Linear Models

STAT 551

Course Notes by Meridith Bartley

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

First printing, March 2013

Contents

Part One

1. Linear Regression

- projection
- orthongonal decomposition
- Gaussian Linear Regression
- prediction (generally of $\hat{y}$)
- different types of errors
- influence
- lack of fit
- R^2
- Multicollinearity

1.1 Projection in Euclidean Space

Monday August 22

Definition 1.1.1 — Euclidian Space. One way to think of the Euclidean plane is as a set of points satisfying certain relationships, expressible in terms of distance and angle. **Euclidean space** is an abstraction detached from actual physical locations, specific reference frames, measurement instruments, and so on.

Let Euclidian Space be denoted by $\mathbb{R}^P$.

$$\mathbb{R}X \ldots X\mathbb{R} = \{(x_1, \ldots, x_p) : x_1 \in \mathbb{R} \ldots, x_p \in \mathbb{R}^P\}$$

Definition 1.1.2 — Inner Product. In linear algebra, an inner product space is a vector space with an additional structure called an inner product. This additional structure associates each pair of vectors in the space with a scalar quantity known as the inner product of the vectors. **Inner products** allow the rigorous introduction of intuitive geometrical notions such as the length of a vector or the angle between two vectors. They also provide the means of defining orthogonality between vectors (zero inner product).

Let $a \in \mathbb{R}^P, b \in \mathbb{R}^P$

$$a^T b = \sum_{i=1}^{P} a_i b_i$$

$$a^T b = < a, b >$$

Definition 1.1.3 — Hilbert Space. The mathematical concept of a Hilbert space generalizes the notion of Euclidean space. It extends the methods of vector algebra and calculus from the two-dimensional Euclidean plane and three-dimensional space to spaces with any finite or infinite number of dimensions. A Hilbert space is an abstract vector space possessing the structure of an inner product that allows length and angle to be measured. Furthermore, Hilbert spaces are complete: there are enough limits in the space to allow the techniques of calculus to be used.

Hilbert Inner Product Space $\{\mathbb{R}^P, < a, b >\}$

General Inner Product

Let $\Sigma \in \mathbb{R}^{PxP}$ set of all PxP matrices. Assume Σ is a positive definite matrix.

$$x^T \Sigma x < 0$$
$$\forall x \in \mathbb{R}^P$$
$$x \neq 0$$

Then $a^T \Sigma b$ also satisfies the conditions for inner product.

$$a^T \Sigma b = < a, b >_\Sigma$$

$$a^T b = a^T I b = < a, b >_I$$

$\{\mathbb{R}^P, <,>_\Sigma\}$ is a more general inner product space.

Linear Transformation

A matrix, $A, \in \mathbb{R}^{PxP}$ can be viewed as linear transformation
$T_A : \mathbb{R}^P \to \mathbb{R}^P, x \mapsto Ax$

R Bing Li will denote T_A as A.
$\to$ means maps to for a domain.
$\mapsto$ means maps to for a value.
$\Rightarrow$ means implies.

If $A : \mathbb{R}^P \to \mathbb{R}^P$,

$$ker(A) = \{x \in \mathbb{R}^P, Ax = 0\}$$
$$ran(A) = \{Ax : x \in \mathbb{R}^P\}$$

Definition 1.1.4 — Kernel. In linear algebra, the kernel, or sometimes the null space, is the set of all elements v of V for which L(v) = 0, where 0 denotes the zero vector in W.

In coordinate plane, think of a function that crosses the x-axis. The kernel would be all points on x where $y = 0$.

Definition 1.1.5 — Range. In coordinate plane, how much of the y axis is reached with the function? Now extend this idea to more dimensions.

A linear transformation is **idempotent** if

$$A = A^2$$

$$Ax = A(A(x))$$

$$\forall x \in \mathbb{R}^P$$

If A were a number it could only be 1 or 0.

Wednesday August 24

Let $T \in \mathbb{R}^{PxP}$ then there exists a unique operator $R \in \mathbb{R}^{PxP}$ such that $\forall x, y \in \mathbb{R}^P$,

$$< x, Ty > = < Rx, y >$$

(general inner product, $a^T \Sigma b$). Aside: What this states is that if you give me any operator in the first you can find one in the second.

R is called the **adjoint operator** of T. Written as T^*, that is,

$$< x, Ty > = < T^* x, y >$$

Derived Facts

$$\begin{aligned} < x, Ty > &= < T^*, y > && \text{(by the definition)} \\ &= < y, T^* x > && \text{(inner products the order doesn't matter)} \\ &= < (T^*)^* y, x > && \text{(Use the definition again)} \\ &= < x, (T^*)^* y > && \text{(swap order)} \end{aligned}$$

So, $T = (T^*)^*$.

It is easy to see in our case

$$\begin{aligned} < x, Ty >_\Sigma &= x^T \Sigma T y \\ &= x^T \Sigma T \Sigma^{-1} \Sigma y \\ &= (\Sigma^{-1} T^T \Sigma x)^T \Sigma y \\ &= < \Sigma^{-1} T^T \Sigma x, y >_\Sigma \end{aligned}$$

So, $T^* = \Sigma^{-1}T^T\Sigma$ when $\Sigma = I_P$ (identity) and $T^* = T^T$.

Derived Facts

An operator is **self adjoint** if its adjoint is itself. (i.e. if $T = T^*$ or $< x, Ty >=< Tx, y >$). In the case of $<,>_\Sigma$,

$$T = \Sigma^{-1}T^T\Sigma$$

if

$$\Sigma = I_P,\ T = T^T$$

R Self adjoint implies symmetric. It's a more general case, hence the use of Σ vs I. Useful to remember in following two Theorems

Theorem 1.1.1 If $\boldsymbol{A} \in \mathbb{R}^{PxP}$ is symmetric, then there exists **eigenvalue-eigenvector pairs**. $(\lambda_1, v_1), \ldots (\lambda_P, v_P)$ such that $v_1 \perp \ldots \perp v_P$. Orthoginal basis (ONB) such that

$$\boldsymbol{A} = \sum_{i=1}^{P} \lambda_i v_i v_i^T \text{(spectral decomposition)}$$

More generally, if $\boldsymbol{A}$ is a linear operator in $\mathscr{H}$ (finite dimential inner product such as $(\mathbb{R}^P, <,>_\Sigma)$). its eigen pair (linear operator now) (λ, v) is defined by

$$\begin{cases} \boldsymbol{A}\underline{v} = \underline{\lambda v} \\ < \underline{v}, \underline{v} >= 1 \end{cases}$$

Definition 1.1.6 — Orthogonal Basis. In the following, $(\mathbb{R}^P, <,>_\Sigma) = \mathscr{H}$ (H for Hilbert)

ONB is defined by:

1. $v_i \perp v_j, < v_i, v_j >= 0$
2. $||v_i|| = 1$
3. $\text{span}\{v_1, \ldots, v_P\} = \mathscr{H}$

Theorem 1.1.2 Suppose $\boldsymbol{A} : \mathscr{H} \to \mathscr{H}$ is a self adjoint linear operator. Then $\boldsymbol{A}$ has eigen pairs: $(\lambda_1, v_1, \ldots, (\lambda_P, v_P)$ where $\{v_1, \ldots, v_P\}$ is ONB of $\mathbb{R}$ such that

$$\boldsymbol{A} = \sum_{i=1}^{P} \lambda_i v_i v_i^T \Sigma$$

Proof. (λ, v) is eigen pair of $\boldsymbol{A}$, which means

$$\boldsymbol{A}v = \lambda v$$

$$< v, v >= 1$$

$$v^T\Sigma v = 1$$

Let $u = \Sigma^{\frac{1}{2}}v$.

R Aside: $\Sigma^\alpha = \Sigma\lambda_i^\alpha v_i v_i^T$

Let $v = \Sigma^{-\frac{1}{2}} u$.

$$\begin{aligned} \boldsymbol{A}\Sigma^{-\frac{1}{2}} u &= \lambda \Sigma^{-\frac{1}{2}} u \\ \Sigma^{-\frac{1}{2}} u &= \lambda u \end{aligned}$$

So, (λ, v) is an eigen pair of $\boldsymbol{A}$ in $(\mathbb{R}, <,>_\Sigma) \Leftrightarrow (\lambda, u)$ '...' of $\Sigma^{\frac{1}{2}} \boldsymbol{A} \Sigma^{-\frac{1}{2}}$ in $(\mathbb{R}, <,>_I)$.
Note that, $\boldsymbol{A}$ is self adjoint in $(\mathbb{R}, <,>_\Sigma)$. So, $\boldsymbol{A} = \Sigma^{-1} \boldsymbol{A}^T \Sigma$

$$\begin{aligned} \Sigma^{\frac{1}{2}} \boldsymbol{A} \Sigma^{-\frac{1}{2}} &= \Sigma^{\frac{1}{2}} \boldsymbol{A}^T \Sigma \Sigma^{-\frac{1}{2}} \\ &= \Sigma^{-\frac{1}{2}} \boldsymbol{A}^T \Sigma^{\frac{1}{2}} \\ &= (\Sigma^{\frac{1}{2}} \boldsymbol{A} \Sigma^{-\frac{1}{2}})^T \end{aligned}$$

Note: $\Sigma^{\frac{1}{2}} \boldsymbol{A} \Sigma^{-\frac{1}{2}}$ is symmetric!! So by Theorem 1.1, $\Sigma^{\frac{1}{2}} A \Sigma^{-\frac{1}{2}} = \sum \lambda_i v_i v_i^T$ where (λ_i, v_i) eigen-pairs of $\Sigma^{\frac{1}{2}} \boldsymbol{A} \Sigma^{-\frac{1}{2}}$.

That means $(\lambda_i, \Sigma^{\frac{1}{2}} v_i)$ are eigen pairs of A.

So, $\Sigma^{\frac{1}{2}} \boldsymbol{A} \Sigma^{-\frac{1}{2}} = \sum_{i=1}^{P} \Sigma^{\frac{1}{2}} u_i u_i^T \Sigma^{\frac{1}{2}} \Rightarrow A = \sum_{i=1}^{P} \lambda u_i u_i^T \Sigma$ ■

Definition 1.1.7 — Projection. If P is an operator in $(\mathbb{R}^P, <,>)$ then P is called a **projection** if it is both idempotent ($P = P^2$) and self adjoint ($P = P^*$).

