(\mu_2, \sigma^2)$. Suppose that we wish to test the following hypothesis:
$$H_0 : \mu_1 = \mu_2$$
We can see that we have the following model:
$$U_i = \mu_1 + \epsilon_i \ \ (i = 1, \dots, n_1)$$
$$V_i = \mu_2 + \epsilon_i \ \ (i = n_1 + 1, \dots, n_1 + n_2)$$
where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Thus consider the following matrix representation of our model:
$$\begin{bmatrix} U_1 \\ \vdots \\ U_{n_1} \\ V_1 \\ \vdots \\ V_{n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_1} \\ \epsilon_{n_1+1} \\ \vdots \\ \epsilon_{n_1+n_2} \end{bmatrix}$$
Thus our model is of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. We can also rewrite our hypothesis in the following form:
$$\mathbf{A}\boldsymbol{\beta} = \mathbf{c} = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix}$$

We can see that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{Y}) = \begin{bmatrix} \sum_{i=1}^{n_1} U_i \\ \sum_{i=1}^{n_2} V_i \end{bmatrix}$$

Thus we have that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \bar{U} \\ \bar{V} \end{bmatrix}$. From theorem 26, we know that our test statistic has the following form

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{= (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})}{qS^2}$$

We have that

$$\mathbf{A}\hat{\boldsymbol{\beta}} = \bar{U} - \bar{V}$$

$$\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' = \frac{1}{n_1} - \frac{1}{n_2} = \frac{n_2 - n_1}{n_1 n_2}$$

$$S^2 = \frac{RSS}{n_1 + n_2 - 2} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} U_i^2 + \sum_{i=1}^{n_2} V_i^2 - n_1\bar{U}^2 - n_2\bar{V}^2}{n_1 + n_2 - 2}$$

$$= \frac{\sum_{i=1}^{n_1}(U_i - \bar{U})^2 + \sum_{i=1}^{n_2}(V_i - \bar{V})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Thus we have that

$$F = \frac{(\bar{U} - \bar{V})^2}{\frac{S^2}{n_1} - \frac{S^2}{n_2}} \sim F_{1, n_1 + n_2 - 2}$$

4.3 Multiple Correlation Coefficient

We will start with trying interpret β_p in a geometric sense. Thus lets define $\mathbf{V}_{p-1} = \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_{p-1})$. Let

$$\hat{\mathbf{x}}_p = \mathbf{P}_{\mathbf{V}_{p-1}}\mathbf{x}_p$$

$$\hat{\mathbf{x}}_p^{\perp} = (\mathbf{I} - \mathbf{P}_{\mathbf{V}_{p-1}})\mathbf{x}_p$$

Thus we can see that $\mathbf{x}_p = \hat{\mathbf{x}}_p + \hat{\mathbf{x}}_p^{\perp}$. We can see that we are decomposing $\mathbf{x}_p$ into a part that is in the linear span of $\{\mathbf{x}_1, \ldots, \mathbf{x}_{p-1}\}$ and a part that is not in the span. Notice that

$$||\hat{\mathbf{x}}_p^{\perp}||^2 = \langle \hat{\mathbf{x}}_p^{\perp}, \hat{\mathbf{x}}_p^{\perp} \rangle = \langle \mathbf{x}_p - \mathbf{P}_{\mathbf{V}_{p-1}}\mathbf{x}_p, \hat{\mathbf{x}}_p^{\perp} \rangle$$

$$= \langle \mathbf{x}_p, \hat{\mathbf{x}}_p^{\perp} \rangle - \langle \mathbf{P}_{\mathbf{V}_{p-1}}\mathbf{x}_p, \hat{\mathbf{x}}_p^{\perp} \rangle$$

$$= \langle \mathbf{x}_p, \hat{\mathbf{x}}_p^{\perp} \rangle$$

Since $\mathbf{P}_{\mathbf{V}_{p-1}}\mathbf{x}_p = \hat{\mathbf{x}}_p \perp \hat{\mathbf{x}}_p^{\perp}$. Let $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Thus we can see that

$$\langle \hat{\boldsymbol{\theta}}, \hat{\mathbf{x}}_p^{\perp} \rangle = \langle \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\mathbf{x}}_p^{\perp} \rangle = \langle \hat{\beta}_p \mathbf{x}_p, \hat{\mathbf{x}}_p^{\perp} \rangle = \beta_p ||\hat{\mathbf{x}}_p^{\perp}||^2$$

$$\implies \hat{\beta}_p = \frac{\langle \hat{\boldsymbol{\theta}}, \hat{\mathbf{x}}_p^{\perp} \rangle}{||\hat{\mathbf{x}}_p^{\perp}||^2}$$

Thus we can see that $\hat{\mathbf{x}}_p^{\perp}$ measures the part of $\mathbf{x}_k$ that contributes to the linear relationship of $\mathbf{y}$ and $\mathbf{x}_p$ after accounting for the linear effects of $\mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$. Lets consider the covariance of $\hat{\beta}_j$ and $\hat{\beta}_i$. In order to do that, we have to use the fact that $\hat{\beta}_p = \frac{\langle \mathbf{y}, \hat{x}_p^{\perp} \rangle}{||\hat{x}_p^{\perp}||^2}$.

$$cov(\hat{\beta}_j, \hat{\beta}_i) = cov\left(\frac{\langle \mathbf{y}, \hat{\mathbf{x}}_i^{\perp} \rangle}{||\hat{\mathbf{x}}_i^{\perp}||^2}, \frac{\langle \mathbf{y}, \hat{\mathbf{x}}_j^{\perp} \rangle}{||\hat{\mathbf{x}}_j^{\perp}||^2} \right) = \frac{(\hat{\mathbf{x}}_i^{\perp})' var(\mathbf{y}) \hat{\mathbf{x}}_j^{\perp}}{||\hat{\mathbf{x}}_i^{\perp}||^2 ||\hat{\mathbf{x}}_j^{\perp}||^2} = \frac{(\hat{\mathbf{x}}_i^{\perp})' \sigma^2 \hat{\mathbf{x}}_j^{\perp}}{||\hat{\mathbf{x}}_i^{\perp}||^2 ||\hat{\mathbf{x}}_j^{\perp}||^2}$$

We also know that

$$cov(\hat{\beta}_j, \hat{\beta}_i) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}_{ij}$$

Thus we can see that

$$(\mathbf{X}'\mathbf{X})^{-1}_{ij} = \frac{(\hat{\mathbf{x}}_i^{\perp})'\hat{\mathbf{x}}_j^{\perp}}{||\hat{\mathbf{x}}_i^{\perp}||^2||\hat{\mathbf{x}}_j^{\perp}||^2}$$

Since we know that

$$cos(\alpha) = \frac{\langle a, b\rangle}{||a||||b||}$$

where α is the angle between a and b, we have that

$$(\mathbf{X}'\mathbf{X})^{-1}_{ij} = \frac{(cos(\alpha))}{||\hat{\mathbf{x}}_i^{\perp}||||\hat{\mathbf{x}}_j^{\perp}||}$$

where α is the angle between $\hat{\mathbf{x}}_i^{\perp}$ and $\hat{\mathbf{x}}_j^{\perp}$. Thus as the angle increases, the covariance between the two estimates decreases.

Definition 35 *Let $\mathbf{V}_k = \{\mathbf{x}_1, \ldots \mathbf{x}_k\}$. We define the* ***Multiple Correlation Coefficient*** *as*

$$r = \frac{||\mathbf{P}_{\mathbf{V}_k}\mathbf{y} - \mathbf{P}_{\mathbf{1}}\mathbf{y}||}{||\mathbf{y} - \mathbf{P}_{\mathbf{1}}\mathbf{y}||} = \frac{||\hat{\mathbf{y}}_k - \hat{\mathbf{y}}_0||}{||\mathbf{y} - \hat{\mathbf{y}}_0||}$$

Notice that $\mathbf{y} - \hat{\mathbf{y}}_k \in \mathbf{V}_k^{\perp}$ and $\hat{\mathbf{y}}_k - \hat{\mathbf{y}}_0 \in \mathbf{V}_k$. Thus we have that

$$||\mathbf{y} - \hat{\mathbf{y}}_0||^2 = ||\mathbf{y} - \hat{\mathbf{y}}_k + \hat{\mathbf{y}}_k - \hat{\mathbf{y}}_0||^2 = ||\mathbf{y} - \hat{\mathbf{y}}_k||^2 + ||\hat{\mathbf{y}}_k - \hat{\mathbf{y}}_0||^2$$

Definition 36 *We define the* ***Coefficient of Determination*** *as*

$$R^2 = r^2 = \frac{||\hat{\mathbf{y}}_k - \hat{\mathbf{y}}_0||^2}{||\mathbf{y} - \hat{\mathbf{y}}_0||^2} = 1 - \frac{||\mathbf{y} - \hat{\mathbf{y}}_k||}{||\mathbf{y} - \hat{\mathbf{y}}_0||} = 1 - \frac{SSE}{SSTO} = \frac{SSReg}{SSTO}$$

What is the contribution of $\mathbf{x}_k$ to the reduction of SSE? We know that

$$SSE_{k-1} - SSE_k = \mathbf{y}(\mathbf{I} - \mathbf{P}_{\mathbf{V}_{k-1}})\mathbf{y} - \mathbf{y}(\mathbf{I} - \mathbf{P}_{\mathbf{V}_{k-1}} - \mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}})\mathbf{y} = \mathbf{y}'\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}\mathbf{y}$$

$$\mathbf{y}'\hat{\mathbf{x}}_k^{\perp}((\hat{\mathbf{x}}_k^{\perp})'\hat{\mathbf{x}}_k^{\perp})^{-1}(\hat{\mathbf{x}}_k^{\perp})'\mathbf{y} = \frac{\langle \mathbf{y}, \hat{\mathbf{x}}_k^{\perp}\rangle^2}{||\hat{\mathbf{x}}_k^{\perp}||^2}$$

Since we know that $\langle \mathbf{y}, \hat{\mathbf{x}}_k^{\perp}\rangle = \hat{\beta}_k||\hat{\mathbf{x}}_k^{\perp}||^2$, we have

$$SSE_{k-1} - SSE_k = \hat{\beta}_k^2||\hat{\mathbf{x}}_k^{\perp}||^2$$

Consider the t-test for testing whether $\beta_k - 0$. We know the test statistic takes the following form

$$t = \frac{\hat{\beta}_k}{s\hat{.}e(\hat{\beta}_k)}$$

We will show the that the test statistic above has a relationship with the coefficient of determination. Let $d = t^2/(n-k)$. We will show that

$$R_k^2 = \frac{d}{d+1}(1 - R_{k-1}^2) + R_{k-1}^2$$

Since we know that $\hat{\beta}_k = \mathbf{y}'\hat{\mathbf{x}}_k^{\perp}((\hat{\mathbf{x}}_k^{\perp})'\hat{\mathbf{x}}_k^{\perp})^{-1}((\hat{\mathbf{x}}_k^{\perp})'\hat{\mathbf{x}}_k^{\perp})^{-1}(\hat{\mathbf{x}}_k^{\perp})'\mathbf{y}$, we have that

$$d = \frac{\mathbf{y}'\hat{\mathbf{x}}_k^{\perp}((\hat{\mathbf{x}}_k^{\perp})'\hat{\mathbf{x}}_k^{\perp})^{-1}((\hat{\mathbf{x}}_k^{\perp})'\hat{\mathbf{x}}_k^{\perp})^{-1}(\hat{\mathbf{x}}_k^{\perp})'\mathbf{y}}{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}(n-k)}$$

From above, we know that $(\mathbf{X}'\mathbf{X})^{-1}_{kk} = \frac{1}{||\hat{\mathbf{x}}_k^{\perp}||^2}$, and $\hat{\sigma}^2 = \frac{\mathbf{y}(\mathbf{I}-\mathbf{P}_{\mathbf{V}_k})\mathbf{y}}{n-k}$. Thus we have

$$d = \frac{\mathbf{y}'\hat{\mathbf{x}}_k^{\perp}((\hat{\mathbf{x}}_k^{\perp})'\hat{\mathbf{x}}_k^{\perp})^{-1}((\hat{\mathbf{x}}_k^{\perp})'\hat{\mathbf{x}}_k^{\perp})^{-1}(\hat{\mathbf{x}}_k^{\perp})'\mathbf{y}||\hat{\mathbf{x}}_k^{\perp}||^2}{\mathbf{y}(\mathbf{I}-\mathbf{P}_{\mathbf{V}_k})\mathbf{y}} = \frac{\mathbf{y}'\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}\mathbf{y}}{\mathbf{y}(\mathbf{I}-\mathbf{P}_{\mathbf{V}_k})\mathbf{y}}$$

Thus we have that

$$\frac{d}{1+d} = \frac{\mathbf{y}'\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}\mathbf{y}/\mathbf{y}(\mathbf{I}-\mathbf{P}_{\mathbf{V}_k})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{V}_k})\mathbf{y}+\mathbf{y}'\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}\mathbf{y})/\mathbf{y}(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{y}} = \frac{\mathbf{y}'\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{V}_{k-1}})\mathbf{y}}$$

We know that

$$R^2_{k-1} = \frac{\mathbf{y}'(\mathbf{P}_{\mathbf{V}_{k-1}}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}$$

Thus we have that

$$\frac{d}{d+1}(1-R^2_{k-1})+R^2_{k-1} = \frac{\mathbf{y}'\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{V}_{k-1}})\mathbf{y}}\frac{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{V}_{k-1}})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{y}} + \frac{\mathbf{y}'(\mathbf{P}_{\mathbf{V}_{k-1}}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}$$

$$= \frac{\mathbf{y}'\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{y}} + \frac{\mathbf{y}'(\mathbf{P}_{\mathbf{V}_{k-1}}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{y}} = \frac{\mathbf{y}'(\mathbf{P}_{\mathbf{V}_{k-1}}+\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}$$

Notice that $\mathbf{P}_{\mathbf{V}_{k-1}}+\mathbf{P}_{\hat{\mathbf{x}}_k^{\perp}} = \mathbf{P}_{\mathbf{V}_k}$, thus we have

$$\frac{d}{d+1}(1-R^2_{k-1})+R^2_{k-1} = \frac{\mathbf{y}'(\mathbf{P}_{\mathbf{V}_k}-\mathbf{P}_{\mathbf{1}})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{y}} = R^2_k$$

From a practical point of view, we can see that $\frac{d}{1+d} = \frac{R^2_k - R^2_{k-1}}{1-R^2_{k-1}}$. This can be interpreted as the proportion of the variance explained by the k^{th} predictor that is not explained by the $k-1$ previous predictors. We can see that the proportion only depend on d.

Definition 37 *Let* $\mathbf{v}_1, \mathbf{v}_2, \mathbf{x}_1, \dots \mathbf{x}_k \in \mathbb{R}^n$. *Let* $\mathbf{V} = \mathcal{L}(\mathbf{x}_1, \dots \mathbf{x}_k)$, $\hat{\mathbf{v}}_1 = \mathbf{P}_{\mathbf{V}}\mathbf{v}_1$, *and* $\hat{\mathbf{v}}_2 = \mathbf{P}_{\mathbf{V}}\mathbf{v}_2$. *The* ***Partial Correlation*** *of* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *with the linear effects of* $\mathbf{x}_1, \dots \mathbf{x}_k$ *removed is*

$$r_{\mathbf{v}_1,\mathbf{v}_2.\mathbf{V}} = \frac{\langle \mathbf{v}_1 - \hat{\mathbf{v}}_1, \mathbf{v}_2 - \hat{\mathbf{v}}_2 \rangle}{||\mathbf{v}_1 - \hat{\mathbf{v}}_1||||\mathbf{v}_2 - \hat{\mathbf{v}}_2||}$$

Not that we usually always remove the linear effect of $\mathbf{1}$. If $\mathbf{1} \in \mathbf{V}$ then $r_{\mathbf{v}_1,\mathbf{v}_2.\mathbf{V}}$ is scale invariant and also invariant to translation.

