



Linear Models

STAT 551

Course Notes by Meredith Bartley



Copyright © 2013 John Smith

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, March 2013

Contents

I	Part One	
1	Linear Regression	9
1.1	Projection in Euclidean Space	9
1.2	Cochran's Theorem	16
1.3	Gaussian Linear Regresson Model	17
1.4	Statistical Inference for β , σ^2	20
1.5	Delete One Prediciton	22
1.6	Residuals	24
1.7	Influence and Cook's Distance	24
1.8	Orthogonal Decomposition	25
1.9	Lack of Fit Test	27
1.10	Explicit Intercept	30
1.11	R^2	33
1.12	Multicollinearity	33
1.13	Variable Selection	34
1.14	Non iid Linear Regression	40
2	General Linear Hypothesis & Simultaneous Confidence Intervals	43
2.1	General Linear Model	43
2.2	Hypothesis Testing	45
2.3	Scheffe's Simutaneous Confidence Intervals	46

2.4	Coordinate Version of SCI	50
2.5	Bonferroni's SCI	52
3	One-Way ANOVA	55
3.1	ANOVA Model and Test Statistic	55
3.2	Scheffe's SCI	57
3.3	Bonferonni SCI	59
4	Mutiway ANOVA	61
4.1	Orthogonal Design	61
4.2	Two-Way ANOVA (without Interactions)	63
4.3	Testing Hypotheses	66
4.4	Scheffe's SCI	69
4.5	Non-additive 2-way ANOVA (with Interactions)	70
4.6	Scheffe SCI for 2-Way ANOVA with Interactions	73
4.7	Latin Squares	74
4.8	Orthogonal Nested Model	79

II

Part Two

5	Random Effects Model	85
5.1	Introduction to Random Effects	85
5.2	Sampling Distributions	86
5.3	Restricted MLE - REMLE	89
5.4	Unbalanced Case of One Way Random Effect ANOVA	92
5.5	Balance, Nested Random Effect Model	94
5.6	Nested Mixed Effect Model	98
5.7	Mean Parameterization of Link Function	100
5.8	Bringing in the Predictors	101
5.9	Dispersion Parameters	102
6	Estimation in GLM	105
6.1	Overview	105
6.2	Estimation of ϕ	108
6.3	Statistical Inference for Generalized Linear Model	109
7	Omitted	113
8	Omitted	115
9	Statistically Inference for GLM	117
9.1	Asymptotic Distribution	117

9.2	Estimation of Asymptotic Varariance of $(\hat{\beta})$ and its Confidence Interval	117
9.3	Deviance Function	117
9.4	Residuals	119
9.4.1	Pearson's Residual	119
9.4.2	Anscombe Residual	120
9.5	Deviance Residual	120
10	Omitted - adding to make next chapter 11	123
11	Nature of Predictors	125
11.1	Link to ANOVA	125
11.2	Analysis of Deviance (ANODev)	127
11.3	Numerical Prediction	128
11.4	Testing Hypothesis	129
12	Two Special Cases of GLM	131
12.1	Logistic Regression	131

Part One

1	Linear Regression	9
1.1	Projection in Euclidean Space	
1.2	Cochran's Theorem	
1.3	Gaussian Linear Regression Model	
1.4	Statistical Inference for β , σ^2	
1.5	Delete One Prediction	
1.6	Residuals	
1.7	Influence and Cook's Distance	
1.8	Orthogonal Decomposition	
1.9	Lack of Fit Test	
1.10	Explicit Intercept	
1.11	R^2	
1.12	Multicollinearity	
1.13	Variable Selection	
1.14	Non iid Linear Regression	
2	General Linear Hypothesis & Simultaneous Confidence Intervals	43
2.1	General Linear Model	
2.2	Hypothesis Testing	
2.3	Scheffe's Simultaneous Confidence Intervals	
2.4	Coordinate Version of SCI	
2.5	Bonferroni's SCI	
3	One-Way ANOVA	55
3.1	ANOVA Model and Test Statistic	
3.2	Scheffe's SCI	
3.3	Bonferroni SCI	
4	Multway ANOVA	61
4.1	Orthogonal Design	
4.2	Two-Way ANOVA (without Interactions)	
4.3	Testing Hypotheses	
4.4	Scheffe's SCI	
4.5	Non-additive 2-way ANOVA (with Interactions)	
4.6	Scheffe SCI for 2-Way ANOVA with Interactions	
4.7	Latin Squares	
4.8	Orthogonal Nested Model	

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 .

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

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

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

$$a^T b = \langle a, b \rangle$$

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 $\{\mathbb{R}^P, \langle a, b \rangle\}$

General Inner Product

Let $\Sigma \in \mathbb{R}^{P \times P}$ set of all $P \times P$ 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.

$$a^T \Sigma b = \langle a, b \rangle_{\Sigma}$$

$$a^T b = a^T I b = \langle a, b \rangle_I$$

$\{\mathbb{R}^P, \langle, \rangle_{\Sigma}\}$ is a more general inner product space.