Preposition 1.1 If $\boldsymbol{A}$ is a linear operator then $ker(\boldsymbol{A}) = ran(\boldsymbol{A}^*)^\perp$

Proof. Take $x \in ker(\boldsymbol{A}) (\Rightarrow \boldsymbol{A}x = 0)$.
$\forall y \in ran(\boldsymbol{A}^*), x \perp y$
$\Rightarrow x \perp y \forall y = \boldsymbol{A}^* z, z \in \mathbb{R}^P$
Hence,

$$\begin{aligned} < x, y > &= < x, \boldsymbol{A}^* z > \\ &= < \boldsymbol{A}x, z > \\ &= < 0, z > \\ &= 0 \end{aligned}$$

$$\begin{aligned} &\Rightarrow x \perp y \\ &\Rightarrow x \in ran(\boldsymbol{A}^*)^\perp \end{aligned}$$

Or vice versa. ■

Friday August 26

R $\perp$ means orthogonal complement.

$$\mathscr{S}^\perp = \{ v \in \mathbb{R}^P, v \perp \mathscr{S} \}$$

$$v \perp w \forall w \in \mathscr{S}$$

$$< v, w >= 0 \forall w \in \mathscr{S}$$

$$= \{v \in \mathbb{R}^P, < v, w >= 0 \forall w \in \mathscr{S}\}$$

Recall, $ker(A) = ran(A^*)^{\perp}$

So, if A is self adjoint then this is true and $ran(A)$ is also $span(A)$ which is the subspace spanned all columns of A.

Theorem 1.1.3 If P is a projection, then
1. $Pv = v$, $\forall v \in ran(P)$
2. $Pv = 0$, $\forall v \perp ran(P)$
3. If Q is another projections such that the $ran(Q) = ran(P)$ then $Q = P$. (The range determines the operator, because it is what decomposes the operator.)

Asside: P acts like one on some spaces, and zero on orthogonal space.

Proof. 1. Let $v \in ran(P)$. Since $P^2 = P$ (idempotent) then

$$P^2 v = Pv$$
$$\Rightarrow P^2 v - PV = 0$$
$$\Rightarrow P(Pv - v) = 0$$
$$\Rightarrow Pv - v \in ker(P)$$
$$\Rightarrow Pv - v \perp ran(P)$$
$$\Rightarrow < Pv - v, Pv - v >= 0$$
$$\Rightarrow ||Pv - v|| = 0$$
$$\Rightarrow Pv - v = 0$$
$$\Rightarrow Pv = v$$

2. If

$$v \perp ran(P)$$
$$\Rightarrow v \in ker(P)$$
$$\Rightarrow Pv = 0$$

3. If Q is another operator with $ran(Q) = ran(P) = \mathscr{S}$ then $\forall v \in \mathscr{S}$

$$Qv = v = Pf(\forall v \perp \mathscr{S})$$
$$Qv = 0 = Pv$$
$$Qv = Pv\forall,\ v \in \mathscr{S}$$
$$Q = P$$

■

Theorem 1.1.4 Suppose $\mathscr{S}$ is a subspace of $\mathbb{R}^P$, R $V_1, \ldots, V_m$ is a basis of $\mathscr{S}$.

Let $V = (V_1, \ldots, V_m) \in \mathbb{R}^{xM}$.

Then,
1. $A = V(V^T \Sigma V)^{-1} V^T \Sigma$ is a projection.
2. $ran(A) = \mathscr{S}$

Proof. 1. idempotent.

$$A^2 = V(V^T \Sigma V)^{-1} V^t \Sigma V (V^T \Sigma V)^{-1} V^T \Sigma$$
$$= V(V^T \Sigma V)^{-1} V^T \Sigma$$
$$= A$$

2. Self adjoint.
   Let $x, y \in \mathbb{R}^P$
   $< x, Ay > = x^T \Sigma v (v^T \Sigma v)^{-1} v^{\Sigma} y$
   $= (v(v^T \Sigma v)^{-1} v^T \Sigma x)^T \Sigma y$
   $=< Ax, y >$
3. $ran(A) = \mathscr{S}$?
   Let $x \in \mathbb{R}^P$.
   $Ax = v(v^T \Sigma v)^{-1} v^T \Sigma x \in span(v) = \mathscr{S}$
   So let $x \in \mathscr{S}$,
$$x \in ran(v)$$
$$x = vy$$
   for some $y \in \mathbb{R}^P$
$$= v(v^T \Sigma v)^{-1} v^T \Sigma v y$$
   $\in ran(A)$
   So, $\mathscr{S} \subseteq ran(A)$ and then $\mathscr{S} = ran(A)$.

■

We write A as $P_{\mathscr{S}}(\Sigma)$ (orthogonal projection on to $\mathscr{S}$ with respect to Σ - product).

In the following, let $I : \mathbb{R}^P \to \mathbb{R}^P$ be the identity mapping. ($x \mapsto x$)
Let $\mathscr{S}$ be a subspace in $\mathbb{R}^P$.
Let $Q_{\mathscr{S}}(\Sigma) = I - P_{\mathscr{S}}(\Sigma)$

Proprosition 1.2 $Q_{\mathscr{S}}(\Sigma) = P_{\mathscr{S}^{\perp}}(\Sigma)$

Proof. Show $Q_{\mathscr{S}}(\Sigma)$ is projection.

1. *Idempotent*
   $Q^2_{\mathscr{S}}(\Sigma) = Q_{\mathscr{S}}(\Sigma) Q_{\mathscr{S}}(\Sigma)$
   $= (I - P_{\mathscr{S}}(\Sigma))(I - P_{\mathscr{S}}(\Sigma))$
   $= I - P_{\mathscr{S}}(\Sigma) - P_{\mathscr{S}}(\Sigma) + P_{\mathscr{S}} P_{\mathscr{S}}$
   $= Q_{\mathscr{S}}(\Sigma)$
2. *Self-adjoint*
$$x, y \in \mathbb{R}^P$$
   $< x, Q_{\mathscr{S}}(\Sigma) y > =< x, (I - P_{\mathscr{S}}(\Sigma)) y >$
   $=< x, y > - < x, P_{\mathscr{S}}(\Sigma)) y >$
   $=< x, y > - < P_{\mathscr{S}}(\Sigma)) x, y >$
   $=< (I - P_{\mathscr{S}}(\Sigma)) x, y >$
   $=< Q_{\mathscr{S}}(\Sigma) x, y >$
3. *Range*
   $ran(Q_{\mathscr{S}}(\Sigma)) = \mathscr{S}^{\perp}$. Take $x \perp \mathscr{S} = ran(P_{\mathscr{S}}(\Sigma))^{\perp} = ker(P_{\mathscr{S}}(\Sigma))$.

$$\Rightarrow P_{\mathscr{S}}(\Sigma) = 0$$
$$\Rightarrow Q_{\mathscr{S}}(\Sigma)x = x - P_{\mathscr{S}}(\Sigma)x = x$$
$$X \in ran(Q_{\mathscr{S}}(\Sigma))$$
$$\Rightarrow \mathscr{S}^{\perp} \subseteq ran(Q_{\mathscr{S}}(\Sigma))$$
$$\text{Take } x \in ran(Q_{\mathscr{S}}(\Sigma)),\ \forall y \in \mathscr{S} = ran(P_{\mathscr{S}}(\Sigma))$$
$$y = P_{\mathscr{S}}(\Sigma)z \text{ for some } z \in \mathbb{R}^P$$
$$< x, y >=< x, P_{\mathscr{S}}(\Sigma)z >=< P_{\mathscr{S}}(\Sigma)x, z >= 0$$
$$\Rightarrow x \in \mathscr{S}^{\perp}$$
$$\Rightarrow ran(Q_{\mathscr{S}}(\Sigma)) = \mathscr{S}^{\perp}$$

■

1.2 Cochran's Theorem

This section will be about the distribution of the squared norm of a projection of a Gaussian random vector.

Preposition 1.3 If A is idempotent, then its eigenvalues are either 0 or 1.

Proof. λ is eigenvalue of A.

$$\Rightarrow Av = \lambda v(||v|| = 1)$$

$$\lambda = Av = A^2 v = \lambda Av = \lambda^2$$

So, λ is 0 or 1. ■

Monday August 29

Lemma 1.1 Suppose $V \sim N(0, \sigma^2 I_P)$.
P is projection with I_P- inner product. Then $V^T PV \sim \sigma^2 \chi^2_S$ where df = rank(P).

Proof. P is symmetric, and it has spectral decomposisition,

$$ARA^T$$

where the A's are orthogonal and R is diagonal with diagonal entries 0 or 1.

Then,

$$A^T V \sim N_P(0, A^T(\sigma^2 I_P)A) = N_P(0, \sigma I_P)$$

Let,

$$Z = RA^T V$$

then,

$$Z \sim N_P(0, \sigma^2 R^2) = N_P(0, \sigma^2 R)$$

That means among the components of Z, some are distributied as N(0, 1) and the rest are zero and they are independant. So,

$$Z^T Z \sim \chi^2_S = V^T PV$$

■

Corollary 1.2.1 Suppose $X \sim N(0, \Sigma)$. Consider the Hilbert space $(\mathbb{R}^P, <, >_{\Sigma^{-1}})$.

$$< a, b >_{\Sigma^{-1}} = a^T \Sigma^{-1} b$$

Let $\mathscr{S}$ be a subspace of $\mathbb{R}^P$ and $P_{\mathscr{S}}(\sigma^{-1})$ be the projection onto $\mathscr{S}$ with respect to $<, >_{\Sigma}^{-1}$ (special case of Fisher information inner product)

Then,

$$||P_{\mathscr{S}}(\Sigma^{-1})x||^2_{\Sigma^{-1}} \sim \chi^2_r$$

where $r = dim(\mathscr{S})$.

Proof. Let V be a basis matrix of $\mathscr{S}$ (i.e. the col of V form basis in $\mathscr{S}$).

$$\begin{aligned}
||P_{\mathscr{S}}(\Sigma^{-1})X||^2_{\Sigma^{-1}} &= < P_{\mathscr{S}}(\Sigma^{-1})X, P_{\mathscr{S}}(\Sigma^{-1})X > \\
&= X^T P_{\mathscr{S}}(\Sigma^{-1})\Sigma^{-1} P_{\mathscr{S}}(\Sigma^{-1})X \\
&= X^T (V(V^T\Sigma^{-1}V)^{-1}V^T\Sigma^{-1})^T\Sigma^{-1}(V(V^T\Sigma^{-1}V)^{-1}V^T\Sigma^{-1})X \\
&= X^T\Sigma^{-1}V(V^T\Sigma^{-1}V)^{-1}v^T\Sigma^{-1}V(V^T\Sigma^{-1}V)^{-1}V^T\Sigma^{-1})X \\
&= (\Sigma^{-\frac{1}{2}}X)^T[\Sigma^{-\frac{1}{2}}V(V^T\Sigma^{-1}V)^{-1}(\Sigma^{-\frac{1}{2}}V)^T](\Sigma^{-\frac{1}{2}}X)
\end{aligned}$$

But,

$$\Sigma^{-\frac{1}{2}}x \sim N(0, I_P)$$

So,

$$\Sigma^{-\frac{1}{2}}V(V^T\Sigma^{-1}V)^{-1}(V^T\Sigma^{-\frac{1}{2}})^T \quad (*)$$

is a projection with repect to I_P-inner producted (idempotent, self adjoint, YES).