Consider a multiple regression model of $\mathbf{y}$ on $\mathbf{x}_1, \dots, \mathbf{x}_k$. Let us define the following

$$\mathbf{V}_{k-1} = \mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_{k-1})$$

$$\mathbf{V}_k = \mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_k)$$

$$\hat{\mathbf{y}}_{k-1} = \mathbf{P}_{\mathbf{V}_{k-1}}\mathbf{y}$$

$$\hat{\mathbf{y}}_k = \mathbf{P}_{\mathbf{V}_k}\mathbf{y}$$

$$\mathbf{x}_k^{\perp} = \mathbf{x}_k - \mathbf{P}_{\mathbf{V}_{k-1}}\mathbf{x}_k$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}_k$$

We can see that $\hat{\mathbf{y}}_k = \hat{\mathbf{y}}_{k-1} + \hat{\beta}_k\mathbf{x}_k^{\perp}$. We can decompose $\mathbf{y}$ and $\mathbf{x}_k$ into orthogonal components as follows:

$$\mathbf{y} = \hat{\mathbf{y}}_{k-1} + \hat{\beta}_k\mathbf{x}_k^{\perp} + (\mathbf{y} - \hat{\mathbf{y}}_k)$$

$$\mathbf{x}_k = \mathbf{P}_{\mathbf{V}_{k-1}}\mathbf{x} + \mathbf{x}_k^{\perp}$$

We can use this to find the correlation coefficient of $\mathbf{y}$ and $\mathbf{x}_k$ when accounting for the linear effects of $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Thus we have

$$r_{\mathbf{y},\mathbf{x}_k.\mathbf{x}_1,\ldots,\mathbf{x}_{k-1}} = \frac{\langle \mathbf{y} - \hat{\mathbf{y}}_{k-1}, \mathbf{x}_k^{\perp} \rangle}{||\mathbf{y} - \hat{\mathbf{y}}_{k-1}||||\mathbf{x}_k^{\perp}||} = \frac{\langle \hat{\beta}_k \mathbf{x}_k^{\perp} + \mathbf{e}, \mathbf{x}_k^{\perp} \rangle}{||\hat{\beta}_k \mathbf{x}_k^{\perp} + \mathbf{e}||||\mathbf{x}_k^{\perp}||}$$

Since $\mathbf{e} \perp \mathbf{x}_k^{\perp}$, we have

$$r_{\mathbf{y},\mathbf{x}_k.\mathbf{x}_1,\ldots,\mathbf{x}_{k-1}} = \frac{\langle \hat{\beta}_k \mathbf{x}_k^{\perp}, \mathbf{x}_k^{\perp} \rangle}{||\hat{\beta}_k \mathbf{x}_k^{\perp} + \mathbf{e}||||\mathbf{x}_k^{\perp}||} = \frac{\hat{\beta}_k ||\mathbf{x}_k^{\perp}||^2}{||\hat{\beta}_k \mathbf{x}_k^{\perp} + \mathbf{e}||||\mathbf{x}_k^{\perp}||}$$

Since $\mathbf{e} \perp \mathbf{x}_k^{\perp}$, we know that $||\mathbf{e} + \mathbf{x}_k^{\perp}|| = ||\mathbf{e}|| + ||\mathbf{x}_k^{\perp}||$. Thus we have

$$r_{\mathbf{y},\mathbf{x}_k.\mathbf{x}_1,\ldots,\mathbf{x}_{k-1}} = \frac{\hat{\beta}_k ||\mathbf{x}_k^{\perp}||}{||\hat{\beta}_k \mathbf{x}_k^{\perp}|| + ||\mathbf{e}||} = \frac{\hat{\beta}_k ||\mathbf{x}_k^{\perp}||}{\sqrt{||\hat{\beta}_k \mathbf{x}_k^{\perp}||^2 + ||\mathbf{e}||^2}} = \frac{\hat{\beta}_k ||\mathbf{x}_k^{\perp}||}{||\mathbf{e}||\sqrt{\frac{||\hat{\beta}_k \mathbf{x}_k^{\perp}||^2}{||\mathbf{e}||^2} + 1}}$$

Thus we can see that the partial correlation of $\mathbf{y}$ and $\mathbf{x}_k$ with the linear effects of $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ removed is $\frac{t/\sqrt{n-k}}{\sqrt{1+t^2/(n-k)}}$.

Lets look at the relationship between $r_{\mathbf{x}_1,\mathbf{x}_2.\mathbf{V}_1}$ and $r_{\mathbf{x}_1,\mathbf{x}_2.\mathbf{V}_2}$, where $\mathbf{V}_1 = \mathcal{L}(\mathbf{x}_3, \ldots, \mathbf{x}_k)$ and $\mathbf{V}_2 = \mathcal{L}(\mathbf{x}_4, \ldots, \mathbf{x}_k)$. Thus we have

$$r_{\mathbf{x}_1,\mathbf{x}_2.\mathbf{V}_1} = \frac{\langle \mathbf{x}_1 - (\mathbf{P}_{\mathbf{V}_2}\mathbf{x}_1 + \mathbf{P}_{\mathbf{x}_3^{\perp}}\mathbf{x}_1), \mathbf{x}_2 - (\mathbf{P}_{\mathbf{V}_2}\mathbf{x}_2 + \mathbf{P}_{\mathbf{x}_3^{\perp}}\mathbf{x}_2) \rangle}{||\mathbf{x}_1 - (\mathbf{P}_{\mathbf{V}_2}\mathbf{x}_1 + \mathbf{P}_{\mathbf{x}_3^{\perp}}\mathbf{x}_1)||||\mathbf{x}_2 - (\mathbf{P}_{\mathbf{V}_2}\mathbf{x}_2 + \mathbf{P}_{\mathbf{x}_3^{\perp}}\mathbf{x}_2)||}$$

Letting $\mathbf{x}_i^{\perp} = \mathbf{x}_i - \mathbf{P}_{\mathbf{V}_2}\mathbf{x}_i$, we have

$$\mathbf{x}_i - (\mathbf{P}_{\mathbf{V}_2}\mathbf{x}_i + \mathbf{P}_{\mathbf{x}_3^{\perp}}\mathbf{x}_i) = \mathbf{x}_i^{\perp} - \frac{\langle \mathbf{x}_3^{\perp}, \mathbf{x}_i \rangle \mathbf{x}_3^{\perp}}{||\mathbf{x}_3^{\perp}||}$$

We can also see that $\langle \mathbf{x}_3^{\perp}, \mathbf{x}_i \rangle = \langle \mathbf{x}_3^{\perp}, \mathbf{x}_i^{\perp} + \mathbf{P}_{\mathbf{V}_2}\mathbf{x}_i \rangle = \langle \mathbf{x}_3^{\perp}, \mathbf{x}_i^{\perp} \rangle$. Thus we have

$$\mathbf{x}_i - (\mathbf{P}_{\mathbf{V}_2}\mathbf{x}_i + \mathbf{P}_{\mathbf{x}_3^{\perp}}\mathbf{x}_i) = \mathbf{x}_i^{\perp} - \frac{\langle \mathbf{x}_3^{\perp}, \mathbf{x}_i^{\perp} \rangle \mathbf{x}_3^{\perp}}{||\mathbf{x}_3^{\perp}||}$$

WLOG, we can assume that $||\mathbf{x}_i^{\perp}|| = 1$ for $i = 1, 2, 3$. Thus we have

$$r_{\mathbf{x}_1,\mathbf{x}_2.\mathbf{V}_1} = \frac{\langle \mathbf{x}_1^{\perp} - \langle \mathbf{x}_3^{\perp}, \mathbf{x}_1^{\perp} \rangle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} - \langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle \mathbf{x}_3^{\perp} \rangle}{||\mathbf{x}_1^{\perp} \langle \mathbf{x}_3^{\perp}, \mathbf{x}_1^{\perp} \rangle \mathbf{x}_3^{\perp}||||\mathbf{x}_2^{\perp} \langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle \mathbf{x}_3^{\perp}||}$$

$$= \frac{\langle \mathbf{x}_1^{\perp}, \mathbf{x}_2^{\perp} \rangle - 2\langle \mathbf{x}_1^{\perp}, \mathbf{x}_3^{\perp} \rangle \langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle + \langle \mathbf{x}_1^{\perp}, \mathbf{x}_3^{\perp} \rangle \langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle}{\sqrt{\langle \mathbf{x}_1^{\perp}, \mathbf{x}_1^{\perp} \rangle - 2\langle \mathbf{x}_3^{\perp}, \mathbf{x}_1^{\perp} \rangle^2 + \langle \mathbf{x}_3^{\perp}, \mathbf{x}_1^{\perp} \rangle^2}\sqrt{\langle \mathbf{x}_2^{\perp}, \mathbf{x}_2^{\perp} \rangle - 2\langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle^2 + \langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle^2}}$$

$$= \frac{\langle \mathbf{x}_1^{\perp}, \mathbf{x}_2^{\perp} \rangle - \langle \mathbf{x}_1^{\perp}, \mathbf{x}_3^{\perp} \rangle \langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle}{\sqrt{1 - \langle \mathbf{x}_3^{\perp}, \mathbf{x}_1^{\perp} \rangle^2}\sqrt{1 - \langle \mathbf{x}_3^{\perp}, \mathbf{x}_2^{\perp} \rangle^2}}$$

Since we know that $r_{\mathbf{x}_1,\mathbf{x}_2.\mathbf{V}_2} = \langle \mathbf{x}_1^{\perp}, \mathbf{x}_2^{\perp} \rangle$, $r_{\mathbf{x}_1,\mathbf{x}_3.\mathbf{V}_2} = \langle \mathbf{x}_1^{\perp}, \mathbf{x}_3^{\perp} \rangle$, and $r_{\mathbf{x}_2,\mathbf{x}_3.\mathbf{V}_2} = \langle \mathbf{x}_2^{\perp}, \mathbf{x}_3^{\perp} \rangle$, we have

$$r_{\mathbf{x}_1,\mathbf{x}_2.\mathbf{V}_1} = \frac{r_{\mathbf{x}_1,\mathbf{x}_2.\mathbf{V}_2} - r_{\mathbf{x}_1,\mathbf{x}_3.\mathbf{V}_2} r_{\mathbf{x}_2,\mathbf{x}_3.\mathbf{V}_2}}{\sqrt{1 - r^2_{\mathbf{x}_1,\mathbf{x}_3.\mathbf{V}_2}}\sqrt{1 - r^2_{\mathbf{x}_2,\mathbf{x}_3.\mathbf{V}_2}}}$$

4.4 Simultaneous Confidence Intervals and Regions

Consider the following model

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

We know from previous classes that a confidence interval for β_j would take take the following form

$$\hat{\beta}_j \pm t_{n-p}^{\alpha/2}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}$$

Suppose we do this for each $\hat{beta}_i$. Thus even though the probability that any one confidence interval contains the true β_j is $1-\alpha$, the probability that all the confidence intervals contain the true parameters is not $1-\alpha$. Let E_i be the event that the i^{th} $100(1-\alpha_i)\%$ confidence interval contains the true value of β. Thus we know that $P(E_i) = 1-\alpha_i$. Thus if we want the probability that all confidence intervals contain the true parameter value, we have:

$$1-\delta = p(\cap_{i=1}^m E_i) = 1 - P(\cup_{i=1}^m E_i^c)$$
$$\geq 1 - \sum_{i=1}^m P(E_i^c) = 1 - \sum_{i=1}^m \alpha_i$$

Definition 38 *We call the value δ the **Familywise Error Rate**.*

4.4.1 Bonferroni

Consider the case when $\alpha_1 = \cdots = \alpha_m = \alpha$. Thus we have that

$$1-\delta \geq 1 - m\alpha$$

Thus if we want $\delta = \tilde{\delta}$, then we can pick $\alpha = \frac{\delta}{m}$. Therefore, we can see that we will achieve this Familywise Error rate at the minimum. This method is known as the **Bonferroni Adjustment**. While this controls the FWE, this can become overly conservative when m is large.

4.4.2 Scheffe's Method

Consider the standard linear model

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

Suppose we want to test $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, where rank of $\mathbf{A}$ is q. Suppose we want to construct $100(1-\alpha)$ confidence intervals for making inference on $\mathbf{a}_1\boldsymbol{\beta}, \ldots, \mathbf{a}_q\boldsymbol{\beta}$, where $\mathbf{A}' = (\mathbf{a}_1, \ldots, \mathbf{a}_q)$. Thus we know that

$$\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c} = \mathcal{N}_q(\mathbf{A}\boldsymbol{\beta} - \mathbf{c}, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')$$

Thus we know that

$$(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c} - (\mathbf{A}\boldsymbol{\beta} - \mathbf{c}))'(\sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c} - (\mathbf{A}\boldsymbol{\beta} - \mathbf{c})) \sim \chi_q^2$$

Simplifying, we have

$$\frac{(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))}{\sigma^2} \sim \chi_q^2$$

If σ^2 is known, then $\{\boldsymbol{\beta} | (\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \leq \chi_{q,\alpha}^2\sigma^2\}$ is a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\beta}$. If σ^2 is not known, then we have that $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2$. Thus we have

$$F = \frac{(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))/q\sigma^2}{(n-p)\hat{\sigma}^2/(\sigma^2(n-p))} \sim F_{q,n-p}$$

Simplifying, we get

$$F = \frac{(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))}{q\hat{\sigma}^2} \sim F_{q,n-p}$$

Thus we have that $\boldsymbol{\beta}|\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))q\hat{\sigma}^2 \leq F_{q,n-p,\alpha}$ is a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\beta}$.

Suppose we wish to make inference on $h'\boldsymbol{\phi}$ where $\boldsymbol{\phi} = \mathbf{A}\boldsymbol{\beta}$. Likewise, we have

$$\frac{(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})}{qS^2} \sim F_{q,n-p,\alpha}$$

Let $L = \mathbf{A}(\mathbf{X}'\mathbf{X})\mathbf{A}'$. Thus we have

$$1 - \alpha = P\left((\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \leq qS^2F_{q,n-p,\alpha}\right)$$

$$= P\left(\mathbf{b}'\mathbf{L}^{-1}\mathbf{b} \leq qS^2F_{q,n-p,\alpha}\right)$$

$$= P\left(\max_{\mathbf{h}} \frac{(\mathbf{h}'\mathbf{b})^2}{\mathbf{h}'\mathbf{L}\mathbf{h}} \leq qS^2F_{q,n-p,\alpha} \;\; \mathbf{h} \neq 0\right)$$

$$= P\left(\frac{(\mathbf{h}'\mathbf{b})^2}{\mathbf{h}'\mathbf{L}\mathbf{h}} \leq qS^2F_{q,n-p,\alpha} \;\; \forall\, \mathbf{h} \neq 0\right)$$

$$= P\left((\mathbf{h}'(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}))^2 \leq \mathbf{h}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')\mathbf{h}qS^2F_{q,n-p,\alpha} \;\; \forall\, \mathbf{h} \neq 0\right)$$

$$= P\left(\mathbf{h}'\boldsymbol{\phi} \in \mathbf{h}'\hat{\boldsymbol{\phi}} \pm \sqrt{\mathbf{h}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')\mathbf{h}qS^2F_{q,n-p,\alpha}} \;\; \forall\, \mathbf{h} \neq 0\right)$$

Thu we arrive at our confidence intervals derived by using Scheffe's Method

$$= P\left(\mathbf{h}'\mathbf{A}\boldsymbol{\beta} \in \mathbf{h}'\mathbf{A}\hat{\boldsymbol{\beta}} \pm \sqrt{\mathbf{h}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')\mathbf{h}qS^2F_{q,n-p,\alpha}} \;\; \forall\, \mathbf{h} \neq 0\right)$$

4.4.3 Studentized Range Distribution and Tukey's Method

Definition 39 *Let $z_1, \ldots, z_k$ and u be independent random variables with $z_i \sim \mathcal{N}(0,1)$ and $u \sim \chi^2_m(0)$. Then*

$$q = \max_{1 \leq i \neq j \leq k} \frac{|z_i - z_j|}{\sqrt{u/m}}$$

*is said to have a **Studentized Range Distribution** with k and m degrees of freedom (denoted $q_{k,m}$).*

Lemma 26 *Let $y_1, \ldots, y_k$ and S^2 be independent random variables with $y_i \sim \mathcal{N}(\mu, a\sigma^2)$ and $\frac{mS^2}{\sigma^2} \sim \chi^2_m(0)$. Then we have that*

$$\max_{1 \leq i \neq j \leq k} \frac{|y_i - y_j|}{\sqrt{a}S} \sim q_{k,m}$$

Proof: We know that $\frac{y_i-\mu}{\sqrt{a}\sigma} \sim \mathcal{N}(0,1)$. We also know that $\frac{mS^2}{\sigma^2} \sim \chi^2_m(0)$. Thus we know that

$$\max_{1\leq i\neq j\leq k} \frac{|\frac{y_i-\mu}{\sqrt{a}\sigma} - \frac{y_j-\mu}{\sqrt{a}\sigma}|}{\sqrt{\frac{mS^2}{\sigma^2}/m}} \sim q_{k,m}$$

by definition. Since $a > 0$, we have

$$= \max_{1\leq i\neq j\leq k} \frac{|y_i - y_j|}{\sqrt{a\frac{\sigma^2 S^2}{\sigma^2}}} = \max_{1\leq i\neq j\leq k} \frac{|y_i - y_j|}{\sqrt{a}S} \sim q_{k,m}$$

□

Suppose we have a one-way ANOVA model. Thus we have:

$$y_{ij} \sim \mathcal{N}(\mu + \alpha_i, \sigma^2); \quad \sum \alpha_i = 0, \quad i = 1, \ldots, k \;\; j = 1, \ldots, n$$

$$\bar{Y_{i.}} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}, \quad \hat{\sigma}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y_{i.}})^2$$

Notice that $\frac{\bar{y_{i.}}-\alpha_i}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$ and $\frac{(k(n-1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{k(n-1)}$. Thus by the lemma, we have

$$\max_{1\leq i\neq j\leq k} \frac{|\bar{y_{i.}} - \bar{y_{j.}} - (\alpha_i - \alpha_j)|}{\sqrt{\frac{1}{n}\hat{\sigma}^2}} = \max_{1\leq i\neq j\leq k} \frac{\sqrt{n}|\bar{y_{i.}} - \bar{y_{j.}} - (\alpha_i - \alpha_j)|}{\hat{\sigma}} \sim q_{k,k(n-1)}$$

Thus we have that

$$1 - \alpha = P\left(\max_{1\leq i\neq j\leq k} \sqrt{n}|\bar{y_{i.}} - \bar{y_{j.}} - (\alpha_i - \alpha_j)| \leq \hat{\sigma} q_{k,k(n-1),\alpha}\right)$$

$$= P\left(|\bar{y_{i.}} - \bar{y_{j.}} - (\alpha_i - \alpha_j)| \leq \frac{\hat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \quad \forall\, i \neq j\right)$$

$$= P\left(\alpha_i - \alpha_j \in \bar{y_{i.}} - \bar{y_{j.}} \pm \frac{\hat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \quad \forall\, i \neq j\right)$$

Thus a $100(1-\alpha)\%$ simultaneous confidence interval for all pairwise comparisons is

$$\bar{y_{i.}} - \bar{y_{j.}} \pm \frac{\hat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \quad i \neq j$$

Lemma 27 *Let $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$. Then we have*

$$|\alpha_i - \alpha_j| \leq b \quad \forall\, i,j \iff |\sum_{i=1}^{k} c_i a_i| \leq b \sum_{i=1}^{k} \frac{|c_i|}{2}$$

for all c_i's such that $\sum_{i=1}^{k} c_i = 0$.