By Lemme 1.1,

$$(*) \sim \chi^2_r$$

.

■

It is then easy to derive Cocharan's Theorem. (see proof in Homework 1)

Theorem 1.2.2 Let $X \sim N(0, \Sigma)$ and $\mathscr{H} = \{\mathbb{R}^P, <, >_{\Sigma^{-1}}\}$. Let $\mathscr{S}_1, dots, \mathscr{S}_k$ be linear subspaces of $\mathbb{R}^P$ such that $\mathscr{S}_i \perp \mathscr{S}_j$ in $<, >_{\Sigma^{-1}}$

Let $r_i = dim(\mathscr{S}_i)$.

Let $W_i = ||P_{\mathscr{S}_i}(\Sigma^{-1})X||^2_{\Sigma-1}$

Then,

1. $W_i \sim \chi^2_{r_i}$
2. $W_1 \perp\!\!\!\perp, \dots, \perp\!\!\!\perp W_k$ where $\perp\!\!\!\perp$ indicates independence.

1.3 Gaussian Linear Regresson Model

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$X = \begin{pmatrix} x_{11} & \dots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{nxp}$$

Consider the linear model,

$$y = X\beta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

where X has full column rank ($n \geq p$).

Here X is treated as fixed.

Maximum Likelihood Estimator

$$\mathrm{E}(y) = X\beta \in \mathbb{R}^n$$

$$\mathrm{Var}(y) = \sigma^2 I_n$$

$$y \sim \mathrm{N}_p(X\beta, \sigma^2 I_n)$$

Multivariate Normal Density

$$y \sim \mathrm{N}(\mu, \Sigma)$$

$$f_Y(y) = \frac{1}{(2\pi)^{\frac{n}{2}}[det(\Sigma)]^{\frac{1}{2}}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)}$$

In our case,

$$\Sigma = \sigma I_n$$

$$\det(\Sigma) = \det(\sigma^2 I_n) = \sigma^2 \det(I_n) = \sigma^{2n}$$

So,

$$f_Y(y) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\sigma^{2n}}} e^{-\frac{1}{2\sigma^2}||y-\mu||^2}$$

To find the log likelihood and subsequently take the partial derivatives for MLE,

$\log(f_y(\eta)) = \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}||y-\mu||^2 = \ell(\beta, \sigma^2, y)$

$\frac{\partial}{\partial\beta} = \cdots = -\frac{1}{2\sigma^2}2X^T(y - X\beta) = 0$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \in \mathbb{R}^P$$

$\frac{\partial}{\partial\sigma^2} l(\beta, \sigma^2, y) = \cdots = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}||y - X\beta||^2 = 0$

$$\hat{\sigma^2} = \frac{1}{n}||y - X\hat{\beta}||^2$$

In summary, the MLE for (β, σ^2) in Gaussian Linear Model are

$$\hat{\beta} = (X^T x)^{-1} X^T Y$$
$$\hat{\sigma^2} = \frac{1}{n} ||y - X\hat{\beta}||^2$$

Note that

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

So,

$$\hat{y} = P_{\text{span}(x)}(I_P) = P_X y$$

Now,

$$\begin{aligned}
\hat{\sigma^2} &= \frac{1}{n} ||y - \hat{y}||^2 \\
&= \frac{1}{n} ||y - P_X y||^2 \\
&= \frac{1}{n} ||(I_n - P_X) y||^2 \\
&= \frac{1}{n} ||Q_X y||^2
\end{aligned}$$

where $Q_X = (I_n - P_X)$ is projection on to $\text{span}(X)^\perp$.

It turns out that $(X^T y, y^T y)$ is complete, sufficient statistic for this Gaussian linear model (see homework).

Wednesday August 31

Recall,

$$\begin{aligned}
&\hat{\beta} = (X^T x)^{-1} X^T Y \\
&\hat{\sigma^2} = \frac{1}{n} ||y - X\hat{\beta}||^2 \\
&Q_x = I_n - P_x \\
&P_X + X(X^T X)^{-1} X^T
\end{aligned}$$

Several properties,

$$\text{E}(\hat{\beta}) = \beta \quad \text{(unbiased)}$$

$$\text{Var}(\hat{\beta}) = (X^T X)^{-1} X^T (\sigma^2 I_n) X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

Thus,

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

Because P_x has rank p and Q_x has rank $(n-p)$, then

$$||Q_x y||^2 \sim \chi^2_{(n-p)}$$

Let's find an unbiased estimator for σ^2 (needed for UMVUE),

$$\begin{aligned} E(\hat{\sigma^2}) &= E(\frac{1}{n}||Q_x y||^2) \\ &= \frac{n-p}{n}\sigma^2 \\ E\left(\frac{n}{n-p}\hat{\sigma^2}\right) &= \tilde{\sigma}^2 \end{aligned}$$

Moreover, $\hat{\beta}$ has one-to-one transformation with

$$(X^T X)^{-1}X^T y \leftrightarrow X(X^T X)^{-1}X^T y = P_X y$$

$$\begin{aligned} Cov(P_X y, Q_X y) &= P_X \sigma^2 I_n Q_X \\ &= \sigma^2 P_X Q_X \\ &= 0 \end{aligned}$$

$$P_X y \perp\!\!\!\perp Q_X y \quad \text{(due to normality)}$$

$\hat{\beta} \leftrightarrow P_X y$
$\hat{\sigma^2}$ is a funciton of $Q_X y$, so $\hat{\beta} \perp\!\!\!\perp \hat{\sigma^2}$

In your homework, $\hat{\beta}, \hat{\sigma^2} \leftrightarrow$ complete sufficient.

$\hat{\beta}, \tilde{\sigma^2}$ is UMVUE (Lehmann-Sheffe).

Theorem 1.3.1 — Gaussian Regression Model. Under this model:

1. $\hat{\beta}, \tilde{\sigma^2}$ UMVUE for β, σ^2
2. $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$
3. $(n-p)\tilde{\sigma^2} \sim \sigma^2 \chi^2_{(n-p)}$
4. $\hat{\beta} \perp\!\!\!\perp \tilde{\sigma^2}$

1.4 Statistical Inference for β, σ^2

Suppose we want to test
$H_0 : \beta_1 = \beta_{i0}$
Let $M = (X^T X)^{-1}$.

Then,

$$\hat{\beta} \sim N(\beta_i 0, \sigma^2 M_{ii})$$

where, $M_{ii} \leftarrow (i,i)^{th}$ entry of M

Also, $\frac{(n-p)\tilde{\sigma^2}}{\sigma^2} \sim \chi^2_{(n-p)}$

$$\hat{\beta} \perp\!\!\!\perp \tilde{\sigma^2}$$

$$\frac{\frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{\sigma^2 M_{ii}}} \sim N(0,1)}{\sqrt{\frac{(n-p)\tilde{\sigma}^2/\sigma^2 \cap_{k=n}^{\infty} A_k^C)}{n-p}}} \sim t_{(n-p)}$$

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{\tilde{\sigma}\sqrt{M_{ii}}} \sim t_{(n-p)} = (*)$$

Reject H_0 if

$$\left| \frac{\hat{\beta}_i - \beta_{i0}}{\tilde{\sigma}\sqrt{M_{ii}}} \right| > t_{\frac{\alpha}{2}(n-p)}$$

Recall,

$$X \sim N(\mu, 1)$$

$$y \sim \chi^2_r$$

$$X \perp\!\!\!\perp y$$

$$\frac{X}{\sqrt{\frac{y}{r}}} \sim t_n(\mu)$$

Power at β_{i1}

$$\hat{\beta}_i \sim N(\beta_{i1}, \sigma^2 M_{i1})$$

So,

$$\frac{\hat{\beta}_i - \beta_{i0}}{\tilde{\sigma}\sqrt{M_{ii}}} \sim t_{(n-p)}(\frac{\beta_{i1} - \beta_{i0}}{\sigma\sqrt{M_{ii}}})$$

(alternative distrabution of T)

By this (*),

$$P(\ \in (-t_{\frac{\alpha}{2}(n-p)}, t_{\frac{\alpha}{2}(n-p)}))$$

Convert this to put β_{i0} in between $(1-\alpha)100$ percent C.I. for $\beta_{i.}$.

$$(\hat{\beta}_1 - t_{\frac{n}{2}(n-p)}\hat{\sigma}\sqrt{M_{ii}}, \hat{\beta}_1 + t_{\frac{n}{2}(n-p)}\hat{\sigma}\sqrt{M_{ii}})$$

1.5 Delete One Prediciton

Very useful in variable selection, cross validation, diagnostics.

Prediction: $\hat{y} = X\hat{\beta} = P_x y$

But this has a drawback as it favors overfitting. Projectioning onto larger spaces will always decrease the norm, $||Q_X y||^2$. (This can decrease errors which would cause you to think it's better, even though it's not.)

To prevent overfitting, try to be objective, withhold y_i when predicting y_i (inverse of a matrix, rank 1 perpendicular)

Theorem 1.5.1 — Theorem 1.7. Suppose $A \in \mathbb{R}^{PxP}$ is a symmetric, nonsingular matrix. and $v \in \mathbb{R}^P$.

Then,

$$(A \pm vv^T)^{-1} = A^{-1} \pm \frac{A^{-1}vv^tA^{-1}}{1 \pm v^T A^{-1} v}$$

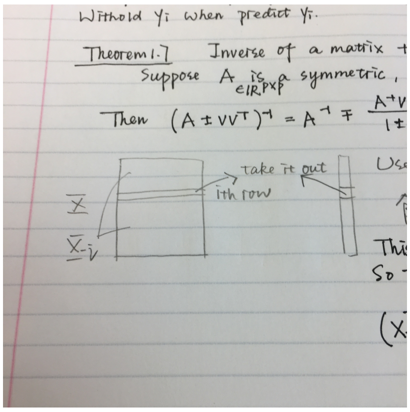


Figure 1.1: Theorem 1.7 Visualization

Use what is left to compute $\hat{\beta_{-i}}$.

$$\hat{\beta_{-i}} = (X_{-1}^T X_{-i})^{-1} X_{-i}^T y_{-i}$$

This can be expanded in simple sum, so that you don't have to do n regressions.

$$\begin{aligned}
(X_{-i}^T X_{-i})^{-1} &= (X^T X - X_i X_i^T)^{-1} \\
&= A^{-1} + \frac{A^{-1} v v^T A^{-1}}{1 - v^t A^{-1} v} \\
&= (X^T X)^{-1} + \frac{(X^T X)^{-1} X_i X_i^T (X^T X)^{-1}}{1 - X_i^T M X_i} \\
X_i^T M X_i &= X_i^T (X^T X)^{-1} \\
&= (P_x)_{ii} \\
&= P_i
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_i &= (X^T X - X_i X_i^T)^{-1} (X^T y - X_i y_i) \\
&= [M + \frac{M X_i X_i^T M}{1 - P_i}](X^T y - X_i y_i) \\
&= M X^T y + \frac{M X_i X_i^T M X^T y}{1 - P_i} - M X_i y_i - \frac{M X_i X_i^T M X_i y_i}{1 - P_i} \\
&= \dots \\
&= \hat{\beta} - \frac{M X_i}{1 - P_i}(y_i - X_i^T \hat{\beta})
\end{aligned}$$

Delete-one regression.
$X_i \hat{\beta_{-i}} = \hat{y}_i - \frac{P_i}{1-P_i}(y_i - \hat{y}_i)$
Friday September 2

Delete- one error

$$y_i - \hat{y}_i^{(-i)}$$

R Recall, you want to leave out y^i so you don't overfit.

The above is equivalent to

$$\begin{aligned}
&y_i - X_i^T \hat{\beta}_{-i} \\
y_i - \hat{y}_i - &\frac{P_i}{1 - P_i}(y_i - \hat{y}_i) \\
(y_i - \hat{y}_i)(1 - &\frac{P_i}{1 - P_i})) \\
&\frac{1}{1 - P_i}(y_i - \hat{y}_i)
\end{aligned}$$

Delete-one cross validation
$\sum_{i=1}^{n} (y_i - \hat{y}_i^{(-i)})^2$
This method is not affected by over fitting.

The following is often used for "tuning" or variable selection (i.e. penalty, bandwidth, regularization, etc). $\sum_{i=1}^{n} \frac{1}{(1-P_i)^2}(y_i - \hat{y}_i)^2$

Note: we will come back to variable selection later.

$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$

$A \subseteq \{1, \ldots, P\}$

Cross validation of A minimizes over $A \in 2^{\{1,\ldots,P\}}$. Best cross validation set.

1.6 Residuals

- Residual

$$\hat{e}_i = y_i - \hat{y}_i$$

- Standardized Residual

$$\mathrm{Var}(\hat{e}_i) = \mathrm{Var}(y_i - \hat{y}_i) = \mathrm{Var}((Q_X)_{ii} y_i)$$

$$= ((Q_X)_{ii} y_i)\sigma^2$$

$$= (1 - P_i)\sigma^2$$

$$sd(\hat{e}_i) = \sqrt{1 - P_i}\sigma$$

$$\hat{sd}(\hat{e}_i) = \sqrt{1 - P_i}\tilde{\sigma}$$

$$\tilde{\sigma} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i^{(-i)})^2}{n - p}$$

- Standardized residual

$$E_i^* = \frac{\hat{e}_i}{\tilde{\sigma}\sqrt{1 - P_i}}$$

- Prediction Error Sum of Squares (PRESS) Residual

$$y_i - \tilde{y}_i^{(-i)} = \frac{1}{1 - P_i}\hat{e}_i = \hat{e}_{iP}$$

$$\hat{e}_{iP} \sim N(0, \frac{\sigma^2}{1 - P_i})$$

- Standardized PRESS Error

$$\frac{\hat{e}_{iP}}{\tilde{\sigma}/\sqrt{1 - _i}} = \frac{\frac{1}{1-P_i}\hat{e}_i}{\tilde{\sigma}(\sqrt{1 - P_i})} = \frac{\hat{e}_i}{\tilde{\sigma}(\sqrt{1 - P_i})} = e_i^*$$

1.7 Influence and Cook's Distance

Definition 1.7.1 — Influence. The difference between predictions with and without a data point.

$$\hat{y}_i - \hat{y}_i^{(-i)}$$

$$\hat{y}_i - \hat{y}_i^{(-i)}$$

$$X_i\hat{\beta} - X_i\hat{\beta}_{-i}$$

$$\propto ||X_i\hat{\beta} - X_i\hat{\beta}_{-i}||^2$$
$$= (X(\hat{\beta} - \hat{\beta}_{-i}))^T(X(\hat{\beta} - \hat{\beta}_{-i}))$$
$$(\hat{\beta}\hat{\beta}_{-i})^T X^T X(\hat{\beta}\hat{\beta}_{-i})$$

Recall,
$\hat{\beta}_{-i} - \hat{\beta} = -\frac{MX_i(y_i - \hat{y}_i)}{1-P_i} = -\frac{MX_i\hat{e}_i}{1-P_i}$
$||X_i\hat{\beta} - X_i\hat{\beta}_{-i}||^2 =$
$=$
Cook's Distance (Technometrics, 1976?)

$$||\frac{\hat{y} - \hat{y}^{(-i)}||^2}{\tilde{\sigma}^2} = \frac{|i\hat{e}_i^2}{(1-P_i)^2\tilde{\sigma}^2}$$

Definition 1.7.2 — Cook's Distance. Cook's distance measures the influence of the i^{th} deservation.

1.8 Orthogonal Decomposition

Recall, $\mathbb{R}^n$ is Euclidean Space.

$\mathscr{S}$ is a subspace ($\mathscr{S} \leq \mathbb{R}^n$)

R $\leq$ is subspace
$\subseteq$ is a subset

For
$\mathscr{S}_1 \leq \mathscr{S}_1\mathscr{S}_2 \leq \mathscr{S}$

$$\mathscr{S}_1 + \mathscr{S}_2 = \{x + y : x \in \mathscr{S}_1, y \in \mathscr{S}_2\}$$

Suppose $\mathscr{S}_1, \mathscr{S}_2 \leq \mathscr{S}$,
$\mathscr{S}_1 + \mathscr{S}_2 = \mathscr{S}$, $\mathscr{S}_1 \perp \mathscr{S}_2$
then,

$$\{\mathscr{S}_1, \mathscr{S}_2\}$$

is called an orthogonal decomposition of $\mathscr{S}$
In this case,

$$\mathscr{S}_1 \oplus \mathscr{S}_2 = \mathscr{S}$$

More generally,

Definition 1.8.1 — Orthogonal Decompositon (O.D.). Let $\mathscr{S}_1, \ldots, \mathscr{S}_k$ be subspaces of $\mathscr{S}$ such that

1. $\mathscr{S}_1, \ldots, \mathscr{S}_k = \{v_1 + \cdots + v_k : v_1 \in \mathscr{S}_1, \ldots, v_k \in \mathscr{S}_k\}$

2. $\mathscr{S}_i \perp \mathscr{S}_j \quad \forall i \neq j$

Then, $\{\mathscr{S}_1, \mathscr{S}_2, \dots, \mathscr{S}_k\}$ is an **orthogonal decomposition** of $\mathscr{S}$. We may write $\mathscr{S} = \mathscr{S}_1 \oplus \mathscr{S}_2 \oplus \cdots \oplus \mathscr{S}_k$.

Proposition 1.5 If $\mathscr{S}_1, \dots, \mathscr{S}_k$ is an O.D. of $\mathscr{S}$, then any $v \in \mathscr{S}$ can be uniquely written as

$$v_1 + \cdots + v_k$$

, where $v_1 \in \mathscr{S}_1, \dots v_k \in \mathscr{S}_k$.

Wednesday September 7

Definition 1.8.2 — Direct Difference. Let $\mathscr{S}_1 \leq \mathscr{S}_2 \leq \mathbb{R}^n$. Then,

$$\mathscr{S}_2 \cap \mathscr{S}_1^{\perp} \equiv \mathscr{S}_2 \ominus \mathscr{S}_1$$

is called **direct difference**. This is almost the same as orthogonal complement, except it is within $\mathscr{S}_2$.

Proposition 1.6 If $\mathscr{S}_1 \leq \mathscr{S}_2$, then

$$\mathscr{S}_2 = \mathscr{S}_1 \oplus (\mathscr{S}_2 \ominus \mathscr{S}_1)$$

Proposition 1.7 - Orthogonal Decomposition and Projection Consider a Hilbert Space, $\mathscr{H} = \{\mathbb{R}^n, <,>_A\}$,

1. If $\mathscr{S} \leq \mathscr{S}_1 \perp \mathscr{S}_2$ in $\mathscr{H}$, then

$$P_{\mathscr{S}_1}(A) P_{\mathscr{S}_2}(A) = 0$$

2. If $\mathscr{S} \leq \mathscr{H}, \dots, \mathscr{S}_k \leq \mathscr{H}$, and $\mathscr{S}_1 \perp \cdots \perp \mathscr{S}_k$, then

$$P_{\mathscr{S}_1, \oplus \cdots \oplus \mathscr{S}_k}(A) = P_{\mathscr{S}_1}(A) + \cdots + P_{\mathscr{S}_k}(A)$$

3. If $\mathscr{S}_1 \leq \mathscr{S}_2 \leq \mathbb{R}^n$, then

$$P_{\mathscr{S}_2 \ominus \mathscr{S}_1}(A) = P_{\mathscr{S}_2}(A) - P_{\mathscr{S}_1}(A)$$

Theorem 1.8.1 — Generalization of the earlier Cochran's Theorem. Suppose $X \sim N(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{nxn}$ is positive definite.

Let $\mathscr{H} = \{<,>_{\Sigma^{-1}}\}$. Suppose $\mathscr{S}_1, \dots \mathscr{S}_k, \mathscr{S} \leq \mathscr{H}$ such that $\mathscr{S} = \mathscr{S}_1 \oplus \cdots \oplus \mathscr{S}_k$.

Let

$$w_i = ||P_{\mathscr{S}_i}(\Sigma^{-1}) X||^2_{\Sigma^{-1}}$$
$$w = ||P_{\mathscr{S}}(\Sigma^{-1}) X||^2_{\Sigma^{-1}}$$

Then,

1. $w = w_1 + \cdots + w_k$
2. $w_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp w_k$
3. $w_i \sim \chi^2_{r_i}$
   $w \sim \chi^2_r$
   where r_i is the $dim(\mathscr{S}_i)$, r is the $dim(\mathscr{S})$, and $r = r_1 + \cdots + r_k$.