Thus we can use this lemma to get a confidence interval for all contrasts in the means.

$$1 - \alpha = P\left(|\bar{y_{i.}} - \bar{y_{j.}} - (\alpha_i - \alpha_j)| \leq \frac{\hat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \quad \forall\, i \neq j\right)$$

$$\iff P\left(|\sum_{i=1}^{k} c_i(\bar{y_{i.}} - \alpha_i)| \leq \frac{\hat{\sigma}}{\sqrt{n}} q_{k,k(n-1),\alpha} \sum_{i=1}^{k} \frac{|c_i|}{2} \quad \forall c_i \; s.t. \sum c_i = 0\right)$$

$$= P\left(\sum_{i=1}^{k} c_i\alpha_i \in \sum_{i=1}^{k} c_i\bar{y_{i.}} \pm \frac{\hat{\sigma}}{\sqrt{n}}q_{k,k(n-1),\alpha}\sum_{i=1}^{k}\frac{|c_i|}{2} \quad \forall c_i \ s.t. \ \sum c_i = 0\right)$$

Thus we have that a $100(1-\alpha)\%$ simultaneous confidence interval for all $\sum_{i=1}^{k} c_i\alpha$ such that $\sum_{i=1}^{k} c_i = 0$ is

$$\sum_{i=1}^{k} c_i\bar{y_{i.}} \pm \frac{\hat{\sigma}}{\sqrt{n}}q_{k,k(n-1),\alpha}\sum_{i=1}^{k}\frac{|c_i|}{2}$$

4.5 Fieller's Theorem

Theorem 27 (Fieller's Theorem) *Let U and V be two normally distributed random variables, with means μ_U and μ_V, and variances $\nu_{11}\sigma^2$ and $\nu_{22}\sigma^2$ and covariance $\nu_{12}\sigma^2$. Then a $100(1-\alpha)\%$ confidence interval for $\frac{\mu_U}{\mu_V}$ is*

$$\frac{1}{(1-g)}\left[\frac{U}{V} - g\frac{\nu_{12}}{\nu_{22}} \pm \frac{\hat{\sigma}t_{m,\alpha/2}}{V}\sqrt{-2\frac{U}{V}\nu_{12} + \frac{g\nu_{12}^2}{\nu_{22}} + \frac{\nu_{22}U^2}{V^2} + (1-g)\nu_{11}}\right]$$

where $g = \frac{t_{m,\alpha}^2\hat{\sigma}^2\nu_{22}^2}{V^2}$. Note that m is the degree of freedom of $\hat{\sigma}^2$.

Proof: Let $W = U - \theta V = U - \frac{\mu_U}{\mu_V} \sim \mathcal{N}(\mathbf{0}, \sigma^2\nu_{11} + \sigma^2\theta^2\nu_{22} - 2\theta\sigma^2\nu_{12})$. Thus we have that

$$\frac{W}{\sqrt{\hat{\sigma}^2(\nu_{11} + \theta^2\nu_{22} - 2\theta\nu_{12})}} \sim t_m$$

Thus we have that

$$1 - \alpha = P\left(\frac{W^2}{\hat{\sigma}^2(\nu_{11} + \theta^2\nu_{22} - 2\theta\nu_{12})} \geq t_{m,\alpha/2}^2\right)$$

$$= P\left((U - \theta V)^2 - t_{m,\alpha/2}^2\hat{\sigma}^2(\nu_{11} + \theta^2\nu_{22} - 2\theta\nu_{12}) \geq 0\right)$$

Let $\psi(\theta) = (U - \theta V)^2 - t_{m,\alpha/2}^2\hat{\sigma}^2(\nu_{11} + \theta^2\nu_{22} - 2\theta\nu_{12})$. Thus we have

$$\psi(\theta) = (U^2 - t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{11}) - \theta(2UV - 2t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{12}) + \theta^2(V^2 - t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{22})$$

If $\psi(\theta) = 0$, then we have

$$\theta = \frac{2UV - 2t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{12} \pm \sqrt{(2UV - 2t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{12})^2 - 4(V^2 - t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{22})(U^2 - t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{11})}}{2(V^2 - t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{22})}$$

$$= \frac{UV - t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{12} \pm \sqrt{-8UVt_{m,\alpha/2}^2\hat{\sigma}^2\nu_{12} + 4(t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{12}) + 4t_{m,\alpha/2}^2\hat{\sigma}^2(U^2\nu_{22} + V^2\nu_{11}) - 4(t_{m,\alpha/2}^2\hat{\sigma}^2)^2\nu_{11}\nu_{22}}}{(V^2 - t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{22})}$$

Letting $g = \frac{t_{m,\alpha}^2\hat{\sigma}^2\nu_{22}^2}{V^2}$, we have

$$= \frac{UV - gV^2\frac{\nu_{12}}{\nu_{22}} \pm \hat{\sigma}t_{m,\alpha/2}V\sqrt{-2\frac{U}{V}\nu_{12} + \frac{t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{12}^2}{V^2} + \frac{U^2}{V^2}\nu_{22} + \nu_{11} - \frac{t_{m,\alpha/2}^2\hat{\sigma}^2\nu_{11}\nu_{22}}{V^2}}}{(1-g)V^2}$$

$$= \frac{U - gV\frac{\nu_{12}}{\nu_{22}} \pm \hat{\sigma}t_{m,\alpha/2}\sqrt{-2\frac{U}{V}\nu_{12} + \frac{g\nu_{12}^2}{\nu_{22}} + \frac{U^2}{V^2}\nu_{22} + (1-g)\nu_{11}}}{(1-g)}$$

$$= \frac{1}{(1-g)}\left[\frac{U}{V} - g\frac{\nu_{12}}{\nu_{22}} \pm \frac{\hat{\sigma}t_{m,\alpha/2}}{V}\sqrt{-2\frac{U}{V}\nu_{12} + \frac{g\nu_{12}^2}{\nu_{22}} + \frac{U^2}{V^2}\nu_{22} + (1-g)\nu_{11}}\right]$$

Thus a $100(1-\alpha)\%$ confidence interval for $\frac{\mu_U}{\mu_V}$ is

$$\frac{1}{(1-g)}\left[\frac{U}{V} - g\frac{\nu_{12}}{\nu_{22}} \pm \frac{\hat{\sigma}t_{m,\alpha/2}}{V}\sqrt{-2\frac{U}{V}\nu_{12} + \frac{g\nu_{12}^2}{\nu_{22}} + \frac{U^2}{V^2}\nu_{22} + (1-g)\nu_{11}}\right]$$

□

We will now show an application of Fieller's Theorem. Consider the model

$$\mathbb{E}(y) = \beta_0 + \beta_1(x_i - \bar{X})$$

Thus we know that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n (x_i - \bar{x})y_i \end{bmatrix}$$

Thus we have

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\sum_{i=1}^n (x_i-\bar{x})y_i}{\sum_{i=1}^n (x_i-\bar{x})^2} \end{bmatrix}$$

Suppose we observe y, but not x that gave the value of y. Suppose we want to construct a $100(1-\alpha)\%$ confidence interval for x (this problem is called the inverse calibration problem). Thus we have the following

$$y = \bar{y} + \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}(x - \bar{x})$$

$$\implies x = \bar{x} + \frac{y - \bar{y}}{\frac{\sum_{i=1}^n (x_i-\bar{x})y_i}{\sum_{i=1}^n (x_i-\bar{x})^2}}$$

Thus we can get a confidence interval for $x - \bar{x}$. Let $U = y - \bar{y}$ and $V = \frac{\sum_{i=1}^n (x_i-\bar{x})y_i}{\sum_{i=1}^n (x_i-\bar{x})^2}$. We have $var(U) = \sigma^2 + \frac{\sigma^2}{n}$ and $var(V) = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}$. Thus we have that $\nu_{12} = 0$, $\nu_{11} = 1 + 1/n$, and $\nu_{22} = \frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2}$. Thus we can directly apply Fieller's Theorem and get a confidence interval for x.

4.6 Case Deletion Diagnostics

Consider the following model

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}; \quad rank(\mathbf{X}) = P$$

Thus we know that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Let

$$h_{ii} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')_{ii} = \mathbf{e}_i\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}\mathbf{e}_i)' = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

Definition 40 *We call the elements h_{ii} the* ***leverage****.*

Notice that $\hat{y}_i = (H\mathbf{y})_i = \sum_{j=1}^n h_{ij}y_j = \sum_{j\neq i} h_{ij}y_j + h_{ii}y_i$. Thus we have that

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii} \qquad \frac{\partial \hat{y}_i}{\partial y_j} = h_{ij}$$

Thus we can see that the leverage represents how changes in the i^{th} outcome will effect the i^{th} prediction. Consider the i^{th} residual. we know that $e_i = y_i - \hat{y}_i$. Thus we have

$$cov(\mathbf{e}) = cov((\mathbf{I}-\mathbf{H})\mathbf{y}) = (\mathbf{I}-\mathbf{H})\sigma^2(\mathbf{I}-\mathbf{H}) = \sigma^2(\mathbf{I}-\mathbf{H})$$

Thus we have

$$var(e_i) = \sigma^2(1-h_{ii})$$

Definition 41 *We call the following quantity the* ***Normalized Residuals****:*

$$\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}} \sim \mathcal{N}(0,1)$$

Definition 42 *Let* $\hat{\sigma}^2 = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p}$. *We call the following quantity the* ***Internally Studentized Residuals****:*

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$$

Definition 43 *Let* $\hat{\sigma}^2_{(i)}$ *be the sample variance calculated without the* i^{th} *observation. We call the following quantity the* ***Externally Studentized Residuals****:*

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}}$$

where $\hat{\sigma}^2_{(i)} = \sum_{j\neq i}(y_j - \mathbf{x}_j'\hat{\boldsymbol{\beta}}_{(i)})$.

Lets take a look into the difference between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$. We will first note that

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^n \mathbf{x}_j\mathbf{x}_j' = \mathbf{x}_i\mathbf{x}_i' + \sum_{j\neq i}\mathbf{x}_j\mathbf{x}_j' = \mathbf{x}_i\mathbf{x}_i' + \mathbf{X}_{(i)}'\mathbf{X}_{(i)}$$

$$\mathbf{X}'\mathbf{y} = \sum_{i=1}^n \mathbf{x}_i y_i = \mathbf{x}_i y_i + \sum_{j\neq i}\mathbf{x}_j y_j = \mathbf{x}_i y_i + \mathbf{X}_{(i)}\mathbf{y}_{(i)}$$

Thus we have that

$$\hat{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'\right)^{-1}(\mathbf{X}\mathbf{y} - \mathbf{x}_i y_i)$$

Using Sherman-Morrison, we have that $\left(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'\right)^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}$. Thus we have

$$\hat{\boldsymbol{\beta}}_{(i)} = \left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right)(\mathbf{X}\mathbf{y} - \mathbf{x}_i y_i)$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}}{1-h_{ii}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i}{1-h_{ii}}$$

$$= \hat{\boldsymbol{\beta}} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}\left[\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y - (1-h_{ii})y_i - h_{ii}y_i\right]$$

Thus we have that

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$$

Lets take a look at how $\hat{\sigma}^2$ relates to $\hat{\sigma}^2_{(i)}$. We know that

$$(n-p-1)\hat{\sigma}^2_{(i)} = \sum_{j\neq i}\left(y_j - \mathbf{x}_j'\left(\hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}\right)\right)^2$$

$$= \sum_{j\neq i}\left(e_j + \frac{\mathbf{x}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}\right)^2 = \sum_{j\neq i}\left(e_j + \frac{h_{ji}e_i}{1-h_{ii}}\right)^2$$

$$= \sum_{j=1}^{n}\left(e_j + \frac{h_{ji}e_i}{1-h_{ii}}\right)^2 - \left(e_i - \frac{h_{ii}e_i}{1-h_{ii}}\right)$$

Since we know that $e_i - \frac{h_{ii}e_i}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}$, we have that

$$(n-p-1)\hat{\sigma}^2_{(i)} = \sum_{j=1}^{n} e_j^2 + 2\frac{e_j h_{ji} e_i}{1-h_{ii}} + \frac{h_{ji}^2 e_i^2}{(1-h_{ii})^2} - \frac{e_i}{1-h_{ii}}$$

Notice that

$$\mathbf{0} = \mathbf{H}(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{He} \implies \sum_{j=1}^{n} h_{ij}e_j = \sum_{j=1}^{n} e_j h_{ji} = 0$$

We also have that

$$\sum_j h_{ij}h_{ji} = (\mathbf{H}^2)_{ii} = (\mathbf{H})_{ii} = h_{ii} \implies \sum_j h_{ij}^2 = h_{ii}$$

Thus we have

$$(n-p-1)\hat{\sigma}^2_{(i)} = \sum_{j=1}^{n} e_j^2 + \frac{h_{ii}e_i^2}{(1-h_{ii})^2} - \frac{e_i}{1-h_{ii}}$$

$$= \sum_{j=1}^{n}(e_j^2) + \frac{e_i^2}{(1-h_{ii})^2}\left[h_{ii}-1\right]$$

$$(n-p-1)\hat{\sigma}^2_{(i)} = (n-p)\hat{\sigma}^2 - \frac{e_i^2}{1-h_{ii}}$$

Using this identity between $\hat{\sigma}^2_{(i)}$ and $\hat{\sigma}^2$, we can derive a relationship between t_i^2 and r_i^2.

$$t_i^2 = \frac{e_i^2}{\hat{\sigma}_{(i)}(1-h_{ii})} = \frac{e_i^2(n-p-1)}{(1-h_{ii})\left[(n-p)\hat{\sigma}^2 - \frac{e_i^2}{1-h_{ii}}\right]}$$

$$= \frac{e_i^2(n-p-1)}{\hat{\sigma}^2(1-h_{ii})\left[n-p-\frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})}\right]} = \frac{r_i^2(n-p-1)}{n-p-r_i^2}$$

4.7 Leave-One-Out Case Diagnostics

Leave-One-Out case diagnostics deals with answering how the i^{th} case affects the volume of the confidence ellipsoid.

Definition 44 ***Cook's Distance*** *is defined as*

$$C_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\hat{\sigma}^2}$$

We can show that $C_i = \frac{h_{ii} r_i^2}{p(1-h_{ii})}$.

Definition 45 *__DFITS__ is defined as*

$$wk_i = \frac{|\mathbf{x}_i(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})|}{\hat{\sigma}_{(i)} h_{ii}}$$

It also goes by the Welsch-Kul distance measure.

Definition 46 *The __Andrew-Pregibon__ test statistic is defined as*

$$AP_i = \frac{(n-p-1)\hat{\sigma}^2_{(i)} |\mathbf{X}'_{(i)}\mathbf{X}_{(i)}|}{(n-p)\hat{\sigma}^2 |\mathbf{X}'\mathbf{X}|}$$

We can interpret the Andrew-Pregibon confidence interval as the ratio of the areas of the confidence intervals.
We know that $\hat{\boldsymbol{\beta}}_j = \frac{\langle \mathbf{y}, \mathbf{x}_j^{\perp}\rangle}{||\mathbf{x}_j^{\perp}||^2}$. Thus we can see that $var(\hat{\boldsymbol{\beta}}_j) = \frac{\sigma^2}{||x_j^{\perp}||^2}$. We will show that $var(\hat{\boldsymbol{\beta}}_j) = \frac{\sigma^2}{(n-1)\widehat{var}(\mathbf{x}_j)[1-R_j^2]}$ where $R_j^2 = \frac{\text{SSReg of } \mathbf{x}_j \text{ on } \mathbf{x}_i,\ i \neq j}{\sum_{i=1}^n (x_{ji} - \bar{x_{j\cdot}})^2}$. Notice that

$$R_j^2 = \frac{\mathbf{x}_j' \left(\mathbf{P}_{\mathbf{X}_{(j)}} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{x}_j}{\mathbf{x}_j' \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{x}_j}$$

Thus we have

$$1 - R_j^2 = \frac{\mathbf{x}_j' \left(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{(j)}}\right)\mathbf{x}_j}{\mathbf{x}_j' \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{x}_j} = \frac{||\mathbf{x}_j^{\perp}||}{\mathbf{x}_j' \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{x}_j}$$

Since we have that $(n-1)\widehat{var}(\mathbf{x}_j) = \mathbf{x}_j' \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{x}_j$, we have

$$var(\hat{\boldsymbol{\beta}}_j) = \frac{\sigma^2}{(n-1)\widehat{var}(\mathbf{x}_j)[1-R_j^2]} = \frac{\sigma^2}{(n-1)\widehat{var}(\mathbf{x}_j)} \frac{1}{1-R_j^2}$$

Definition 47 *We define the __Variance Inflation Factor__ as*

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

4.8 Lack of Fit

Suppose we have the following hypothesis to test

$$H_0 : \mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\eta}$$

What happens if the model is misspecified in one of the following ways?

1. $var(e_i)$ are correlated

2. mean function is misspecified

3. error distribution is misspecified

For 1 and 3, we would look at residual plots and q-q plots. However, lets take a closer look at 2.
Suppose $\mathbb{E}(\mathbf{y}) = \boldsymbol{\gamma} \neq \boldsymbol{\eta}$. Assuming $\boldsymbol{\gamma}$ is known, let $\boldsymbol{\gamma}_0 = \mathbf{P}_{\mathbf{X}}\boldsymbol{\gamma}$. We will define the model residual vector as $\boldsymbol{\gamma} - \boldsymbol{\gamma}_0 = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\gamma}$. Define

$$\boldsymbol{\Lambda}^2 = \boldsymbol{\gamma}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\gamma} = \boldsymbol{\gamma}'\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)'(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)$$

Now, we can see that

$$\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{\eta}} = \mathbf{y} - \mathbf{P}_{\mathbf{X}}\mathbf{y}$$

$$\mathbb{E}(\mathbf{e}) = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbb{E}(y) = \boldsymbol{\gamma} - \boldsymbol{\gamma}_0$$

$$= \mathbb{E}(\mathbf{e}'\mathbf{e}) = \mathbb{E}(\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}) = \boldsymbol{\gamma}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\gamma} + \sigma^2(n-p)$$

Thus we can see that the $\mathbb{E}(MSE) = \sigma^2 + \frac{\boldsymbol{\Lambda}}{n-p}$. We can also see that the expected regression sum of squares is

$$\mathbb{E}(\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}}) = \mathbb{E}(\mathbf{y}'\mathbf{P}_{\mathbf{X}}\mathbf{y}) = \boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0 + \sigma^2 p$$

From this we can see that the total sum of squares will be

$$\boldsymbol{\Lambda} + \sigma^2(n-p) + \sigma^2(p) + \boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}'\boldsymbol{\gamma} + \sigma^2 n$$

Thus we can make an ANOVA table to summarize the results above:

source	SS	df	$\mathbb{E}(MS)$
Regression	$\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}}$	p	$\frac{\boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0}{p} + \sigma^2$
Error	$\mathbf{e}'\mathbf{e}$	$n-p$	$\sigma^2 + \frac{\boldsymbol{\Lambda}}{n-p}$
Total	$\mathbf{y}'\mathbf{y}$	n	$\frac{\boldsymbol{\gamma}'\boldsymbol{\gamma}}{n} + \sigma^2$

In order for us to derive a lack of fit test, the design matrix, $\mathbf{X}$, needs to have some structure/ requirements. We will need enough replications at different locations of the design space. Fro example, let there be g unique points in our design matrix. Let each point have n_i observations. Thus we have n_i observations at $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{ip}]$. Therefore our design and response matrix should take the following form

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{n_1 1} \\ y_{21} \\ \vdots \\ y_{n_g g} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

In order for us to test whether the mean function is misspecified, we must have that $g > p$. Now we can partition the sum of squared error into the following two parts.

$$\mathbf{e}'\mathbf{e} = \sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \hat{y}_{tr})^2 = \sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \bar{y}_{.r} + \bar{y}_{.r} - \hat{y}_{tr})^2$$

$$= \sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \bar{y}_{.r})^2 + \sum_{r=1}^{g}\sum_{t=1}^{n_r}(\bar{y}_{.r} - \mathbf{x}_r\hat{\boldsymbol{\beta}})^2$$

Notice that $\sum_{t=1}^{n_r}\left(\frac{y_{tr}-\bar{y}_{tr}}{\sigma}\right)^2 \sim \chi^2_{n_r-1}(0)$. Thus we have

$$\mathbb{E}\left[\sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \bar{y}_{.r})^2\right] = \mathbb{E}\left[\sum_{r=1}^{g}\sum_{t=1}^{n_r}\left(\frac{y_{tr} - \bar{y}_{tr}}{\sigma^2}\right)^2\sigma^2\right] = (n-g)\sigma^2$$

where $n = \sum_{i=1}^{g} n_i$. Since we know that $\mathbb{E}(\mathbf{e}'\mathbf{e}) = (n-p)\sigma^2 + \mathbf{\Lambda}$, we know that

$$\mathbb{E}(\sum_{r=1}^{g}\sum_{t=1}^{n_r}(\bar{y}_{.r} - \mathbf{x}_r\hat{\boldsymbol{\beta}})^2) = \mathbb{E}(\mathbf{e}'\mathbf{e}) - \mathbb{E}(\sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \bar{y}_{.r})^2) = (g-p)\sigma^2 + \mathbf{\Lambda}$$

Thus we can summarize these results using an updated ANOVA table

source	SS	df	$\mathbb{E}(MS)$
Regression	$\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}}$	p	$\frac{\boldsymbol{\gamma}_0'\boldsymbol{\gamma}_0}{p} + \sigma^2$
SSPE	$\sum_{r=1}^{g}\sum_{t=1}^{n_r}(y_{tr} - \bar{y}_{.r})^2$	$g-p$	σ^2
SSLOF	$\sum_{r=1}^{g}\sum_{t=1}^{n_r}(\bar{y}_{.r} - \mathbf{x}_r\hat{\boldsymbol{\beta}})^2$	$n-g$	$\sigma^2 + \frac{\mathbf{\Lambda}}{n-g}$
Total	$\mathbf{y}'\mathbf{y}$	n	$\frac{\boldsymbol{\gamma}'\boldsymbol{\gamma}}{n} + \sigma^2$

(SSPE stands for sum of squares - pure error, and SSLOF stands for sum of squares - lack of fit). Consider the following matrix

$$\mathbf{U} = \begin{bmatrix} \mathbf{I}_{n_1} - \frac{\mathbf{1}\mathbf{1}'}{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} - \frac{\mathbf{1}\mathbf{1}'}{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_{n_g} - \frac{\mathbf{1}\mathbf{1}'}{n_g} \end{bmatrix}$$

Thus we can write $SSPE = \mathbf{y}'\mathbf{U}\mathbf{y}$. Since $SSE = \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}$, we know that $SSLOF = \mathbf{y}'(\mathbf{I} - \mathbf{P_X} - \mathbf{U})\mathbf{y}$. Since $\mathbf{U}\mathbf{X} = \mathbf{0}$ and $(\mathbf{I} - \mathbf{P_X})\mathbf{X} = \mathbf{0}$,under H_0, we have

$$SSPE = \mathbf{y}'\mathbf{U}\mathbf{y} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{U}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \sim \chi^2_{n-g}(0)$$

$$SSLOF = \mathbf{y}'(\mathbf{I} - \mathbf{P_X} - \mathbf{U})\mathbf{y} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{P_X} - \mathbf{U})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \sim \chi^2_{g-p}(0)$$

Thus we can use the following F-test to test the following null hypothesis:

$$H_0 : \mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\eta}$$

$$F^* = \frac{SSLOF/(g-p)}{SSPE/(n-g)} \sim F_{g-p,n-g}$$

5 Optimal Design

5.1 Introduction

Consider the following linear regression model

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

Optimal design deals with answering the following question: Given a fixed ample size N, how do we select N points from the design interval $\mathbf{X}$ to observe the response $\mathbf{y}$ in some optimal way. In many linear regression settings, this comes down to minimizing the variance of your estimates in some way. Consider the following example.

Consider the simple linear regression model on the interval $X = [a, b]$. What is the best design for estimating both the slope and intercept?
Since we want to minimize the variance of the slope and intercept, we will look at minimizing the confidence ellipse for β_0 and β_1. We know that the ellipsoid will take the form

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \psi(\alpha)$$

Thus we can minimize the axes of the ellipsoid λ_1 and λ_2 (where λ_1 and λ_2 are the eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$). One way to minimize this is to minimize $det((\mathbf{X}'\mathbf{X})^{-1})$ since we know that the determinant is the product of the eigenvalues. In this case, we can see that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

Thus we can see that

$$det((\mathbf{X}'\mathbf{X})^{-1}) = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

WLOG, we can assume $x \in [-1, 1]$ since $\exists\ g : [a, b] \to [-1, 1]$ that is an isomorphic transformation. It is easy to see that

$$det((\mathbf{X}'\mathbf{X})^{-1}) = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \geq \frac{1}{n}$$

Notice that this happens when $n/2$ of them are -1 and $n/2$ of them are 1. Therefore, the best design for estimating both the slope and intercept is to take equal observations at both ends of the interval.

What if we are only interested in estimating the intercept? This means that we want to minimize the variance of $\hat{\beta}_0$. We know that

$$var(\hat{\beta}_0) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{11}$$

$$= \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \sigma^2 \frac{1}{n - \frac{\left(\sum_{i=1}^n x_i\right)^2}{\sum_{i=1}^n x_i^2}} \geq \frac{\sigma^2}{n}$$

Thus we can see that the variance is minimized whenever $\sum_{i=1}^n x_i = 0$. Thus any symmetric design around 0 is optimal for estimating the intercept only.

What if the goal is to estimate the mean response given at a point x_0 inside $[a, b]$? Thus we want to minimize the variance of $\hat{y}(x_0)$. We know that

$$\hat{y}(x_0) = \begin{bmatrix} 1 & x_0 \end{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Thus we have

$$\begin{aligned} var(\hat{y}(x_0)) &= \sigma^2\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0' = \sigma^2\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0' \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} 1 & x_0 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} 1 \\ x_0 \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sum_{i=1}^n x_i^2 - 2x_0\sum_{i=1}^n x_i + x_0^2 n = \frac{\sum_{i=1}^n (x_i - x_0)^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{\sum_{i=1}^n (x_i - x_0)^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n \left((x_i - \bar{x}) + (\bar{x} - x_0)\right)^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - x_0)^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \geq \frac{1}{n} \end{aligned}$$

We can see that they are equal when $\bar{x} = x_0$. Thus any design around x_0 is optimal.

What if we are interested in estimating the mean response given at a point x_0 outside of $[a, b]$? We have the following expression for the variance of $\hat{y}(x_0)$:

$$var(\hat{y}(x_0)) = \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Note that

$$min_{\mathbf{X}} var(\hat{y}(x_0)) \implies min_{\mathbf{X}} \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

WLOG, we can again assume that $X \in [-1, 1]$. Thus we know that $-1 \leq \bar{x} \leq 1$. By putting all observations at the endpoints of the design window, we can achieve any $\bar{x}$ by adjusting the proportion of observations at each end point. Notice that we can also maximize the variance (denominator) by putting all observations on the end of the design window. Thus let δ be the proportion of the observations put on 1 and $(1 - \delta)$ be the proportion on observations put on -1. Thus we can see that $\bar{x} = 2\delta - 1$. Thus we have

$$\begin{aligned} \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} &= \frac{(2\delta - 1 - x_0)^2}{\delta n(2 - 2\delta)^2 + (1 - \delta)n(2\delta)^2} = \frac{(2\delta - 1 - x_0)^2}{\delta n(4 - 8\delta - 4\delta^2) + (1 - \delta)n4\delta^2} \\ &= \frac{(2\delta - 1 - x_0)^2}{4n(\delta - \delta^2)} \\ \frac{\partial}{\partial\delta} &= \frac{4n(\delta - \delta^2)4(2\delta - 1 - x_0) - (2\delta - 1 - x_0)^2 4n(1 - 2\delta)}{16n^2(\delta - \delta^2)^2} \\ &= \frac{4n(2\delta - 1 - x_0)\left(4(\delta - \delta^2) - (2\delta - 1 - x_0)(1 - 2\delta)\right)}{16n^2(\delta - \delta^2)^2} \end{aligned}$$

We can see that the derivative equals zero when $4(\delta - \delta^2) - (2\delta - 1 - x_0)(1 - 2\delta) = 0$. Thus we have

$$4(\delta - \delta^2) - (2\delta - 1 - x_0)(1 - 2\delta) = -2x_0\delta + 1 + x_0 = 0$$

Thus we have that

$$\hat{\delta} = \frac{1 + x_0}{2x_0}$$

for all $x_0 \in \mathbb{R}$.

5.2 Optimal Design Theory

Definition 48 *An **Exact Optimal Design** tells you exactly how many doses you need, where these dose levels are, and how many subjects to assign to each of these doses in some optimal way.*

Definition 49 *An **Approximate Optimal Design** tells you how many doses you need, where these dose levels are, and roughly how many subjects to assign to each of these doses in some optimal way.*

We will mostly deal with Approximate Optimal Designs in this section. When considering an Approximate Optimal Design problem, we usually ask the following questions:

1. How many points are needed to optimize the criterion? (call this number k)
2. Where are the optimal design points? ($x_1, \dots, x_k \in [a, b]$)
3. What is the optimal proportion of the total observations to take at each of these points? ($w_1, \dots, w_k$)

We will denote a generic approximate design by $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} = \begin{bmatrix} x_1 & x_2 & \dots & x_k \\ w_1 & w_2 & \dots & w_k \end{bmatrix}$$

where $0 \le w_i \le 1$ for $i = 1, \dots, k$ and $\sum_{i=1}^k w_i = 1$. For fixed N, the implemented design, $\boldsymbol{\xi}$, assigns approximately Nw_i subjects to x_i for $i = 1, \dots, k$.

Consider the following model

$$y(x) = f'(x)\boldsymbol{\beta} + \epsilon(x)$$

where $\epsilon(x) \sim \mathcal{N}(0, \sigma^2/\lambda(x))$ where $\lambda(x)$ is some known positive function. We will also assume that $f'(x) \in \mathbb{R}^d$. It can be shown that the Fisher information matrix for a k-point approximate design $\boldsymbol{\xi}$ is proportional to

$$M(\boldsymbol{\xi}) = \sum_{i=1}^{k} \lambda(x_i) w_i f(x_i) f'(x_i)$$

when $k \ge d$. We know from previous classes that

$$cov(\hat{\boldsymbol{\beta}}) \propto M^{-1}(\boldsymbol{\xi})$$

For a nonlinear model, we have $\mathbb{E}(y) = f(x, \boldsymbol{\beta})$. Thus we can replace $f(x)$ in the above formula with the gradient of $f(x, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Now that we have found a formula for the covariance of our regression parameters, what does it mean to minimize the covariance?

Definition 50 *We say that $\boldsymbol{\xi}$ achieves **D-Optimality** if $\boldsymbol{\xi}$ minimizes $|M^{-1}(\boldsymbol{\xi})|$, or equivalently $ln|M^{-1}(\boldsymbol{\xi})|$.*

Definition 51 *We say that $\boldsymbol{\xi}$ achieves **A-Optimality** if $\boldsymbol{\xi}$ minimizes $tr(M^{-1}(\boldsymbol{\xi}))$.*

Definition 52 *We say that $\boldsymbol{\xi}$ achieves **I-Optimality** if $\boldsymbol{\xi}$ minimizes the variance over a response region (X). Thus we have*

$$\min_{\boldsymbol{\xi}} \frac{tr\left(RM^{-1}(\boldsymbol{\xi})\right)}{\int_X dx}$$

where $R = \int_X f(x)f'(x)dx$.