Notation 1.1. *We use $\oplus$ for spaces. We can also use $\oplus$ function to stack up matrices. Let $A_1, \ldots, A_k$ be matrices with arbitrary dimensions.*

$$A_1 \oplus \cdots \oplus A_k = \begin{pmatrix} A_1 & \ldots & 0 \\ & \ddots & \\ 0 & \ldots & A_k \end{pmatrix}$$

1.9 Lack of Fit Test

Goodness of Fit

At each x_i you have multiple observations, say $y_{i1}, \ldots, y_{im_i}$. In this case, you may test to see if a linear model,$y_i = x_i^T \beta + \varepsilon_i$, is the correct choice for fitting the data. In general, lack of fit refers to testing whether any (linear, generalized, etc) model is adequately describing the data.

Denote
$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{im_i} \end{pmatrix}$
$1_{m_i} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$
$X = \begin{pmatrix} X_1^T \\ \vdots \\ X_m^T \end{pmatrix}$
Assume

$$y_{ij} = X_i^T \beta + \varepsilon_{ij}$$

where $\varepsilon \sim^{iid} N(0, \sigma^2)$.
The point is that you have $y_{i1} \ldots y_{jm}$ for each X_i.

In matrix form,

$$(1_{m_1} \oplus \cdots \oplus 1_{m_n}) X\beta + \varepsilon$$

So, let N denote a full sample size.

$$N = m_1 + \cdots + m_n$$

this is a special case of linear model, except the design matrix is structured $(1_{m_1} \oplus \cdots \oplus 1_{m_n})X$ instead of X. So the formula for MLE (and so on) is the same.

$$X \leftrightarrow (1_{m_1} \oplus \cdots \oplus 1_{m_n})X$$

So,

$$\hat{\beta} = ([(1_{m_1} \oplus \cdots \oplus 1_{m_n})X])^T([(1_{m_1} \oplus \cdots \oplus 1_{m_n})X])^{-1}[(1_{m_1} \oplus \cdots \oplus 1_{m_n})X]^T y$$

$$
\begin{aligned}
\hat{y} &= (1_{m_1} \oplus \cdots \oplus 1_{m_n})X\hat{\beta} \\
&= (1_{m_1} \oplus \cdots \oplus 1_{m_n})X([(1_{m_1} \oplus \cdots \oplus 1_{m_n})X])^T([(1_{m_1} \oplus \cdots \oplus 1_{m_n})X])^{-1}[(1_{m_1} \oplus \cdots \oplus 1_{m_n})X]^T y \\
&= (1_{m_1} \oplus \cdots \oplus 1_{m_n})X[X^T \begin{pmatrix} m_1 & \dots & 0 \\ & \ddots & \\ 0 & \dots & m_n \end{pmatrix} X]^{-1}X^T(1_{m_1} \oplus \cdots \oplus 1_{m_n})
\end{aligned}
$$

So, in linear model with replication we have our hypotheses for lack of fit test,

$$H_O : E(y_i) = 1_{m_i} X_i^T \beta$$

$$H_1 : E(y_i) = 1_{m_i} \mu_i$$

We are testing whether the arbitrary means, $\mu_1, \dots \mu_n$ sit on the same line.

Friday September 9

Under H_1,

$$y = (1_{m_1} \oplus \cdots \oplus 1_{m_n}) \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} + \varepsilon$$

So the $\hat{y}$ under this model,

$$\hat{y}_{H_1} = P_{1_{m_1} \oplus \cdots \oplus 1_{m_n}} y = (1_{m_1} \oplus \cdots \oplus 1_{m_n}) \begin{pmatrix} m_1 & & \\ & \ddots & \\ & & m_n \end{pmatrix} (1_{m_1} \oplus \cdots \oplus 1_{m_n})^T y$$

but under H_0,

$$\hat{y}_{H_0} = P_{(1_{m_1} \oplus \cdots \oplus 1_{m_n})X} y$$

$$\mathscr{S}_1 = \text{span}\{(1_{m_1} \oplus \cdots \oplus 1_{m_n})X\} \quad \text{(p-dim)}$$

$$\mathscr{S}_2 = \text{span}\{(1_{m_1} \oplus \cdots \oplus 1_{m_n})\} \quad \text{(n-dim)}$$

$$\mathscr{S}_3 = \mathbb{R}^N \quad (N = m_1 + \cdots + m_n)$$

$$\mathscr{S}_1 \leq \mathscr{S}_2 \leq \mathscr{S}_3$$

R Above used the fact that span(AB) $\subseteq$ span(A)

Lemma 1.1 If $\mathscr{S}_1 \leq \mathscr{S}_2 \leq \mathscr{S}_3$ then

1. $\mathscr{S}_3 \ominus \mathscr{S}_2 \leq \mathscr{S}_3 \ominus \mathscr{S}_1$
2. $(\mathscr{S}_3 \ominus \mathscr{S}_1) \ominus \mathscr{S}_2 = \mathscr{S}_3 \mathscr{S}_2$
3. $(\mathscr{S}_3 \ominus \mathscr{S}_1) = (\mathscr{S}_3 \ominus \mathscr{S}_2) \oplus (\mathscr{S}_2 \ominus \mathscr{S}_1)$

Go back to lack of fit,

$$(\mathscr{S}_3 \ominus \mathscr{S}_1) = (\mathscr{S}_3 \ominus \mathscr{S}_2) \oplus (\mathscr{S}_2 \oplus \mathscr{S}_1)$$

$$P_{\mathscr{S}_3 \ominus \mathscr{S}_1} y = P_{\mathscr{S}_3 \ominus \mathscr{S}_3} y + P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y \quad \text{(Orthogonal Decomposition)}$$

$$||P_{\mathscr{S}_3 \ominus \mathscr{S}_1} y||^2 = ||P_{\mathscr{S}_3 \ominus \mathscr{S}_3} y||^2 + ||P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y||^2$$

$$dim(\mathscr{S}_2 \ominus \mathscr{S}_1) = n - p$$

$$dim(\mathscr{S}_3 \ominus \mathscr{S}_2) = N - n$$

Now,

$$E(P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y) = P_{\mathscr{S}_3 \ominus \mathscr{S}_2} E(y) = P_{\mathscr{S}_3 \ominus \mathscr{S}_2} \mu = 0$$

But,
$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \in \mathscr{S}_2$
and,
$(1_{m_1} \oplus \cdots \oplus 1_{m_n})\underline{\mu}$

$$Var(P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y) = P_{\mathscr{S}_3 \ominus \mathscr{S}_2} Var(y) P_{\mathscr{S}_3 \ominus \mathscr{S}_2} = \sigma^2 P^2_{\mathscr{S}_3 \ominus \mathscr{S}_2} = \sigma^2 P_{\mathscr{S}_3 \ominus \mathscr{S}_2}$$

We know that $y \sim N(\mu, \sigma^2 I_n)$. So,

$$P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y \sim N(0, \sigma^2 P_{\mathscr{S}_3 \ominus \mathscr{S}_2}$$

Similarly,

$$E(P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y) = P_{\mathscr{S}_2 \ominus \mathscr{S}_1} E(y)$$

which under H_0 is,

$$P_{\mathscr{S}_2 \ominus \mathscr{S}_1}(1_{m_1} \oplus \cdots \oplus 1_{m_n}) X\beta = 0$$

$$Var(P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y) = \sigma^2 P_{\mathscr{S}_2 \ominus \mathscr{S}_1}$$

$$P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y \sim N(0, \sigma^2 P_{\mathscr{S}_2 \ominus \mathscr{S}_1})$$

By Chochran's Theorem:
Under H_O,

$$||P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y||^2 \sim \chi^2_{(N-n)}$$
$$||P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y||^2 \sim \chi^2_{(n-p)}$$

$$||P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y||^2 \perp\!\!\!\perp ||P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y||^2$$

So our lack of fit test is:

$$\frac{||P_{\mathscr{S}_2\ominus\mathscr{S}_1}y||^2/(n-p)}{||P_{\mathscr{S}_3\ominus\mathscr{S}_2}y||^2/(N-n)} \sim F_{n-p,N-n}$$

1.10 Explicit Intercept

We now apply this $\mathscr{S}_1, dots$ argument to another problem: special linear model.

$$y_i = \alpha + \beta^T X_i + \varepsilon_i \quad i = 1,\dots,n$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$Y = 1_n\alpha + X\beta + \varepsilon = (1_n X)\begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \varepsilon = U\eta + \varepsilon$$

Let $P_{1_n} = 1_n(1_n^T 1_n)^{-1}1_n^T = \frac{1_n 1_n^T}{n}$.

Note that for all $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^n$,

$$P_{1_n}a = \frac{1_n 1_n^T a}{n} = 1_n\bar{a}, \quad \bar{a} = \frac{1}{n}\sum_{i=1}^n a_i$$

which is a mean projection. (?)
$Q_{1_n} = I_n - P_{1_n}$ (projection on$1_n^\perp$)

$$Q_{1_n}a = \begin{pmatrix} a_1 - \bar{a} \\ \vdots \\ a_n - \bar{a} \end{pmatrix}$$

Decompose X:

$$X = P_{1_n}X + Q_{1_n}X$$

$$U\eta = 1_n\alpha + X\beta = 1_n\alpha + P_{1_n}X\beta + Q_{1_n}X\beta = 1_n(\alpha + \frac{1_n^T X\beta}{n}) + Q_{1_n}X\beta = (1_n Q_{1_n}X)\begin{pmatrix} \alpha^\star \\ \beta \end{pmatrix} = (1_n Q_{1_n}X)\eta^\star = U^\star\eta^\star$$

So we do least squres of

$$(y - U^\star\eta^\star)^T(y - U^\star\eta^\star)$$

and minimize this over all $\eta^\star \in \mathbb{R}^{Px1}$

$$\hat{\eta}^\star = (U^{\star T}U^\star)U^{\star T}y$$

$$U^{\star T}U^\star = \begin{pmatrix} 1_n^T \\ (Q_{1_n}X)^T \end{pmatrix}(1_n Q_{1_n}X) = \begin{pmatrix} 1_n^t 1_n & Q_{1_n}X1_n \\ 1_n^T Q_{1_n}X & Q_{1_n}XQ_{1_n}X \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & X^T Q_{1_n}X \end{pmatrix}$$

$$\hat{\eta}^\star = \begin{pmatrix} n^{-1} & 0 \\ 0 & (X^T Q_{1_n}X)^{-1} \end{pmatrix}\begin{pmatrix} 1_n \\ (Q_{1_n}X)^T \end{pmatrix} y$$

Monday September 12
So

$$\hat{\alpha}^\star = n^{-1}1_n^T y$$

$$\hat{\beta} = (X^T QX)^{-1}X^T Qy$$

$$\hat{\alpha} = n^{-1}1_n^T y - n^{-1}X\hat{\beta}^\star$$

For statistical inference, we want to make a decomposition of $\mathbb{R}^n$.
Let, $\mathscr{S}_1 = \text{span}(1_n), \mathscr{S}_2 = \text{span}(1_n, X), \mathscr{S}_3 = \mathbb{R}^n$.