Our goal in optimal design is to find $\boldsymbol{\xi}$ such that it is optimal among all possible $\boldsymbol{\xi}$. We can use the following theorem to help check/ prove that some design is D-Optimal.

Theorem 28 (Equivalence Theorem for D-Optimality) $\boldsymbol{\xi}^*$ *is D-Optimal if and only if* $\lambda(x)f'(x)M^{-1}(\boldsymbol{\xi})f(x) - d \leq 0$ *for all* $x \in X$*, with equality at the design points. In this case,* d *is the dimension of* $f(x)$*.*

When dealing with these problems in real life, sometimes the optimal design is not actually desired. This could be due to some uncertainty in your assumptions, or perhaps due to other reasons. In this case, it may be useful to characterize how much worse this model is compared to the optimal model. We call this the Design Efficiency.

Definition 53 *We define the following quantity as **A-efficiency**:*

$$A\text{-}\mathit{eff}(\boldsymbol{\xi}) = \frac{tr\left(M(\boldsymbol{\xi})\right)}{tr\left(M(\boldsymbol{\xi}^*)\right))}$$

where $\boldsymbol{\xi}^*$ *is the optimal design.*

We can interpret this as sum of the variances of the parameters under the design $\boldsymbol{\xi}$, divided by the sum of the variances of the parameters under the optimal design, $\boldsymbol{\xi}^*$.

Definition 54 *We define the following quantity as **D-efficiency**:*

$$D\text{-}\mathit{eff}(\boldsymbol{\xi}) = \left(\frac{|M(\boldsymbol{\xi})|}{|M(\boldsymbol{\xi}^*)|}\right)^{1/p}$$

where $\boldsymbol{\xi}^*$ *is the optimal design.*

We can interpret this as some ratio of the area of the confidence ellipsoid of the parameters under the design $\boldsymbol{\xi}$, divided by the area of the confidence ellipsoid of the parameters under the optimal design, $\boldsymbol{\xi}^*$.

Thus if $\boldsymbol{\xi}$ has an efficiency of 0.5, $\boldsymbol{\xi}$ needs to be replicated twice in order to do as well as $\boldsymbol{\xi}^*$.

6 Shrinkage Estimators

6.1 Ridge Regression

Ridge regression can be used as a means for improving the estimation of regression coefficients when the predictors are highly correlated. We can also prove that the prediction accuracy with the ridge estimate. The ridge estimate solves the following problem

$$\min_{\hat{\boldsymbol{\beta}}_R} ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R||^2 + k||\hat{\boldsymbol{\beta}}_R||_2^2 \qquad k \geq 0$$

We can show that the solution to this minimization problem will have the following form

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

Using the spectral decomposition of $\mathbf{X}'\mathbf{X}$ ($\mathbf{X}'\mathbf{X} = \mathbf{TDT}'$), we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_R &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= (\mathbf{T}(\mathbf{D} + k\mathbf{I})\mathbf{T}')^{-1}\mathbf{TDT}'\hat{\boldsymbol{\beta}} \\
&= \mathbf{T}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{T}'\mathbf{TDT}'\hat{\boldsymbol{\beta}} \\
&= \mathbf{T}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{T}'\mathbf{TDT}'\hat{\boldsymbol{\beta}} \\
&= \mathbf{T}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{DT}'\hat{\boldsymbol{\beta}} \\
&= \sum_{i=1}^{p} \mathbf{t}_i \frac{d_i \mathbf{t}_i'\hat{\boldsymbol{\beta}}}{d_i + k}
\end{aligned}$$

We can now look for the covariance of $\hat{\boldsymbol{\beta}}_R$.

$$\begin{aligned}
cov(\hat{\boldsymbol{\beta}}_R) &= \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \\
&= \sigma^2\mathbf{T}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{T}'\mathbf{TDT}'\mathbf{T}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{T}' \\
&= \sigma^2\mathbf{T}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{D}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{T}' \\
&= \sigma^2 \sum_{i=1}^{p} \frac{d_i}{(d_i + \lambda)^2}\mathbf{t}_i'\mathbf{t}_i
\end{aligned}$$

Now lets take a look at $\mathbb{E}(\hat{\boldsymbol{\beta}}_R) - \boldsymbol{\beta}$.

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_R) - \boldsymbol{\beta} &= \mathbf{T}(\mathbf{D} + k\mathbf{I})^{-1}\mathbf{DT}'\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= \sum_{i=1}^{p} \left(\frac{d_i}{d_i + k} - 1\right)\mathbf{t}_i\mathbf{t}_i'\boldsymbol{\beta} \\
&= -\sum_{i=1}^{p} \left(\frac{k}{d_i + k}\right)\mathbf{t}_i\mathbf{t}_i'\boldsymbol{\beta}
\end{aligned}$$

We can define the Mean Squared Error of $\hat{\boldsymbol{\beta}}_R$ as $MSE = (bias)(bias)' + Cov$. Thus we have that

$$MSE(\hat{\boldsymbol{\beta}}_R) = \mathbf{T}((\mathbf{D}+k\mathbf{I})^{-1}\mathbf{D}+I)\mathbf{T}'\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{T}((\mathbf{D}+k\mathbf{I})^{-1}\mathbf{D}+I)\mathbf{T}'+\sigma^2\mathbf{T}(\mathbf{D}+k\mathbf{I})^{-1}\mathbf{D}(\mathbf{D}+k\mathbf{I})^{-1}\mathbf{T}'$$

Thus we can use the following to compare $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}_R$ using any one of the following metrics:

1. $tr(Cov(\hat{\boldsymbol{\beta}}_R)) - tr(Cov(\hat{\boldsymbol{\beta}}))$
2. $tr(MSE(\hat{\boldsymbol{\beta}}_R)) - tr(MSE(\hat{\boldsymbol{\beta}}))$
3. $||\mathbf{X}\hat{\boldsymbol{\beta}}_R - \mathbf{X}\hat{\boldsymbol{\beta}}||_2^2$

6.2 Stein Estimation

Suppose that we have that

$$\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

Thus we have that $\hat{\boldsymbol{\mu}} = \mathbf{P_X}\mathbf{y}$, $\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{P_X})\mathbf{y}$, and $rank(\mathbf{X}) = p$. We know that

$$||\hat{\boldsymbol{\mu}}||^2 = ||\mathbf{P_X}\mathbf{y}||^2 = \mathbf{y}'\mathbf{P_X}\mathbf{y}$$

We have that

$$\mathbb{E}(||\hat{\boldsymbol{\mu}}||^2) = \boldsymbol{\mu}'\boldsymbol{\mu} + \sigma^2 tr(\mathbf{P_X}) = ||\boldsymbol{\mu}||^2 + p\sigma^2 > ||\boldsymbol{\mu}||^2$$

This suggests that at least some elements of the estimate are too large. Thus consider shrinking the estimate in the following from $\tilde{\boldsymbol{\mu}} = c\hat{\boldsymbol{\mu}}$, where $0 < c < 1$. It is apparent that this will be biased, but is it possible to choose a c such that $\tilde{\boldsymbol{\mu}}$ has a smaller standardized square error loss?

$$L(\tilde{\boldsymbol{\mu}}, (\boldsymbol{\mu}, \sigma^2)) = \frac{||\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2}{\sigma^2}$$

We can define the risk as the averaging of the standardized square error loss.

$$R(\tilde{\boldsymbol{\mu}}, (\boldsymbol{\mu}, \sigma^2)) = \frac{\mathbb{E}||\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2}{\sigma^2}$$

We can see that

$$R(\hat{\boldsymbol{\mu}}, (\boldsymbol{\mu}, \sigma^2)) = \frac{\mathbb{E}\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} - 2\hat{\boldsymbol{\mu}}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu}}{\sigma^2} = \frac{\boldsymbol{\mu}'\boldsymbol{\mu} + p\sigma^2 - 2\boldsymbol{\mu}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu}}{\sigma^2} = p$$

Lets take a look at the risk function of the proposed estimator

$$R(\tilde{\boldsymbol{\mu}}, (\boldsymbol{\mu}, \sigma^2)) = \frac{\mathbb{E}\tilde{\boldsymbol{\mu}}'\tilde{\boldsymbol{\mu}} - 2\tilde{\boldsymbol{\mu}}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu}}{\sigma^2} = \frac{c^2\boldsymbol{\mu}'\boldsymbol{\mu} + c^2p\sigma^2 - 2c\boldsymbol{\mu}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\mu}}{\sigma^2} = \frac{c^2p\sigma^2 + (1-c)^2||\boldsymbol{\mu}||^2}{\sigma^2}$$

Lets minimize the above risk to try and find what c should be

$$\frac{\partial}{\partial c}\left(c^2p\sigma^2 + (1-c)^2||\boldsymbol{\mu}||^2\right) = 2cp\sigma^2 - 2(1-c)||\boldsymbol{\mu}||^2 = 0$$

Thus we have that the optimal is

$$c^* = \frac{||\boldsymbol{\mu}||^2}{||\boldsymbol{\mu}||^2 + p\sigma^2} = 1 - \frac{p\sigma^2}{||\boldsymbol{\mu}||^2 + p\sigma^2}$$

However, we do not know $||\boldsymbol{\mu}||^2$ since we do not know $\boldsymbol{\mu}$. Thus we can consider estimators of the following form

$$\tilde{\boldsymbol{\mu}} = \left(1 - \frac{c\hat{\sigma}^2}{||\hat{\boldsymbol{\mu}}||^2}\right)\hat{\boldsymbol{\mu}}$$

Lemma 28 *1. If $X \sim \chi^2_n$, $\mathbb{E}(X^{-1}) = \frac{1}{n-2}$, $n > 2$.*

2. If $\mathbf{u} \sim \mathcal{N}_p(\boldsymbol{\theta}, \mathbf{I})$ and $k \sim Poisson(||\boldsymbol{\theta}||^2/2)$, then

a) $\mathbb{E}\left(\frac{1}{||\mathbf{u}||^2}\right) = \mathbb{E}\left(\frac{1}{p-2+2k}\right)$

b) $\mathbb{E}\left(\frac{\mathbf{u}'(\mathbf{u}-\boldsymbol{\theta})}{||\mathbf{u}||^2}\right) = \mathbb{E}\left(\frac{p-2}{p-2+2k}\right)$

Proof:

1. We know that
$$\mathbb{E}(X^{-1}) = \int_0^\infty \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-2} e^{-x/2} dx$$
Since we have that $\Gamma(x+1) = x\Gamma(x)$, we have
$$= \frac{1}{2(n-2)} \int_0^\infty \frac{2}{2^{n-2/2}\Gamma(n-2/2)} x^{n-2/2-1} e^{-x/2} dx$$
Since what is inside the integrand is the pdf of a chi-squared distribution with $n-2$ degrees of freedom, we know that it integrates to 1. Thus we have
$$\mathbb{E}(X^{-1}) = \frac{1}{n-2}$$
for $n > 2$.

2. a) We will start with the fact that if $V|k \sim \chi^2_{p+2k}(0)$, then $V \sim \chi^2_p(||\boldsymbol{\theta}||^2)$. By definition of a non-central χ^2, we know that $||\boldsymbol{\mu}||^2 = \boldsymbol{\mu}'\boldsymbol{\mu} \sim \chi^2_p(||\boldsymbol{\theta}||^2)$. Thus we have
$$\mathbb{E}\left(\frac{1}{||\boldsymbol{\mu}||^2}\right) = \mathbb{E}\left(\mathbb{E}\frac{1}{||\boldsymbol{\mu}||^2}|k\right) = \mathbb{E}\left(\frac{1}{p+2k-2}\right)$$
by part 1.

   b) Let $\mathbf{u}' = (u_1, \dots, u_p)$ and $\boldsymbol{\theta}' = (\theta_1, \dots, \theta_p)$. Thus we have
$$\mathbb{E}\left(\frac{1}{||\boldsymbol{\mu}||^2}\right) = \int \frac{1}{\boldsymbol{\mu}'\boldsymbol{\mu}} e^{-(1/2)(\boldsymbol{\mu}-\boldsymbol{\theta})'(\boldsymbol{\mu}-\boldsymbol{\theta})} d\boldsymbol{\mu}$$
$$\frac{\partial}{\partial\theta_i}\mathbb{E}\left(\frac{1}{||\boldsymbol{\mu}||^2}\right) = \int \frac{1}{\boldsymbol{\mu}'\boldsymbol{\mu}} \frac{\partial}{\partial\theta_i} e^{-(1/2)(\boldsymbol{\mu}-\boldsymbol{\theta})'(\boldsymbol{\mu}-\boldsymbol{\theta})} d\boldsymbol{\mu}$$
$$= \int \frac{\mu_i - \theta_i}{\boldsymbol{\mu}'\boldsymbol{\mu}} e^{-(1/2)(\boldsymbol{\mu}-\boldsymbol{\theta})'(\boldsymbol{\mu}-\boldsymbol{\theta})} d\boldsymbol{\mu}$$
Thus we have that
$$(*) \quad \frac{\partial}{\partial\theta_i}\mathbb{E}\left(\frac{1}{||\boldsymbol{\mu}||^2}\right) = \mathbb{E}\left(\frac{\mu_i - \theta_i}{||\boldsymbol{\mu}||^2}\right)$$
From part (a), we know that $\mathbb{E}\left(\frac{1}{||\boldsymbol{\mu}||^2}\right) = \mathbb{E}\left(\frac{1}{p+2k-2}\right)$, thus we can look at the following
$$\frac{\partial}{\partial\theta_i}\mathbb{E}\left(\frac{1}{p+2k-2}\right) = \frac{\partial}{\partial\theta_i}\sum_{i=1}^{\infty} \frac{e^{||\boldsymbol{\theta}||^2/2}}{(p+2k-2)k!}\frac{||\boldsymbol{\theta}||^{2k}}{2^k}$$
$$= \sum_{i=1}^{\infty} \frac{\theta_i(2k - ||\boldsymbol{\theta}||^2)e^{||\boldsymbol{\theta}||^2/2}}{||\boldsymbol{\theta}||^2(p+2k-2)k!}\frac{2||\boldsymbol{\theta}||^{2k}}{2^k}$$
$$(**) \quad \frac{\partial}{\partial\theta_i}\mathbb{E}\left(\frac{1}{p+2k-2}\right) = \frac{\theta_i}{||\boldsymbol{\theta}||^2}\mathbb{E}\left(\frac{2k - ||\boldsymbol{\theta}||^2}{p+2k-2}\right) \qquad k \sim Poisson(||\boldsymbol{\theta}||^2/2)$$
Thus from $(*)$ and $(**)$, we have that
$$\mathbb{E}\left(\frac{\boldsymbol{\mu}-\boldsymbol{\theta}}{||\boldsymbol{\mu}||^2}\right) = \frac{\boldsymbol{\theta}}{||\boldsymbol{\theta}||^2}\mathbb{E}\left(\frac{2k - ||\boldsymbol{\theta}||^2}{p+2k-2}\right)$$
Thus we have
$$\mathbb{E}\left(\frac{\boldsymbol{\mu}'(\boldsymbol{\mu}-\boldsymbol{\theta})}{||\boldsymbol{\mu}||^2}\right) = \mathbb{E}\left(\frac{\boldsymbol{\mu}'\boldsymbol{\mu} - \boldsymbol{\theta}'\boldsymbol{\theta} - \boldsymbol{\theta}'(\boldsymbol{\mu}-\boldsymbol{\theta})}{||\boldsymbol{\mu}||^2}\right)$$

$$= 1 - ||\boldsymbol{\theta}||^2\mathbb{E}\left(\frac{1}{||\boldsymbol{\mu}||^2}\right) - \boldsymbol{\theta}'\mathbb{E}\left(\frac{\boldsymbol{\mu}-\boldsymbol{\theta}}{||\boldsymbol{\mu}||^2}\right)$$

$$= 1 - ||\boldsymbol{\theta}||^2\mathbb{E}\left(\frac{1}{p+2k-2}\right) - \frac{\boldsymbol{\theta}'\boldsymbol{\theta}}{||\boldsymbol{\theta}||^2}\mathbb{E}\left(\frac{2k-||\boldsymbol{\theta}||^2}{p+2k-2}\right)$$

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

Thus we have that

$$\mathbb{E}\left(\frac{\boldsymbol{\mu}'(\boldsymbol{\mu}-\boldsymbol{\theta})}{||\boldsymbol{\mu}||^2}\right) = \mathbb{E}\left(\frac{p-2}{p+2k-2}\right)$$

□

Using this lemma above, we show that if $\tilde{\boldsymbol{\mu}} = \left(1 - \frac{c\hat{\sigma}^2}{||\hat{\boldsymbol{\mu}}||^2}\right)\hat{\boldsymbol{\mu}}$, then $R(\tilde{\boldsymbol{\mu}},(\mu,\sigma^2)) = p - [2c(p-2) - c^2\frac{n-p+2}{n-p}]\mathbb{E}\left(\frac{1}{p+2k-2}\right)$, where $k \sim Poisson(||\boldsymbol{\mu}||^2/2\sigma^2)$.
We know that

$$R(\tilde{\boldsymbol{\mu}},(\mu,\sigma^2)) = \mathbb{E}\left(\frac{||\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}||^2}{\sigma^2}\right) = \mathbb{E}\left(\frac{||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}-c\frac{\hat{\sigma^2}}{||\hat{\boldsymbol{\mu}}||^2}\hat{\boldsymbol{\mu}}||^2}{\sigma^2}\right)$$

$$= \frac{\mathbb{E}||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}||^2}{\sigma^2} - \frac{2c}{\sigma^2}\mathbb{E}(\hat{\sigma}^2)\mathbb{E}\left(\frac{(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})'\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||^2}\right) + \frac{c^2}{\sigma^2}\mathbb{E}(\hat{\sigma}^4)\mathbb{E}\left(\frac{1}{||\hat{\boldsymbol{\mu}}||^2}\right)$$

We know the following:

1. $\frac{\mathbb{E}||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}||^2}{\sigma^2} = p$
2. $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}(0)$
3. $\mathbb{E}\hat{\sigma}^2 = \sigma^2$
4. $\mathbb{E}\hat{\sigma}^4 = var(\hat{\sigma}^2) + (\mathbb{E}\hat{\sigma}^2)^2 = \frac{2(n-p)\sigma^4}{(n-p)^2} + \sigma^4 = \frac{(n-p+2)\sigma^4}{(n-p)}$

Let $\mathbf{Z}$ be an orthonormal basis for $\mathcal{C}(\mathbf{X})$. Let $\mathbf{u} = \frac{\mathbf{Z}'\mathbf{y}}{\sigma}$ and $\boldsymbol{\theta} = \frac{\mathbf{Z}'\boldsymbol{\mu}}{\sigma}$. Then we know that $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\theta},\mathbf{I})$. Since $\mu \in \mathcal{C}(\mathbf{X})$, we know that

$$\boldsymbol{\mu} = \mathbf{P_X}\boldsymbol{\mu} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\mu} = \mathbf{Z}\mathbf{Z}'\boldsymbol{\mu} = \sigma\mathbf{Z}\boldsymbol{\theta}$$

$$\hat{\boldsymbol{\mu}} = \mathbf{P_X}\mathbf{y} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{Z}\mathbf{Z}'\mathbf{y} = \sigma\mathbf{Z}\mathbf{u}$$

Thus we have that

$$\frac{||\boldsymbol{\mu}||^2}{2\sigma^2} = \frac{||\boldsymbol{\theta}||^2}{2}, \quad \frac{1}{||\hat{\boldsymbol{\mu}}||^2} = \frac{1}{\sigma^2||\mathbf{u}||^2}$$

and that

$$\frac{(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})'\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||^2} = \frac{(\mathbf{u}-\boldsymbol{\theta})'\mathbf{u}}{||\mathbf{u}||^2}$$

Therefore, we have

$$\mathbb{E}\left(\frac{(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})'\hat{\boldsymbol{\mu}}}{||\hat{\boldsymbol{\mu}}||^2}\right) = \mathbb{E}\left(\frac{(\mathbf{u}-\boldsymbol{\theta})'\mathbf{u}}{||\mathbf{u}||^2}\right) = \mathbb{E}\left(\frac{p-2}{p+2k-2}\right)$$

and

$$\mathbb{E}\left(\frac{1}{||\hat{\boldsymbol{\mu}}||^2}\right) = \frac{1}{\sigma^2}\mathbb{E}\left(\frac{1}{||\mathbf{u}||^2}\right) = \frac{1}{\sigma^2}\mathbb{E}\left(\frac{1}{p+2k-2}\right)$$

Thus we have that

$$R(\tilde{\boldsymbol{\mu}}, (\mu, \sigma^2)) = p - \left[2c(p-2) + \frac{c^2(n-p+2)}{n-p}\right] \mathbb{E}\left(\frac{1}{p+2k-2}\right)$$

If we minimize this with respect to c, will get that

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

It can be shown that this provides a smaller risk function that the least squares estimate for all $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. Thus we can say that $\tilde{\boldsymbol{\mu}}$ is uniformly better than $\hat{\boldsymbol{\mu}}$ in terms of the risk function. Thus the James Stein estimator is the following

$$\tilde{\boldsymbol{\mu}} = \left(1 - \frac{c\hat{\sigma}^2}{||\hat{\boldsymbol{\mu}}||^2}\right)\hat{\boldsymbol{\mu}}$$

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

7 ANOVA and Linear Mixed Effects Models

7.1 One-way ANOVA

The typical model for a one-way ANOVA model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad i = 1, \dots, a, \quad j = 1, \dots, n$$

where $\sum_{i=1}^{a} \tau_i = 0$ and $\epsilon_{ij} = \mathcal{N}(0, \sigma^2)$. Suppose we want to test the following

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

$$H_1 : \tau_i \neq \tau_j \quad 1 \leq i \neq j \leq a$$

First we will find the least squares estimate. We can see that we have to minimize the following function:

$$\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \mu - \tau_i)^2$$

We can show that the least squares estimate will lead to

$$\hat{\mu} = \bar{y}_{..} \quad \hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..} \quad i = 1, \dots, a$$

We know that $SSTO = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$. we can decompose this into SSE and $SSTr$ in the following way

$$\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_{i.} - \bar{y}_{..})^2$$

let $\mathbf{y}_{i.}$ be the vector of responses for treatment group i. Thus we have

$$\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 = \mathbf{y}_{i.}' \frac{(\mathbf{I} - \frac{\mathbf{11}'}{n})}{\sigma^2} \mathbf{y}_{i.} \sigma^2$$

Since $\mathbf{y}_i \sim \mathcal{N}(\mathbf{1}(\mu + \tau_i), \sigma^2 \mathbf{I})$, we know that $\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 \sim \sigma^2 \chi^2_{n-1}(0)$. Thus we have that

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 \sim \sigma^2 \chi^2_{a(n-1)}$$

We can also let

$$\mathbf{U} = \begin{bmatrix} \frac{\mathbf{11}'}{n} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{11}'}{n} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \frac{\mathbf{11}'}{n} \end{bmatrix}$$

Thus we have that, under H_0,

$$SSE = \mathbf{y}'(\mathbf{I} - \mathbf{U})\mathbf{y} = \mathbf{y}' \frac{(\mathbf{I} - \mathbf{U})}{\sigma^2} \mathbf{y} \sigma^2 \sim \sigma^2 \chi^2_{g(n-1)}(0)$$

Thus we know that $SSTr = \mathbf{y}'(\mathbf{U} - \frac{\mathbf{11}'}{n})\mathbf{y}$. Thus we have

$$SSTr = \mathbf{y}'(\mathbf{U} - \frac{\mathbf{11}'}{N})\mathbf{y} = \mathbf{y}'\left(\frac{(\mathbf{U} - \frac{\mathbf{11}'}{N})}{\sigma^2}\right)\mathbf{y}\sigma^2 \sim \sigma^2\chi^2_{a-1}(0)$$

We can also find the expected value of $SSTr$. Thus we have

$$\mathbb{E}(\mathbf{y}'(\mathbf{U} - \frac{\mathbf{11}'}{N})\mathbf{y}) = \boldsymbol{\mu}'(\mathbf{U} - \frac{\mathbf{11}'}{N})\boldsymbol{\mu} + tr((U - \frac{\mathbf{11}'}{N})\mathbf{I}\sigma^2) = n\sum_{i=1}^{a}\tau_i^2 + \sigma^2(a-1)$$

Thus we know the following facts

1. $\mathbb{E}(SSTr) = n\sum_{i=1}^{a}\tau_i^2 + \sigma^2(a-1)$
2. Under H_0, $\frac{g(n-1)MSE}{\sigma^2} \sim \chi^2_{g(n-1)}(0)$
3. Under H_0, $\frac{(a-1)MSTr}{\sigma^2} \sim \chi^2_{a-1}$

Thus we can test H_0 using the following F-test:

$$F^* = \frac{((a-1)MSTR)/(\sigma^2(a-1))}{((a(n-1))MSE)/(\sigma^2(a(n-1)))} = \frac{MSTR}{MSE} \sim F_{a-1,a(n-1)}$$

7.2 Mixed Effects ANOVA Model

Suppose we treat the difference in treatment effect as a random variable. Thus we would have the following model:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}; \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\epsilon) \quad \tau_i \sim \mathcal{N}(0, \sigma^2_\tau) \quad \epsilon_{ij} \perp\!\!\!\perp \tau_i$$

Suppose we want to test the following hypothesis

$$H_0 : \sigma^2_\tau = 0, \qquad H_1 : \sigma^2_\tau > 0$$

We can see that

$$\bar{y}_{i.} = \mu + \tau_i + \bar{\epsilon}_{.i}$$
$$\bar{y} = \mu + \bar{\tau} + \bar{\epsilon}$$

Thus we can see that

$$SSE = \sum\sum(y_{ij} - \bar{y_i})^2 = \sum\sum\frac{(\epsilon_{ij} - \bar{\epsilon_{i.}})^2}{\sigma^2_\epsilon}\sigma^2_\epsilon$$

Therefore, we can see that $SSE \sim \sigma^2_\epsilon\chi^2_{a(n-1)}$. We can also see that we have

$$SSTr = \sum\sum(\bar{y}_{i.} - \bar{y})^2 = \sum\sum(\tau_i - \bar{\tau} + \bar{\epsilon}_{i.} - \bar{\epsilon})^2$$

Under H_0, we know that $\tau_i - \bar{\tau} = 0$. Thus we have that

$$SSTr = \sum\sum(\bar{\epsilon}_{i.} - \bar{\epsilon})^2 = n\sum_{i=1}^{a}(\bar{\epsilon}_{i.} - \bar{\epsilon})^2 = \sum_{i=1}^{a}\frac{(\bar{\epsilon}_{i.} - \bar{\epsilon})^2}{\sigma^2/n}\sigma^2$$

Since $\frac{(\bar{\epsilon}_{i.} - \bar{\epsilon})^2}{\sigma^2/n} \sim \chi^2_1(0)$, we have that $SSTr \sim \chi^2_{a-1}(0)$. Thus to test H_0, we can use the following test statistic:

$$F^* = \frac{(a(n-1))MSE/(\sigma^2 a(n-1))}{(a-1)MSTr/(\sigma^2(a-1))} = \frac{MSE}{MSTr} \sim F_{a(n-1),(a-1)}$$

Now, how do we estimate σ_τ^2? Lets look at the following

$$\mathbb{E}(SSTr) = \mathbb{E}\left[n\sum_{i=1}^{a}\frac{(\tau_i-\bar{\tau})^2}{\sigma_\tau^2}\sigma_\tau^2 + \sum_{i=1}^{a}\frac{n(\bar{\epsilon}_{i.}-\bar{\epsilon})^2}{\sigma_\epsilon^2}\sigma_\epsilon^2\right]$$

Since we know that

$$\mathbb{E}\left(\sum_{i=1}^{a}\frac{n(\bar{\epsilon}_{i.}-\bar{\epsilon})^2}{\sigma_\epsilon^2}\sigma_\epsilon^2\right) = (a-1)\sigma_\epsilon^2$$

and

$$\mathbb{E}\left(\sum_{i=1}^{a}\frac{(\tau_i-\bar{\tau})^2}{\sigma_\tau^2}\sigma_\tau^2\right) = \sigma_\tau^2(a-1)$$

Thus we know that

$$\mathbb{E}(SSTr) = n(a-1)\sigma_\tau^2(a-1) + (a-1)\sigma_\epsilon^2$$

Since $\mathbb{E}(MSTr) = \frac{\mathbb{E}(SSTr)}{a-1}$, we have that

$$\mathbb{E}(MSTr) = n\sigma_\tau^2 + \sigma_\epsilon^2$$

Since $\mathbb{E}(MSE) = \sigma_\epsilon^2$, we have that

$$\hat{\sigma_\tau^2} = \frac{MSTr - MSE}{n}$$

Can we derive a distribution for σ_τ^2? Not really, but we can define a distribution on the quantity known as the intraclass correlation

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

This quantity is important because it reflects the proportion of the variance of an observation that is the result of differences between treatments. We can show that

$$F^* = \frac{MSTr/(n\sigma_\tau^2+\sigma_\epsilon^2)}{MSE/\sigma_\epsilon^2} \sim F_{a-1,a(n-1)}$$

Thus we have

$$\begin{aligned}
1-\alpha &= \left(F_{a-1,a(n-1),1-\alpha/2} \le \frac{MSTr}{MSE}\frac{\sigma_\epsilon^2}{n\sigma_\tau^2+\sigma_\epsilon^2} \le F_{a-1,a(n-1),\alpha/2}\right)\\
&= \left(\frac{MSE}{MSTr}F_{a-1,a(n-1),1-\alpha/2} \le \frac{\sigma_\epsilon^2}{n\sigma_\tau^2+\sigma_\epsilon^2} \le \frac{MSE}{MSTr}F_{a-1,a(n-1),\alpha/2}\right)\\
&= \left(\frac{MSTr}{MSE}\frac{1}{F_{a-1,a(n-1),1-\alpha/2}} \le \frac{n\sigma_\tau^2+\sigma_\epsilon^2}{\sigma_\epsilon^2} \le \frac{MSTr}{MSE}\frac{1}{F_{a-1,a(n-1),\alpha/2}}\right)\\
&= \left(\frac{MSTr}{MSE}\frac{1}{F_{a-1,a(n-1),1-\alpha/2}} - 1 \le \frac{n\sigma_\tau^2}{\sigma_\epsilon^2} \le \frac{MSTr}{MSE}\frac{1}{F_{a-1,a(n-1),\alpha/2}} - 1\right)\\
&= \left(\frac{1}{n}\left(\frac{MSTr}{MSE}\frac{1}{F_{a-1,a(n-1),1-\alpha/2}} - 1\right) \le \frac{\sigma_\tau^2}{\sigma_\epsilon^2} \le \frac{1}{n}\left(\frac{MSTr}{MSE}\frac{1}{F_{a-1,a(n-1),\alpha/2}} - 1\right)\right)\\
&= \left(L \le \frac{\sigma_\tau^2}{\sigma_\epsilon^2} \le U\right)
\end{aligned}$$

Notice that

$$\frac{\sigma_\tau^2/\sigma_\epsilon^2}{1+\sigma_\tau^2/\sigma_\epsilon^2} = \frac{\sigma_\tau^2}{\sigma_\epsilon^2+\sigma_\tau^2}$$

Thus we have that

$$1-\alpha = \left(\frac{L}{1+L} \le \frac{\sigma_\tau^2}{\sigma_\tau^2+\sigma_\epsilon^2} \le \frac{U}{1+U}\right)$$

7.3 Rules for Expected Mean Squares

Suppose we have the following model:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}; \quad i = 1, \dots, a; \quad j = 1, \dots, b; \quad k = 1, \dots n$$

in this model, we have that τ_i is a fixed effect, β_j is a random effect, and $(\tau\beta)_{ij}$ is also a random effect. Thus, using the following rules, we can derive the expected sum of squares

1. The error term in the model, $\epsilon_{ij\dots m}$, is written as $\epsilon_{(ij\dots)m}$ where m is the subscript that denotes the replication.

2. For each term in the model, divide the subscripts into the following three classes:
   a) live - subscripts that are present in the term and are not in parentheses
   b) dead - subscripts that are present in the term and are in the parentheses
   c) absent - those subscripts that are present in the model, but not in that particular term

   For example, in $(\tau\beta)_{ij}$, i and j are live and k is absent. In $\epsilon_(ij)k$, k is live and i and j are dead.

3. The number of degrees of freedom for any term in the model is the product of the number of levels associated with each dead subscript and the number of levels minus 1 associated with each live subscript. Thus for $(\tau\beta)_{ij}$ is $(a-1)(b-1)$, and $\epsilon_{(ij)k}$ is $ab(n-1)$.

4. Each term in the model either has a variance component or a fixed factor associated with it. If an interaction contains at least one random effect, the entire interaction is considered to be random. Thus the variance component for β is σ_β and the effect of the fixed effect is represented by the sum of squares of the model components associated with that factor, divided by the associated degrees of freedom. Thus the effect for A is $\frac{\sum_{i=1}^{a} \tau_i^2}{a-1}$.