Then,

$$(\mathscr{S}_3 \ominus \mathscr{S}_1) = (\mathscr{S}_3 \ominus \mathscr{S}_2) \oplus (\mathscr{S}_2 \ominus \mathscr{S}_1)$$

Then,

$$||P_{\mathscr{S}_3 \ominus \mathscr{S}_1}y||^2 = ||P_{\mathscr{S}_3 \ominus \mathscr{S}_2}y||^2 + ||P_{\mathscr{S}_2 \ominus \mathscr{S}_1}y||^2$$

Or,

$$SSTotal = SSError + SSRegression$$

We may compute these terms,

$$\begin{aligned} P_{\mathscr{S}_3 \ominus \mathscr{S}_1} = P_{\mathscr{S}_3} - P_{\mathscr{S}_1} \\ &= I_n - \frac{1_n 1_n}{1_n^T 1_n} \\ &= Q_1 n \\ \mathscr{S}_2 \ominus \mathscr{S}_1 &= \text{span}(Q_{1_n}X) \\ P_{\mathscr{S}_2 \ominus \mathscr{S}_1} &= QX(X^T QX)^{-1}QX^T \\ P_{\mathscr{S}_3 \ominus \mathscr{S}_2} &= Q - QX(X^T QX)^{-1}X^T Q \end{aligned}$$

By Cochran's Theorem, (these are orthogonalized projections, etc),

$$\begin{aligned} ||P_{\mathscr{S}_3 \ominus \mathscr{S}_1}y||^2 &\sim \chi^2(n-1) \\ ||P_{\mathscr{S}_2 \ominus \mathscr{S}_1}y||^2 &\sim \chi^2_{(p-1)} \\ ||P_{\mathscr{S}_3 \ominus \mathscr{S}_2}y||^2 &\sim \chi^2_{(n-p-1)} \end{aligned}$$

R
$$dim(\mathscr{S}_3) = n$$

$dim(\mathscr{S}_2) = p+1$ $dim(\mathscr{S}_3) = 1$

We also know that these are all independent of each other. So we can test regression effect with the following hypothesis:

$$H_0 : \beta - 0$$

$$\frac{||P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y||^2/(p-1)}{||P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y||^2/(n-p-1)} = \frac{y^T QX(X^TQX)^{-1}QX^Ty/(p-1)}{y^T(Q-QX(X^TQX)^{-1}X^TQ)y/(n-p-1)} \sim F_{p-1,n-p-1}$$

Distributions

$$\hat{\beta}(X^TQX)^{-1}X^TQy$$

$E(\hat{\beta}) = (X^TQX)^{-1}X^TQ(1_{n\alpha} + X\beta = (X^TQX)^{-1}X^TQX\beta = \beta$
$\text{Var}(\hat{\beta}) = (X^TQX)^{-1}X^TQ(\sigma^2 I_n)QX(X^TQX)^{-1} = \sigma^s(X^TQX)^{-1}$
$\hat{\alpha} = \hat{\alpha}^\star - X^T\hat{\beta}$

Because $\hat{\beta}$ is a function of Qy and $\hat{\alpha}^\star$ is a function of $P_{1_n}y$ (and these are orthogonal to each other and thus by normality also independent).

$$\text{Var}(\hat{\alpha} = \text{Var}(\hat{\alpha}^\star) + \text{Var}(\bar{X}^T\hat{\beta}) = \text{Var}(\bar{y}) + \text{Var}(\bar{X}^T\hat{\beta}) = \tfrac{\sigma^2}{n} + \sigma^2\bar{X}^T(X^TQX)^{-1}\bar{X}$$

$$\hat{\alpha}\ N(\alpha, \frac{\sigma^2}{n} + \sigma^2\bar{X}^T(X^TQX)^{-1}\bar{X})$$

$$\begin{aligned} Cov(\hat{\alpha},\hat{\beta}) &= Cov(\hat{\alpha}^\star - \bar{X}^T\hat{\beta},\hat{\beta}) \\ &= -\bar{X}^T Var(\hat{\beta}) \\ &== \bar{X}^T\sigma^2(X^TQX)^{-1} \end{aligned}$$

$$\begin{pmatrix} al\hat{p}ha \\ \hat{\beta} \end{pmatrix} \sim N[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} + \sigma^2\bar{X}^T(X^TQX)^{-1}\bar{X} & -\sigma^2\bar{X}^T(X^TQX)^{-1}\bar{X} \\ -\sigma^2\bar{X}^T(X^TQX)^{-1}\bar{X} & \frac{\sigma^2}{n} + \sigma^2\bar{X}^T(X^TQX)^{-1}\bar{X} \end{pmatrix}]$$

Estimate σ^2

$$||P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y||^2 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}^T X_i)^2 \sim \sigma^2\chi^1_{n-p-1}$$

So,

$$E(||P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y||^2) = \sigma^2(n-p-1)$$

Thus,

$$\hat{\sigma}^2 = \frac{||P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y||^2}{n-p-1}$$

Theorem 1.10.1 Under the explicit intercept model,

1. $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^1)$ is UMVUE of $(\alpha, \beta, \sigma^2)$ by Lehmann-Sheffe.
2. $$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim N[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} + \sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} & -\sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} \\ -\sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} & \frac{\sigma^2}{n} + \sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} \end{pmatrix}]$$
3. $(n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{(n-p-1)}$
4. $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \perp\!\!\!\perp \hat{\sigma}^2$

1.11 R^2

Proportion of Sum of Squares (SS) explained by regression (i.e. by β).

$$R^2 = \frac{SSR}{SST} = \frac{||P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y||^2}{||P_{\mathscr{S}_3 \ominus \mathscr{S}_1} y||^2}$$

But we know that,

$$R^2 = \frac{||P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y||^2}{||P_{\mathscr{S}_2 \ominus \mathscr{S}_1} y||^2 + ||P_{\mathscr{S}_3 \ominus \mathscr{S}_2} y||^2} = \frac{SSR}{SSR + SSE} = \frac{SSR/SSE}{SSR/SSE + 1}$$

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{(n-p-1)}{p} \frac{SSR}{SSE}$$

$$\frac{SSR}{SSE} = \frac{p}{(n-p-1)} F$$

$$R^2 = \frac{\alpha F}{\alpha F - 1}$$

where $\alpha = \frac{p}{n-p-1}$

This is how we compute the null distribution of R^2.

1.12 Multicollinearity

Wednesday September 14

$y = C_1 \beta_1 + \cdots + C_p \beta_p$

$$X = (C_1, \ldots, C_P) = \begin{matrix} X_1^T \\ \vdots \\ X_n^T \end{matrix}$$

In an extreme case, multicolinearity simply means that the $C_1, \ldots, C_p$ are linearly dependent. In this case β is not identifiable.

We have C_1, C_2, C_3.

$$C_1 = a$$
$$C_2 = 2a$$
$$C_3 = b$$

$$\begin{aligned} y &= a\beta_1 + 2a\beta_2 + b\beta_3 + \varepsilon \\ &= a(\beta_1 + 2\beta_2) + b\beta_3 + \varepsilon \end{aligned}$$

$\beta_1 \& \beta_2$ cannot be split.

In the less extreme case, $X^T X$ is nearly singular, meaning it has small eigenvalues. In this case, although β is identifiable, they have large variance For example, if $C_1 = aC_2x2a$ then β_1, β_2 have large variance which means your parameterization is not good. So may define new parameterization.

$$\gamma_1 = \beta_1 + 2\beta_2$$
$$\gamma_2 = \beta_3$$

If you run regression against these then the variance would be 'normal'.

S0, how to wee out redundant varaibles? One way Variance Inflation Factor (VIF) which for each $i = 1, 2, \ldots, p$ regresses C_i on $\{C_1, \ldots, C_p \setminus C_i\}$ then you get R^2 for this regression call it R_i^2.

If C_i is redunent then R_i^2 would be close to 1.

$$VIF_i = \frac{1}{1 - R_i^2}$$

1.13 Variable Selection

$$y = C_1\beta_1 + \cdots + C_p\beta_p + \varepsilon$$

Some of these β's are zero.

Let us define an active set of parameters,

$$A_0 = \{i : \beta_i \neq 0\}$$

To estimate A_0 is the goal of variable selction.

Mallow's C_p criterion

The fundamental issue is variable selction, penalty - penalizing the number of parameters, so you cannot use something like $y - \hat{y}$ as criterion. The more variables you have the smaller $||\hat{y} - y||^2$ is. So we want to penalize the number of parameters in a reasonable way.

Let any subset $A \subset \{1, \ldots, p\}$,

$$X_A = \{C_i : i \in A\}$$

Notation 1.2. *While we often use X for iid variables (a vector), but here X is a matrix and X_i were referring to its columns. We've changed X_i to C_i to better reflect that we are dealing with columns of X.*

So, $A = \{1,3,5\}$,

$$X_A = \begin{pmatrix} C_1 \\ C_3 \\ C_5 \end{pmatrix}$$

Let P_{X_A}, Q_{X_A} be the projection on to span(X_A), span$(X_A)^{\perp}$. For example,

$$P_{X_A} = X_A(X_A^T X_A)^{-1} X_A^T$$

Let $\mu = E(y) = X\beta = X_{A_0}\beta_{A_0}$.

Mallow says we minimize

$$\frac{E||P_A y - \mu||^2}{\sigma^2}$$

among all $A \subset \{1,\ldots,p\}$.

But we do not know what σ^2 or μ are. If so, we would already know A_0. We must estimate these.

$$E||P_{X_A} y - \mu||^2 = tr(E(P_{X_A} y - \mu)(P_{X_A} y - \mu)^T)$$

$$\begin{aligned}
E(P_{X_A} y - \mu)(P_{X_A} y - \mu)^T &= E[(P_{X_a} y - P_{X_a}\mu) + (P_{X_a}\mu - \mu)][(P_{X_a} y - \mu) + (P_{X_a}\mu - \mu)]^T \\
&= \text{expand, two terms are zero} \\
&= E(P_{X_a} y - P_{X_a}\mu)(P_{X_a} y - P_{X_a}\mu) + (P_{X_a}\mu - \mu)(¶_{X_a}\mu - \mu)^T \\
&= Var(P_{X_a} y) \\
&= P_{X_a}\sigma^2 I_n P_{X_a} = \sigma^2 P_{X_a} \\
&= tr(\sigma^2 P_{X_a} + Q_{X_a}\mu\mu^T Q_{X_a}) \\
&= \sigma * 2tr(P_{X_a}) + tr(Q_{X_a}\mu\mu^t Q_{X_a}) \\
&= \sigma^2(\#(A)) + tr(\mu^T Q_{X_a}\mu)
\end{aligned}$$

$$\Rightarrow E\frac{||P_{X_A} y - \mu||^2}{\sigma^2} = \#(A) + \frac{tr(\mu^T Q_{X_a}\mu)}{\sigma^2}$$

Now let's estimate $\frac{\mu^T Q_{X_a}\mu}{\sigma^2}$.