5. To obtain the expected mean squares, prepare the following table. There is a row for each model component and a column for each subscript. Over each subscript, write the number of levels of the factor associated with that subscript and whether the factor is fixed (F) or random (R). Replicates (associated with ϵ) are always considered to be random.
   a) In each row, write a 1 if one of the dead subscripts in the row components matches the subscripts in the column:

	F	F	R
	a	b	n
Factor	i	j	k
τ_i			
β_j			
$(\tau\beta)_{ij}$			
$\epsilon_{(ij)k}$	1	1	

   b) In each row, if any of the subscripts on the row component match the subscript in the column, write 0 if the column is headed by a fixed factor and a 1 if the column is headed by a random factor

	F	F	R
	a	b	n
Factor	i	j	k
τ_i	0		
β_j		0	
$(\tau\beta)_{ij}$	0	0	
$\epsilon_{(ij)k}$	1	1	1

c) In the remaining empty row positions, write the number of levels shown above the column heading

	F	F	R
	a	b	n
Factor	i	j	k
τ_i	0	b	n
β_j	a	0	n
$(\tau\beta)_{ij}$	0	0	n
$\epsilon_{(ij)k}$	1	1	1

d) To obtain the expected mean square for any model component, first cover all columns headed by live subscripts on that component. Then, in each row that contains at least the same subscripts as those on the component being considered, take the product of the visible numbers and multiply by the appropriate fixed of random factor. The sum of these quantities is the expected means square of the model component being considered.

	F	F	R	
	a	b	n	
Factor	i	j	k	MSE
τ_i	0	b	n	$\sigma^2_\epsilon + bn\left(\frac{\sum \tau_i^2}{a-1}\right) + n\sigma^2_{\tau\beta}$
β_j	a	0	n	$an\sigma^2_\beta + \sigma^2_\epsilon$
$(\tau\beta)_{ij}$	0	0	n	$n\sigma^2_{\tau\beta} + \sigma^2_\epsilon$
$\epsilon_{(ij)k}$	1	1	1	σ^2_ϵ

Suppose we want to test if $H_0 : \tau_1 = \cdots = \tau_a$ using the expected means calculated above. Thus we can try to isolate the term $\sum \tau_i^2$. Thus we will reject H_0 if

$$\frac{MSA}{MSAB} > F_{a-1,(a-1)(b-1),\alpha}$$

Suppose we want to test $H_0 : \sigma^2_\beta = 0$. Thus we would reject H_0 if

$$\frac{MSB}{MSE} > F_{b-1,ab(n-1),\alpha}$$

Suppose we want to test $H_0 : \sigma^2_{\tau\beta} = 0$. Thus we would reject H_0 if

$$\frac{MSAB}{MSE} > F_{(a-1)(b-1),(n-1)ab,\alpha}$$

Consider the following model:

$$y_{ijkl} = \mu + \tau_i + \beta_j + \alpha_k + (\tau\beta)_{ij} + (\tau\alpha)_{ik} + (\beta\alpha)_{jk} + (\tau\beta\alpha)_{ijk} + \epsilon_{(ijk)l}$$

where A is fixed, and B and C are random.

	F	R	R	R	
	a	b	c	n	
Factor	i	j	k	l	$\mathbb{E}(MSE)$
τ_i	0	b	c	n	$bcn\left(\frac{\sum \tau^2}{a-1}\right) + cn\sigma^2_{\tau\beta} + bn\sigma^2_{\tau\alpha} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_\epsilon$
β_j	a	1	c	n	$acn\sigma^2_\beta + cn\sigma^2_{\tau\beta} + an\sigma^2_{\beta\alpha} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_\epsilon$
α_k	a	b	1	n	$abn\sigma^2_\alpha + bn\sigma^2_{\tau\alpha} + an\sigma^2_{\beta\alpha} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_\epsilon$
$(\tau\beta)_{ij}$	1	1	c	n	$cn\sigma^2_{\tau\beta} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_\epsilon$
$(\tau\alpha)_{ik}$	1	b	1	n	$bn\sigma^2_{\tau\alpha} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_\epsilon$
$(\beta\alpha)_{jk}$	a	1	1	n	$an\sigma^2_{\beta\alpha} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_\epsilon$
$(\tau\beta\alpha)_{ijk}$	1	1	1	n	$n\sigma^2_{\tau\beta\alpha} + \sigma^2_\epsilon$
$\epsilon_{(ijk)l}$	1	1	1	1	σ^2_ϵ

Suppose we wish to test $H_0 : \tau_1 = \cdots = \tau_a = 0$. We can see from the expected MSEs that we cannot isolate the $\sum \tau^2$ term. Thus no exact F-test exists. However, we can use an approximate test. Consider the following

$$\gamma_1 = MSA + MSABC \implies \mathbb{E}(\gamma_1) = bcn \left(\frac{\sum \tau^2}{a-1} \right) + cn\sigma^2_{\tau\beta} + bn\sigma^2_{\tau\alpha} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_{\epsilon} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_{\epsilon}$$

$$\gamma_2 = MSAB + MSAC = cn\sigma^2_{\tau\beta} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_{\epsilon} + bn\sigma^2_{\tau\alpha} + n\sigma^2_{\tau\beta\alpha} + \sigma^2_{\epsilon}$$

Thus we can see that $\gamma_1 - \gamma_2 = bcn \left(\frac{\sum \tau^2}{a-1} \right)$. Therefore we can use the following approximation to test H_0

$$F^* = \frac{MSA + MSABC}{MSAB + MSAC} \sim^{approx} F_{p,q}$$

However, how do we find p and q?

Theorem 29 (Satterthwaite's Approximation) *Let $u_i \sim \chi^2_{n_i}$ for $i = 1, \ldots, k$. Let $U = \sum_{i=1}^{K} a_i u_i$. Then we can say that $U \sim^{approx} \chi^2_{a*}$, where*

$$a^* = \frac{(\sum a_i u_i)^2}{\sum \left(\frac{(a_i u_i)^2}{n_i} \right)}$$

Thus in the above question, we know that the degrees of freedom associated with MSA are $a-1$, the degrees of freedom associated with $MSABC$ are $(a-1)(b-1)(c-1)$, the degrees of freedom associated with $MSAB$ are $(a-1)(b-1)$, and the degrees of freedom associated with $MSAC$ are $(b-1)(c-1)$. Thus we have

$$p \approx \frac{(MSA + MSABC)^2}{(MSA^2/(a-1)) + (MSABC^2/((a-1)(b-1)(c-1)))}$$

$$q \approx \frac{(MSAB + MSAC)^2}{(MSAB^2/((a-1)(b-1))) + (MSAC^2/((b-1)(c-1)))}$$

7.4 Linear Mixed Models

Suppose we have the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}); \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}); \quad \mathbf{u} \perp\!\!\!\perp \boldsymbol{\epsilon}$$

Lets use the MLE framework to find the MLE of β and BLUP of $\mathbf{u}$. We know that $\mathbf{y}|\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R})$. Thus we can get the joint distribution of $\mathbf{y}$ and $\mathbf{u}$ in the following way:

$$f(\mathbf{y}, \mathbf{u}) = f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{u}}(\mathbf{u}) \propto exp\left\{ -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \right\} exp\left\{ -\frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u} \right\}$$

$$\ell \propto -\frac{1}{2}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}\mathbf{D}^{-1}\mathbf{u} \right]$$

Differentiating with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$, we get

$$(1) \quad \frac{\partial \ell}{\partial \boldsymbol{\beta}} \propto \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{X}\mathbf{R}^{-1}\mathbf{Z}\mathbf{u} - \mathbf{X}\mathbf{R}^{-1}\mathbf{y}$$

$$(2) \quad \frac{\partial \ell}{\partial \mathbf{u}} \propto \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{D}^{-1}\mathbf{u}$$

We can rewrite it in the following way

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

These set of equations above are known as "Henderson's Normal Equations". Let $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ be the solutions to these equations. From the second equation, we have

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})\tilde{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}$$

$$\tilde{\mathbf{u}} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})$$

Notice that $\mathbf{V} = cov(\mathbf{y}) = cov(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}+\boldsymbol{\epsilon}) = \mathbf{Z}'\mathbf{D}\mathbf{Z}+\mathbf{R}^{-1}$. We can show that $\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$. Thus looking at the first equation, we have

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\tilde{\mathbf{u}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}$$

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}$$

$$\mathbf{X}'(\mathbf{R}^{-1}-\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1})\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{X}(\mathbf{R}^{-1}-\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1})\mathbf{y} = 0$$

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}}\mathbf{X}\mathbf{V}^{-1}\mathbf{y}$$

Thus we arrive at the MLE of $\boldsymbol{\beta}$

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

To obtain $\tilde{\mathbf{u}}$, we have

$$(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})\mathbf{D}\mathbf{Z}' = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{Z}'$$

$$= \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Z}\mathbf{D}\mathbf{Z}'+\mathbf{R}) = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{V}$$

Thus we have

$$\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$$

Therefore, we have

$$\tilde{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})$$

Lets now derive the variances and covariances of the estimates. We know that $cov(\mathbf{u}) = \mathbf{D}$ and $cov(\boldsymbol{\epsilon}) = \mathbf{R}$.

$$cov(\mathbf{y},\mathbf{u}') = cov(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}+\boldsymbol{\epsilon},\mathbf{u}') = \mathbf{Z}cov(\mathbf{u},\mathbf{u}') = \mathbf{Z}\mathbf{D}$$

$$cov(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}cov(\mathbf{y})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

Since $cov(\mathbf{y}) = \mathbf{V}$, we have that

$$cov(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

Let $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. Then we have

1. $\mathbf{P}' = \mathbf{P}$
2. $\mathbf{P}\mathbf{X} = \mathbf{0}$
3. $\mathbf{P}\mathbf{V}\mathbf{P} = \mathbf{P}$

We can re-write $\tilde{\boldsymbol{\mu}}$ as

$$\tilde{\boldsymbol{\mu}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

$$= \mathbf{D}\mathbf{Z}'(\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}$$

$$= \mathbf{D}\mathbf{Z}'\mathbf{P}\mathbf{y}$$

Thus we can see that

1. $cov(\tilde{\boldsymbol{\mu}}) = \mathbf{DZ'PVPZD'} = \mathbf{DZ'PZD}$
2. $cov(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\mu}}) = cov((\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{y}, \mathbf{DZ'Py}) = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{VPZD}$

   $$= (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'PZD} = 0$$

   since $\mathbf{PX} = \mathbf{0} \implies \mathbf{X'P} = \mathbf{0}'$.
3. $cov(\tilde{\boldsymbol{\beta}}, \mathbf{u}) = cov((\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{y}, \mathbf{u}) = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}cov(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}, \mathbf{u}) = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{ZD}$
4. $cov(\tilde{\mathbf{u}}, \mathbf{u}) = cov(\mathbf{DZ'Py}, \mathbf{u}) = \mathbf{DZ'PZD}$
5. $cov(\tilde{\mathbf{u}} - \mathbf{u}) = cov(\tilde{\mathbf{u}}) + cov(\mathbf{u}) - 2cov(\mathbf{u}\tilde{,}\mathbf{u}) = \mathbf{DZ'PZD} + \mathbf{D} - 2\mathbf{DZ'PZD} = \mathbf{D} - \mathbf{DZ'PZD}$
6. $cov(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}} - \mathbf{u}) = cov(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{u}}) - cov(\tilde{\boldsymbol{\beta}}, \mathbf{u}) = -(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{ZD}$

8 Multivariate Statistics

8.1 Graphical Gaussian Models

Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. We can specify the joint distribution by factoring it in the following way.

$$P(\mathbf{Y}) = P(Y_1)P(Y_2|Y_1)\dots P(Y_n|Y_1,\dots Y_{n-1})$$

We can use a DAG or Bayesian Network to represent the dependencies

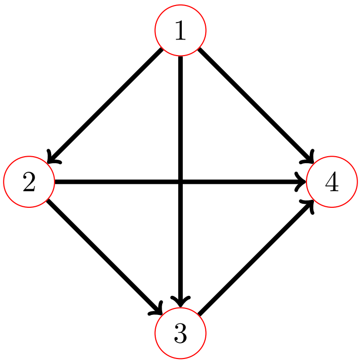

Figure 8.1: A Bayesian network or DAG

We can express this DAG as the following linear models:

$$Y_1 = 0 + \eta_1, \quad \eta_1 \sim \mathcal{N}(0, d_1) \;\; d_1 = \sigma_{11}$$

$$Y_2 = a_{21}y_1 + \eta_2, \quad \eta_2 \sim \mathcal{N}(0, d_2) \;\; d_2 = var(Y_2|Y_1)$$

$$Y_3 = a_{31}y_1 + a_{32}y_2 + \eta_3, \quad \eta_3 \sim \mathcal{N}(0, d_3) \;\; d_3 = var(Y_3|Y_2, Y_1)$$

$$\vdots$$

$$Y_i = \sum_{j=1}^{i-1} a_{ij}y_j + \eta_i, \quad \eta_i \sim \mathcal{N}(0, d_i) \;\; d_i = var(Y_i|Y_1,\dots,Y_{i-1})$$

Since the η_i completely specify $P(Y_i|Y_1,\dots,Y_{i-1})$, and since the densities factor, we know that the η_i are independent. Thus we can write our model as the following:

$$(1) \;\; \mathbf{Y} = \mathbf{AY} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \;\; \mathbf{D} = diag(d_1, d_2, \dots, d_n)$$

We also know that $\mathbf{A}$ will have the following form:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ * & 0 & 0 & \dots & 0 \\ * & * & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & \dots & 0 \end{bmatrix}$$

From (1), we have

$$(\mathbf{I}-\mathbf{A})\mathbf{Y}=\boldsymbol{\eta} \implies var(\mathbf{Y})=(\mathbf{I}-\mathbf{A})^{-1}\mathbf{D}(\mathbf{I}-\mathbf{A})^{-T}$$

$$(\mathbf{I}-\mathbf{A})^{-1}\mathbf{D}(\mathbf{I}-\mathbf{A})^{-T}=\boldsymbol{\Sigma}$$
$$=\mathbf{LDL}'=\mathbf{LD}^{1/2}\mathbf{D}^{1/2}\mathbf{L}'=\tilde{\mathbf{L}}\tilde{\mathbf{L}}'$$

We can see that this is the Cholesky decomposition. How do you get A_{ij} from $\boldsymbol{\Sigma}$? We know that

$$Y_i=\mathbf{a}_i'\mathbf{Y}_{<i}+\eta_i$$

From this, we know that

$$\mathbb{E}[Y_i|\mathbf{Y}_{<i}]=\mathbf{a}_i'\mathbf{Y}_{<i}$$

From properties of a Normal distribution, we have

$$\mathbb{E}[Y_i|\mathbf{Y}_{<i}]=\mu_{Y_i}+\boldsymbol{\Sigma}_{(i,<i)}\boldsymbol{\Sigma}^{-1}_{(<i,<i)}(Y_{<i}-\mu_{Y_{<i}})=\boldsymbol{\Sigma}_{(i,<i)}\boldsymbol{\Sigma}_{(<i,<i)}Y_{<i}$$

Thus we can see that

$$\mathbb{E}[Y_i|\mathbf{Y}_{<i}]=\mathbf{a}_i'\mathbf{Y}_{<i}=\boldsymbol{\Sigma}_{(i,<i)}\boldsymbol{\Sigma}^{-1}_{(<i,<i)}Y_{<i} \implies \mathbf{a}_i'=\boldsymbol{\Sigma}_{(i,<i)}\boldsymbol{\Sigma}^{-1}_{(<i,<i)}$$

where

$$var\left(\begin{bmatrix}\mathbf{Y}_{<i}\\ \mathbf{Y}_i\end{bmatrix}\right)=\begin{bmatrix}\boldsymbol{\Sigma}_{(<i,<i)} & \boldsymbol{\Sigma}_{(<i,i)}\\ \boldsymbol{\Sigma}_{(i,<i)} & \boldsymbol{\Sigma}_{(i,i)}\end{bmatrix}$$

We also know that

$$var(Y_i|\mathbf{Y}_{<i})=\boldsymbol{\Sigma}_{(i,i)}-\boldsymbol{\Sigma}_{(i,<i)}\boldsymbol{\Sigma}^{-1}_{(<i,<i)}\boldsymbol{\Sigma}_{(<i,i)}=d_i$$

Therefore, we have a way to find $\mathbf{A}$ and $\mathbf{D}$. One important property of this model is that if $a_{ij}=0$, then y_i is conditionally independent of y_j given $y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_{i-1}$. Thus, this gives us a way to induce sparsity from graphs.

Definition 55 *A matrix* $\mathbf{A}$ *is sparse if it has at most* m *non-zero elements in each row.*

Alternatively we can think of this as no node having more than m parents.

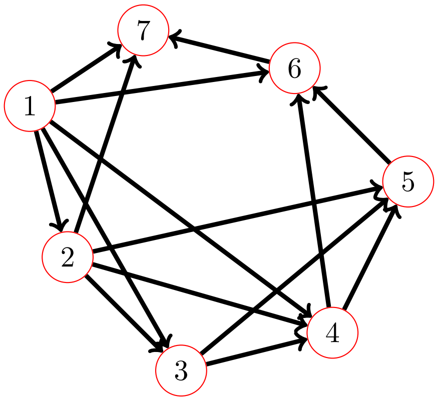

Figure 8.2: A Bayesian network of 7 variables with at most 3 parents (or "neighbors") for each node.

Since $\mathbf{A}$ is sparse, this means that $(\mathbf{I}-\mathbf{A})$ is sparse. However it is important to not that $(\mathbf{I}-\mathbf{A})^{-1}$ need not be sparse. Since $(\mathbf{I}-\mathbf{A})'\mathbf{D}(\mathbf{I}-\mathbf{A})=\boldsymbol{\Sigma}$, we know that $\boldsymbol{\Sigma}^{\mathbf{-1}}$ will be sparse.

8.2 Matrix Normal, Inverse Wishart, and Bayesian Regression

Let $\mathbf{Y} \in \mathbb{R}^{n \times m}$ be a response matrix with m dependent variables and n observations. Let $\mathbf{X} \in \mathbf{n} \times \mathbf{p}$. How would we perform linear regression using this response matrix?
We can vectorize $\mathbf{Y}$ and then perform standard linear regression

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_m \end{bmatrix} \implies \tilde{\mathbf{Y}} = (\mathbf{I} \otimes \mathbf{X})\mathbf{B} + \mathbf{e}$$

How do we model $\mathbf{e}$?
Suppose we modeled it such as:

$$cov(\mathbf{e}_i, \mathbf{e}_j) = u_i j \mathbf{V} \;\; i = 1, \dots, m \;\; j = 1, \dots, m$$

Thus we have

$$cov(\mathbf{e}) = \{cov(\mathbf{e}_i, \mathbf{e}_j)\} = \begin{bmatrix} u_{11}\mathbf{V} & u_{12}\mathbf{V} & \dots & u_{1m}\mathbf{V} \\ u_{21}\mathbf{V} & u_{22}\mathbf{V} & \dots & u_{2m}\mathbf{V} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1}\mathbf{V} & u_{m2}\mathbf{V} & \dots & u_{mm}\mathbf{V} \end{bmatrix} = \mathbf{U} \otimes \mathbf{V}$$

Definition 56 $\mathbf{U} \otimes \mathbf{V}$ *is called the* ***kronecker product****.*

We can explore the properties of the kronecker product by looking at

$$(\mathbf{x}' \otimes \mathbf{A})(\mathbf{y} \otimes \mathbf{B})$$

$$= \begin{bmatrix} x_1\mathbf{A} & x_2\mathbf{A} & \dots & x_n\mathbf{A} \end{bmatrix} \begin{bmatrix} y_1\mathbf{B} \\ y_2\mathbf{B} \\ \vdots \\ y_n\mathbf{B} \end{bmatrix} = \sum_{i=1}^{n} x_i y_i \mathbf{AB} = (\mathbf{x}'\mathbf{y})\mathbf{AB}$$

We can generalize this by looking at

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D})$$

$$= \left(\begin{bmatrix} \mathbf{a}_{1*} \bigotimes \mathbf{B} \\ \vdots \\ \mathbf{a}_{n*} \bigotimes \mathbf{B} \end{bmatrix} \right) \left(\begin{bmatrix} \mathbf{c}_{*1} \otimes \mathbf{D} & \mathbf{c}_{*2} \otimes \mathbf{D} & \dots & \mathbf{c}_{*r} \bigotimes \mathbf{D} \end{bmatrix} \right)$$

We can see that the $(i, j)^{th}$ block is

$$(\mathbf{a}'_{i*} \otimes \mathbf{B})(\mathbf{c}_{*j} \otimes \mathbf{D}) = (\mathbf{a}'_{i*}\mathbf{c}_{*j})\mathbf{BD}$$

Therefore we have that

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$$

Other properties that we have is that

1. $(\mathbf{A} \otimes \mathbf{B}) = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$

2. Cholesky Decomposition $(\mathbf{L}_A\mathbf{L}'_A \otimes \mathbf{L}_B\mathbf{L}'_B) = (\mathbf{L}_A \otimes \mathbf{A}_B)(\mathbf{L}'_A \otimes \mathbf{L}'_B)$

3. $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$

4. $\mathbf{P}_A\mathbf{T}_A\mathbf{P}_A' \otimes \mathbf{P}_B\mathbf{T}_B\mathbf{P}_B' = (\mathbf{P}_A \otimes \mathbf{P}_B)(\mathbf{T}_A \otimes \mathbf{T}_B)(\mathbf{P}_A' \otimes \mathbf{P}_B')$
5. If $\mathbf{A} \in \mathbb{R}^{m\times m}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$, then $det(\mathbf{A}\otimes\mathbf{B}) = det(\mathbf{A})^n det(\mathbf{B})^m$
6. If $\mathbf{A}$ and $\mathbf{B}$ are positive definite, then $\mathbf{A}\otimes\mathbf{B}$ is positive definite.

Back to the model, we have

$$\tilde{\mathbf{Y}} = (\mathbf{I}\otimes\mathbf{X})\mathbf{B} + \mathbf{e} \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{U}\otimes\mathbf{V})$$

Let $\mathbf{z} = \tilde{\mathbf{Y}} - (\mathbf{I}\otimes\mathbf{X})\mathbf{B}$. Thus we have

$$\begin{aligned} p(\mathbf{z}) &\propto \frac{1}{|\mathbf{U}\otimes\mathbf{V}|^{1/2}} exp\left\{-\frac{1}{2}\mathbf{z}'(\mathbf{U}\otimes\mathbf{V})\mathbf{z}\right\} \\ &= \frac{1}{|\mathbf{U}|^{n/2}|\mathbf{V}|^{m/2}} exp\left\{-\frac{1}{2}\mathbf{z}'(\mathbf{U}\otimes\mathbf{V})\mathbf{z}\right\} \end{aligned}$$

Assuming $\mathbf{B} = 0$, we have that

$$p(vec(\mathbf{Y})) = \frac{1}{(2\pi)^{nm/2}|\mathbf{U}|^{n/2}|\mathbf{V}|^{m/2}} exp\left\{-\frac{1}{2}vec(\mathbf{Y})'(\mathbf{U}^{-1}\otimes\mathbf{V}^{-1})vec(\mathbf{Y})\right\}$$

Therefore, we have $vec(\mathbf{Y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{U}\otimes\mathbf{V})$.

Definition 57 *The term $vec(\mathbf{Y})'(\mathbf{U}^{-1}\otimes\mathbf{V}^{-1})vec(\mathbf{Y})$ is known as a **Tensor System**.*

We will digress slightly to look at products of the following form:

$$(\mathbf{A}\otimes\mathbf{B})vec(\mathbf{X}) = vec(\mathbf{C})$$

where $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix}$.

$$(\mathbf{A}\otimes\mathbf{B})vec(\mathbf{X}) = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix} \begin{bmatrix}\mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n\end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n a_{ij}\mathbf{B}\mathbf{x}_j \\ \vdots \\ \sum_{j=1}^n a_{mj}\mathbf{B}\mathbf{x}_j \end{bmatrix}$$

We can see that

$$\sum_{j=1}^n a_{ij}\mathbf{B}\mathbf{x}_j = \mathbf{B}(\sum_{j=1}^n a_{ij}\mathbf{x}_j) = \mathbf{B}\mathbf{X}\mathbf{a}_{i*}$$

Therefore, we have that

$$(\mathbf{A}\otimes\mathbf{B})vec(\mathbf{X}) = \begin{bmatrix} \mathbf{B}\mathbf{X}\mathbf{a}_{1*} \\ \mathbf{B}\mathbf{X}\mathbf{a}_{2*} \\ \vdots \\ \mathbf{B}\mathbf{X}\mathbf{a}_{m*} \end{bmatrix} = vec(\mathbf{B}\mathbf{X}\mathbf{A}')$$

Another Digression: lets now look at

$$vec(\mathbf{Y})'vec(\mathbf{X})$$

where $\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \dots & \mathbf{y}_m\end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_m\end{bmatrix}$. Thus we have that

$$vec(\mathbf{Y})'vec(\mathbf{X}) = \sum_{i=1}^m \mathbf{y}_i'\mathbf{x}_i = \sum_{i=1}^m tr(\mathbf{y}_i'\mathbf{x}_i) = \sum_{i=1}^m tr(\mathbf{x}_i\mathbf{y}_i') = tr(\sum_{i=1}^m \mathbf{x}_i\mathbf{y}_i') = tr(\mathbf{X}\mathbf{Y}')$$

Going back to the multivariate normal model, we have that

$$vec(\mathbf{Y})'(\mathbf{U}^{-1}\otimes\mathbf{V}^{-1})vec(\mathbf{Y}) = vec(\mathbf{Y})'vec(\mathbf{V}^{-1}\mathbf{Y}\mathbf{U}^{-1}) = tr(\mathbf{V}^{-1}\mathbf{Y}\mathbf{U}^{-1}\mathbf{Y}') = tr(\mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y}\mathbf{U}^{-1})$$

Definition 58 *The* ***Matrix-Variate Normal*** *is a random matrix* $\mathbf{Y} \in \mathbb{R}^{n\times m}$*, denoted* $\mathbf{Y} \sim MN(\mathbf{M}, \mathbf{V}, \mathbf{U})$*, such that*

$$p(\mathbf{Y}) = \frac{1}{(2\pi)^{nm/2}|\mathbf{U}|^{n/2}|\mathbf{V}|^{m/2}} exp\left\{-\frac{1}{2}tr((\mathbf{Y}-\mathbf{M})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{M})\mathbf{U}^{-1})\right\}$$

From this, we can see that if $\mathbf{Y} \sim MN(\mathbf{M}, \mathbf{V}, \mathbf{U}) \iff vec(\mathbf{Y}) \sim \mathcal{N}(vec(\mathbf{M}), \mathbf{U}\otimes\mathbf{V})$.

We can now discuss the Bayesian conjugate Matrix Normal-Inverse Wishart model. Suppose we have the following setup

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times m}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times m}$, and $\mathbf{E} \in \mathbb{R}^{n\times m}$. Let $\mathbf{E} \sim MN(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma})$. Lets take a look at the LS-Estimate of this model. We have that

$$\mathbf{X}\alpha \perp (\mathbf{Y}-\mathbf{XB})\forall\alpha \implies \langle \mathbf{Y}-\mathbf{XB}, \mathbf{X}\alpha\rangle = 0 \;\; \forall\alpha$$

$$\implies \alpha\langle \mathbf{Y}-\mathbf{XB}, \mathbf{X}\alpha\rangle = \langle (\mathbf{Y}-\mathbf{XB})\alpha, \mathbf{X}\alpha\rangle = 0 \;\; \forall\alpha$$

$$\implies \mathbf{X}'(\mathbf{Y}-\mathbf{XB}) = 0$$

$$\implies \mathbf{X}'\mathbf{XB} = \mathbf{X}'\mathbf{Y}$$

Thus we have $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Definition 59 *The* ***Inverse-Wishart Distribution*** *is a probability distribution valid over the cone of positive definite matrices. If* $\mathbf{\Sigma}$ *has a Inverse-Wishart distribution, denoted* $\mathbf{\Sigma} \sim IW(\nu, \mathbf{S})$*, then it has the following pdf:*

$$\frac{\mathbf{S}^{\nu/2}}{2^{\nu p/2}\Gamma_p(\nu/2)}|\mathbf{\Sigma}|^{-(\nu+p+1)/2} exp\left\{-\frac{1}{2}tr(\mathbf{S}\mathbf{\Sigma}^{-1})\right\}$$

where $\mathbf{S}, \mathbf{\Sigma} \in \mathbb{R}^{p\times p}$.

Definition 60 *We say that* $(\mathbf{B}, \mathbf{\Sigma})$ *is distributed* ***Matrix-Normal Inverse-Wishart*** *, denoted* $(\mathbf{B}, \mathbf{\Sigma}) \sim MNIW(\mathbf{B}, \mathbf{\Sigma}|\mathbf{M}, \mathbf{V}, \nu, \mathbf{S})$*, if*

$$p(\mathbf{B}, \mathbf{\Sigma}) = IW(\mathbf{\Sigma}|\nu, \mathbf{S}) \times MN(\mathbf{B}|\mathbf{M}, \mathbf{V}, \mathbf{\Sigma})$$

Consider the following Bayesian Model

$$\mathbf{Y}|\mathbf{B}, \mathbf{\Sigma} \sim MN(\mathbf{XB}, \mathbf{I}_n, \mathbf{\Sigma})$$

$$\mathbf{B}|\mathbf{\Sigma} \sim MN(\mathbf{C}, \mathbf{V}, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} \sim IW(\nu, \mathbf{S})$$

We can see that $\mathbf{B}, \mathbf{\Sigma}$ is distributed Matrix-Normal Inverse-Wishart. Suppose that we wish to draw posterior samples from $(\mathbf{B}, \mathbf{\Sigma}|\mathbf{Y})$. We know that

$$p(\mathbf{B}, \mathbf{\Sigma}|\mathbf{Y}) \propto p(\mathbf{\Sigma})p(\mathbf{B}|\mathbf{\Sigma})p(\mathbf{Y}|\mathbf{B}, \mathbf{\Sigma})$$

$$\begin{aligned}\propto |\mathbf{\Sigma}|^{-(\nu+m+1)/2} exp\left\{-\frac{1}{2}tr(\mathbf{S}\mathbf{\Sigma}^{-1})\right\} |\mathbf{\Sigma}|^{p/2} exp\left\{-\frac{1}{2}tr((\mathbf{B}-\mathbf{C})'\mathbf{V}^{-1}(\mathbf{B}-\mathbf{C})\mathbf{\Sigma}^{-1})\right\} \\ \times |\mathbf{\Sigma}|^{n/2} exp\left\{-\frac{1}{2}tr((\mathbf{Y}-\mathbf{XB})'(\mathbf{Y}-\mathbf{XB})\mathbf{\Sigma}^{-1})\right\} \quad (8.1)\end{aligned}$$

$$= |\mathbf{\Sigma}|^{-(\nu+m+1+n+p)/2} exp\left\{-\frac{1}{2}tr((\mathbf{S}+\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}+\mathbf{Y}'\mathbf{Y})\mathbf{\Sigma}^{-1})\right\}$$
$$\times\ exp\left\{-\frac{1}{2}tr((\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}+\mathbf{B}'\mathbf{X}'\mathbf{X}\mathbf{B}-(\mathbf{Y}'\mathbf{X}+\mathbf{C}'\mathbf{V}^{-1})\mathbf{B})\mathbf{\Sigma}^{-1})\right\} \quad (8.2)$$

$$= |\mathbf{\Sigma}|^{-(\nu+m+1+n+p)/2} exp\left\{-\frac{1}{2}tr((\mathbf{S}+\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}+\mathbf{Y}'\mathbf{Y})\mathbf{\Sigma}^{-1})\right\}$$
$$\times\ exp\left\{-\frac{1}{2}tr((\mathbf{B}'(\mathbf{V}^{-1}+\mathbf{X}\mathbf{X}')\mathbf{B}-(\mathbf{Y}'\mathbf{X}+\mathbf{C}'\mathbf{V}^{-1})\mathbf{B})\mathbf{\Sigma}^{-1})\right\} \quad (8.3)$$

let $\mathbf{M} = (\mathbf{V}^{-1}+\mathbf{X}\mathbf{X}')^{-1}$ and $\mathbf{m} = \mathbf{X}'\mathbf{Y}+\mathbf{V}^{-1}\mathbf{C}$. Thus we have

$$= |\mathbf{\Sigma}|^{-(\nu+m+1+n+p)/2} exp\left\{-\frac{1}{2}tr((\mathbf{S}+\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}+\mathbf{Y}'\mathbf{Y})\mathbf{\Sigma}^{-1})\right\}$$
$$\times\ exp\left\{-\frac{1}{2}tr((\mathbf{B}'\mathbf{M}^{-1}\mathbf{B}-\mathbf{m}'\mathbf{B})\mathbf{\Sigma}^{-1})\right\} \quad (8.4)$$

$$= |\mathbf{\Sigma}|^{-(\nu+m+1+n)/2} exp\left\{-\frac{1}{2}tr((\mathbf{S}+\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}+\mathbf{Y}'\mathbf{Y}-\mathbf{m}'\mathbf{M}\mathbf{m})\mathbf{\Sigma}^{-1})\right\}$$
$$\times\ |\mathbf{\Sigma}|^{-p/2} exp\left\{-\frac{1}{2}tr(((\mathbf{B}-\mathbf{m})'\mathbf{M}^{-1}(\mathbf{B}-\mathbf{m})\mathbf{\Sigma}^{-1})\right\} \quad (8.5)$$

Thus we can see that $\mathbf{\Sigma} \sim IW(\nu+n, \mathbf{S}+\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}+\mathbf{Y}'\mathbf{Y}-\mathbf{m}'\mathbf{M}\mathbf{m})$ and $\mathbf{B}|\mathbf{\Sigma} \sim MN(\mathbf{M}\mathbf{m}, \mathbf{M}, \mathbf{\Sigma})$.