Recall, if U is a random vector with multivariate normal distribution so

$$E(U) = e$$

$$Var(U) = Q_{X_A}$$

$$U^T U \sim \chi^2_{(rank(Q)_{X_A})}(||e||^2)$$

Also, $W \sim \chi^2_{(r)}(\delta)$ where $E(W) = r + \delta$.

Go back to our problem of estimating $\frac{\mu^T Q_{X_a}\mu}{\sigma^2}$.

What about $y^t Q_{X_A} y$? We know that

$$E(\frac{Q_{X_A} y}{\sigma}) = \frac{Q_{X_A}\mu}{\sigma}$$

and

$$Var(\frac{Q_{X_A} y}{\sigma}) = \frac{1}{\sigma^2} Q_{X_A} \sigma^2 I_n = 0$$

So,

$$\frac{Q_{X_A} y}{\sigma} \sim N(\frac{Q_{X_A}\mu}{\sigma}, 0)$$

So

$$(\frac{Q_{X_A} y}{\sigma})^T (\frac{Q_{X_A} y}{\sigma}) \sim \chi^2_{(n-\#(A))}((\frac{Q_{X_A}\mu}{\sigma})^T (\frac{Q_{X_A}\mu}{\sigma})) = \chi^2_{(n-\#(A))}(\frac{\mu^T Q_{X_A}\mu}{\sigma^2})$$

Thus,

$$E(\frac{y^T Q_{X_A} y}{\sigma^2}) = n - \#(A) + \frac{\mu^T Q_{X_A}\mu}{\sigma^2}$$

Which, if you subtract over the n and #(A) you get an unbiased estimator of $\frac{\mu^T Q_{X_A}\mu}{\sigma^2}$.
But σ^2 is still unkown, but we use full model,

$$\hat{\sigma}^2 = \frac{y^T Q_{X_A} y}{n-p}$$

Now we can estimate $\frac{\mu^T Q_{X_A}\mu}{\sigma^2}$ by

$$\frac{y^T Q_{X_A} y}{\frac{y^T Q_X y}{n-p}} - n + 2\#(A) = (n-p)\frac{y^T Q_{X_A} y}{y^T Q_X y} - n + 2\#(A)$$

So to recap,

$$E\frac{||P_{X_A} y - \mu||^2}{\sigma^2} = (n-p)\frac{y^T Q_{X_A} y}{y^T Q_X y} - n + 2\#(A)$$

Friday September 16

Akaike/Bayesian Information Criteria (AIC/BIC)

Suppose we have some generic (i.e. not related to the design/covariance in regression context) $X_1, \ldots, X_n$, a sample of independent random vectors with joint density $f_\theta(x_1, \ldots, x_n)$.

$\theta \in \Theta \subset \mathbb{R}^P$

$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_P \end{pmatrix}$

Let $M_0 \subset \{1,\ldots,P\}$ be the true active set, that is $\{i : \theta_i \neq 0\} = M_0$. Also, let $M_0 \subset \{1,\ldots,P\}$. We want to recover the true active set M_0.

Let Θ_M be the paramter space, corresponding to M.

$$\Theta_M = \{\theta \in \Theta : \theta_i = 0 iff i \notin M\}$$

Of course we also thave Θ_{M_0}.

for each $M \subset \{1,\ldots,P\}$ define,

$$L_M = \sup_{\theta \in M} f_\theta(x_1,\ldots,x_n)$$

Then,

$$\text{AIC}(M) = -2\log L_M + 2(\#M)$$

$$\text{BIC}(M) = -2\log L_M + (\log n)(\#M)$$

Use them

$$\hat{M} = \arg\min\{\text{AIC}(M) : M \in 2^{\{1,\ldots,P\}}\}$$
$$\hat{M} = \arg\min\{\text{BIC}(M) : M \in 2^{\{1,\ldots,P\}}\}$$

When P is large, this is called **forward backward selection** instead ov **Best Set Selection**.

Specialized to Gaussian Linear Regression Model

Here there is no variable selection,
$\theta = \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}$
$\Theta \subset \mathbb{R}^{P+1}$ (M is in Θ)
$\beta \in B$
$subset \mathbb{R}^P$ (A is in B)

In our case,

$$y \sim N(X_A\beta_A, \sigma^2 I_n)$$

where $A \subseteq B$ which is where β is.
Note we are again using the notation $X_A = \{C_i : i \in \beta\}$ and that

$$\#A + 1 = \#M$$

If A is an active set of β then $M = A \cup \{p+1\}$ is the active set of θ because σ^2 is always active.
But $L_M = ?$

Recall, MLE for β is (under A),

$$\hat{\beta}_A = (X_A^T X_A)^{-1} X_A^T y$$

$$\hat{\sigma}_A^2 = \frac{y^T Q_{X_A} y}{n}$$

So the likelihood at $(\hat{\beta}_A, \hat{\sigma}_A^2)$,

$$\begin{aligned}
f_{\hat{\theta}_A}(x_1, \dots, x_n) &= \frac{1}{(2\pi)^{\frac{n}{2}} \det(\hat{\sigma}_A^2 I_n)} e^{\frac{-1}{2\hat{\sigma}_A^2} ||y - X_A \hat{\beta}_A||^2} \\
&= L_M \qquad = \dots e^{-\frac{1}{2\hat{\sigma}_A^2} ||y - X_A \hat{\beta}_A||^2||^2} \\
&= \dots e^{-\frac{1}{2\frac{y^T Q_{X_A} y}{n}} ||y - X_A (X_A^T X_A)^{-1} X_A^T y||^2||^2} \\
&= \dots e^{-\frac{1}{2n} ||y - P_{X_A} y||^2||^2} \\
&= \dots e^{-\frac{n}{2} ||Q_{X_A} y||^2||^2}
\end{aligned}$$

$$\begin{aligned}
L_M &= \frac{1}{(2\pi)^{\frac{n}{2}} (\frac{y^T Q_{X_A} y}{n})^{\frac{n}{1}}} e^{-\frac{n}{2}} \\
\log L_M &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\frac{y^t Q_A y}{n}) - \frac{n}{2}
\end{aligned}$$

$$\text{AIC}(A) = -2 \log L_M + 2(\#M) =$$

It is equivalent ti minimize,

$$n \dots$$

For BIC, replace $2(\#A+1)$ by $(\log n)(\#A+1)$.
Variable Selection Consistency
(as opposed to estimation consistency)

You have data, $(X_1, \dots, X_n) = \mathscr{X}$ (again, generic X, not in regresssion context). An estimator,

$$\mathscr{X} \to \Theta$$

Variable selection,

$$\hat{A} : \mathscr{X} \to 2^{\{1,\dots,P\}}$$

that is to say,

$$(x_1, \dots, x_n) \mapsto M$$

Definition 1.13.1 — Variable Selector Consistancy. A variable selector, $\hat{A}$ is said to be **consistant** if

$$P(\hat{A} = A_0) \to 1$$

where A_0 is the true action set.

Next, BIC in variable seleciton consistancy.

Ordering of sequences, $\{a_n\}, \{b_n\}$ 2 sequences in $\mathbb{R}$, positive...

Notation 1.3 (Asymptotic Order of Magnitude). $a_n \prec b_n$ *if* $\frac{a_n}{b_n} \to 0$ *as* $n \to \infty$.

■ **Example 1.1** • $a_n \prec 1 \Leftrightarrow a_n \to 0$
• $a_n \prec n \Leftrightarrow \frac{a_n}{n} \to 0$

■

■ **Example 1.2** •
$$\begin{aligned} a_n &\succ 1 \\ &\Rightarrow 1 \prec a_n \\ &\Rightarrow \frac{1}{a_n} \to 0 \\ &\Rightarrow a_n \to \infty \end{aligned}$$
• $n^{\frac{1}{2}}$

■

The symbol $\sim$ *means both* $\prec$ *and* $\succ$.

Monday September 19
Lemma 1.3

Under some regularity conditions (identifiability, smoothness of log likelihood, support doesn't depend on parameters, ...) then

1. $\Theta_{M_0} \subseteq \Theta_M$
$$2(\log L_M - \log L_{M_0}) \to^{\mathscr{D}} \chi^2_{(\#M - \#M_0)}$$
Here, recall,
$$L_M = \sup_{\theta \in \Theta} f_\theta(x_1, \dots, x_n)$$
2. If $\Theta_M \subseteq \Theta_{M_0}$ then,
$$n^{-1} 2(\log L_M - \log L_{M_0}) \to^{P} 2(\sup_{\theta\in\Theta} E\log f_\theta(x_1,\dots,x_n) - E\log f_{\theta_0}(x_1,\dots,x_n))$$
Moreover, if $M \subset M_0$ then
$$\lim_{n\to\infty}(2(\sup_{\theta\in\Theta} E\log f_\theta(x_1,\dots,x_n) - E\log f_{\theta_0}(x_1,\dots,x_n))) < 0$$

Theorem 1.13.1 Let BIC(M) = $-2\log L_M + (cn)(\#M)$ where $1 \prec c(n) \prec n$. This generalizes BIC so that $c(n)$ replaces $\log(n)$ but still converges slower than n (as does log).

Let $\hat{M} = \arg\min_{M \in 2^{\{1,2,\dots,p\}}} BIC(M)$ then
$$P(\hat{M} = M_0) = 1$$

Proof. Consider the difference,
$$BIC(M) - BIC(M_0) = 2(\log L_{M_0} - \log L_M) + c(n)(\#M - \#M_0)$$

We want to show (with probability going to 1) that

$$BIC(M) - BIC(M_0) > 0 \quad \forall M \neq M_0$$

Case 1 $M \supset M_0$
Then $c(n)(\#M - \#M_0) \to \infty$

Meanwhile, $2(\log L_{M_0} - \log L_M) = O_p(1)$.

R Fact. If $U_n = O_p(1), \alpha_n \to \infty$ then

$$P(U_n + \alpha_n > 0) \to 1$$

So,

$$P(BIC(M) - BIC(M_0)) \to 1$$

Case 2 $M \subseteq M_0$
$n^{-1}2(\log L_{M_0} - \log L_M) \to c(n) > 0$

R Fact. $n^{-1}U_n \to c > 0, \alpha_n \prec n$ and $n^c \prec n$ then

$$P(U_n + \alpha_n > 0) \to 1$$

So again,

$$P(BIC(M) - BIC(M_0)) \to 1$$

Thus, $P(BIC(M)$ is uniquely minimized at $M_0) \to 1$.

■

1.14 Non iid Linear Regression

Suppose

$$y = X\beta + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2\Sigma)$ with arbetrary but known matrix $\Sigma > 0$.
Then MLE for $\hat{\beta}$ is

$$\hat{\beta} = (X^T\Sigma X)^{-1}X^T\Sigma^{-1}X$$

MLE for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = ||Q_X(\Sigma^{-1})y||^2/n$$

But remember $||Q_X(\Sigma^{-1})y||^2_{\Sigma^{-1}} \sim \sigma^2\chi^2_{n-p}$, so now we have

$$E\left(||Q_X(\Sigma^{-1})y||^2_{\Sigma^{-1}}\right) = \sigma^2(n-p)$$

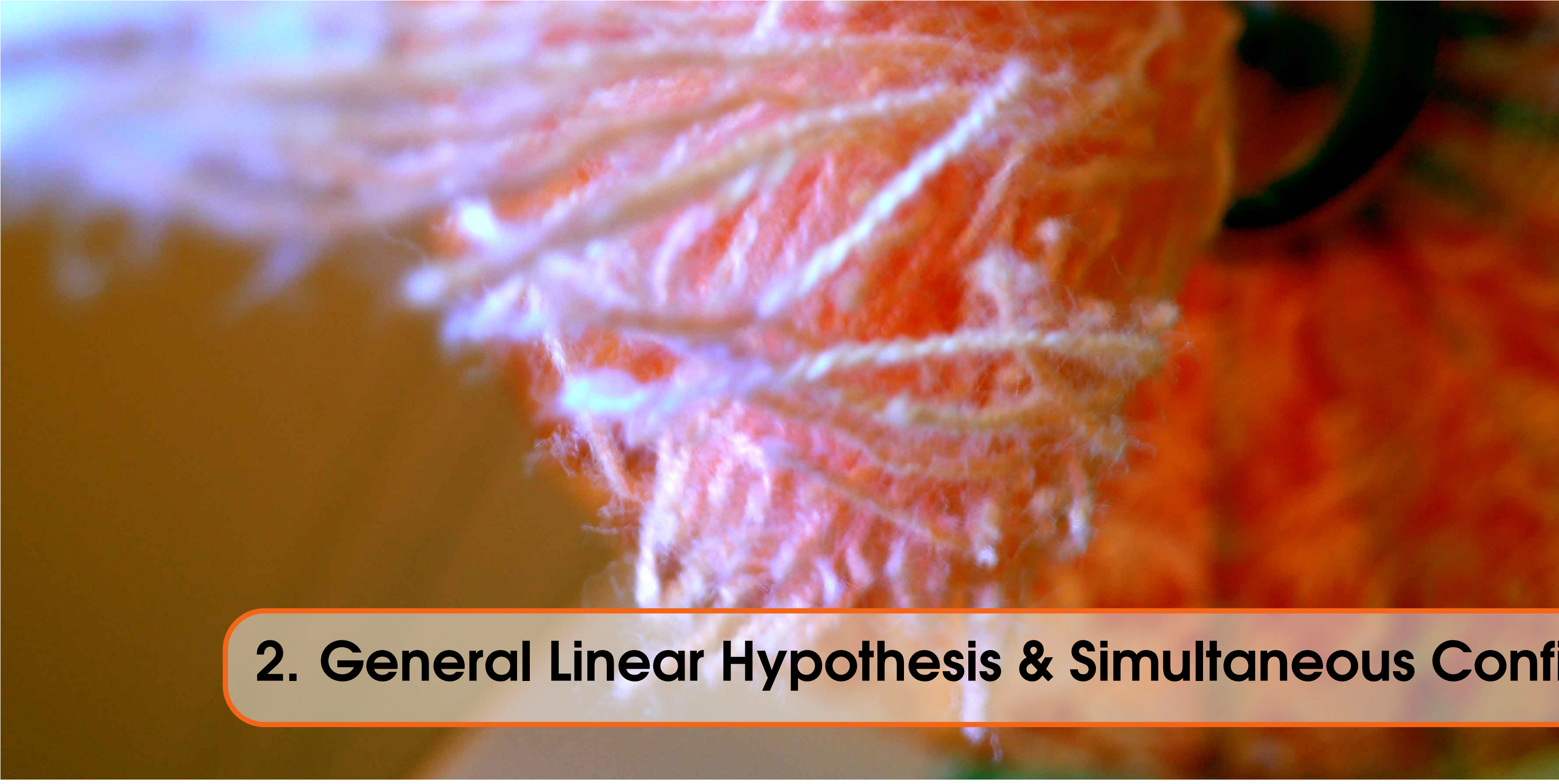
so the unbiased estimator is,

$$\tilde{\sigma}^2 = \frac{||Q_X(\Sigma^{-1})y||^2_{\Sigma^{-1}}}{n-p}$$

Theorem 1.14.1 Under $y = X\beta + \varepsilon$ with ε as above, we have

1. $\hat{\beta}, \tilde{\sigma}^2$ are UMVUE
2. $\hat{\beta} \sim N(\beta, \sigma^2(x^T\Sigma^{-1}X)^{-1})$
3. $\hat{\sigma}^2 \sim \sigma^2(n-p)^{-1}\chi^2_{n-p}$
4. $\tilde{\sigma}^2 \perp\!\!\!\perp \hat{\beta}$

All theories developed previously for $\varepsilon \sim N(0, \sigma^2 I_n)$ can be generalized here in a straightforward manner.



2. General Linear Hypothesis & Simultaneous Confi

2.1 General Linear Model

Definition 2.1.1 — General Linear Models. General Linear Models are the same as linear Gaussian Model, except it is stated in a coordinate-free way.

3. Mutiway ANOVA

3.1 Overview

- Orthogonal design
- Additive 2 way ANOVA
- simultaneous intervals
- nonadditive
- decomposition of sum of squares
- Latin square
- nested design

4. Nonorthogonal Design

4.1 Overview

- $\bar{X}_{\substack{\vdots \\ i}} - \bar{X}_{\substack{:\cdot \\ i}}$

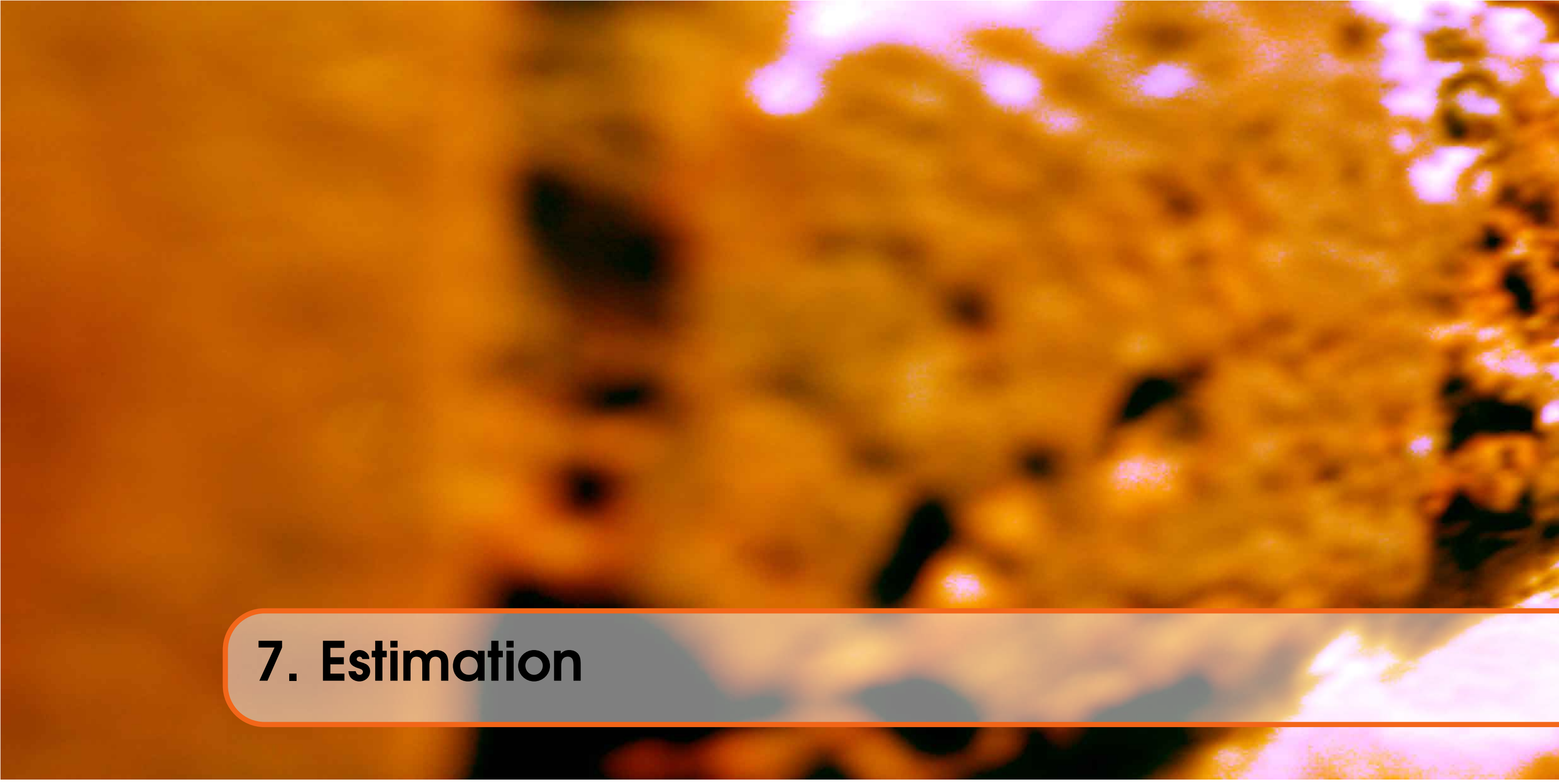
5. Random Effects Model

5.1 Overview

Part Two

6. Basic Concepts

6.1 Overview



7. Estimation

7.1 Overview

8. Inference

8.1 Overview

- deviance <-> sum of squares

9. Residuals

9.1 Overview

10. Cetegorical Prediction

10.1 Overview

11. Some Important GLM

11.1 Overview

12. Multivariate GLM

12.1 Overview

Part Three

13. Principle Componant Analysis

13.1 Overview

14. Canonical Correlation Analysis

14.1 Overview

15. Independent Componant Analysis

15.1 Overview

Index