Linear Transformation

A matrix, $A, \in \mathbb{R}^{P \times P}$ can be viewed as linear transformation

$$T_A : \mathbb{R}^P \rightarrow \mathbb{R}^P, x \mapsto Ax$$



Bing Li will denote T_A as A .

\rightarrow means maps to for a domain.

\mapsto means maps to for a value.

\Rightarrow means implies.

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

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

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

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

$$\langle x, Ty \rangle = \langle Rx, y \rangle$$

(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,

$$\langle x, Ty \rangle = \langle T^*x, y \rangle$$

Derived Facts

$$\begin{aligned} \langle x, Ty \rangle &= \langle T^*, y \rangle \\ &= \langle y, T^*x \rangle \\ &= \langle (T^*)^*y, x \rangle \\ &= \langle x, (T^*)^*y \rangle \end{aligned}$$

(by the definition)
(inner products the order doesn't matter)
(Use the definition again)
(swap order)

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

It is easy to see in our case

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

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 $\langle x, Ty \rangle = \langle Tx, y \rangle$). In the case of \langle, \rangle_Σ ,

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

if

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



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 $A \in \mathbb{R}^{P \times P}$ is symmetric, then there exists **eigenvalue-eigenvector pairs**. $(\lambda_1, v_1), \dots, (\lambda_P, v_P)$ such that $v_1 \perp \dots \perp v_P$. Orthogonal basis (ONB) such that

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

More generally, if A is a linear operator in \mathcal{H} (finite dimensional inner product such as $(\mathbb{R}^P, \langle, \rangle_\Sigma)$). its eigen pair (linear operator now) (λ, v) is defined by

$$\begin{cases} Av = \lambda v \\ \langle v, v \rangle = 1 \end{cases}$$

Definition 1.1.6 — Orthogonal Basis. In the following, $(\mathbb{R}^P, \langle, \rangle_\Sigma) = \mathcal{H}$ (H for Hilbert)

ONB is defined by:

1. $v_i \perp v_j, \langle v_i, v_j \rangle = 0$
2. $\|v_i\| = 1$
3. $\text{span}\{v_1, \dots, v_P\} = \mathcal{H}$

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

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

Proof. (λ, v) is eigen pair of A , which means

$$Av = \lambda v$$

$$\langle v, v \rangle = 1$$

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

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



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

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

$$A\Sigma^{-\frac{1}{2}}u = \lambda\Sigma^{-\frac{1}{2}}u$$

$$\Sigma^{-\frac{1}{2}}u = \lambda u$$

So, (λ, v) is an eigen pair of A in $(\mathbb{R}, <, >_\Sigma) \Leftrightarrow (\lambda, u)$ '...' of $\Sigma^{\frac{1}{2}}A\Sigma^{-\frac{1}{2}}$ in $(\mathbb{R}, <, >_I)$.

Note that, A is self adjoint in $(\mathbb{R}, <, >_\Sigma)$. So, $A = \Sigma^{-1}A^T\Sigma$

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

Note: $\Sigma^{\frac{1}{2}}A\Sigma^{-\frac{1}{2}}$ 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 $\Sigma^{\frac{1}{2}}A\Sigma^{-\frac{1}{2}}$.

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

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

■

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 ($P = P^*$).

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

Proof. Take $x \in \ker(A) (\Rightarrow Ax = 0)$.

$$\begin{aligned}\forall y \in \text{ran}(A^*), x \perp y \\ \Rightarrow x \perp y \forall y = A^*z, z \in \mathbb{R}^P\end{aligned}$$

Hence,

$$\begin{aligned}\langle x, y \rangle &= \langle x, A^*z \rangle \\ &= \langle Ax, z \rangle \\ &= \langle 0, z \rangle \\ &= 0\end{aligned}$$

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

Or vice versa.

■

Friday August 26

 \perp means orthogonal complement.

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

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

$$\langle v, w \rangle = 0 \forall w \in \mathcal{S}$$

$$= \{v \in \mathbb{R}^P, \langle v, w \rangle = 0 \forall w \in \mathcal{S}\}$$

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

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

Theorem 1.1.3 If P is a projection, then

1. $Pv = v, \forall v \in \text{ran}(P)$
 2. $Pv = 0, \forall v \perp \text{ran}(P)$
 3. If Q is another projections such that the $\text{ran}(Q) = \text{ran}(P)$ 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 $v \in \text{ran}(P)$. Since $P^2 = P$ (idempotent) then
 $P^2v = Pv$

$$\Rightarrow P^2v - Pv = 0$$

$$\Rightarrow P(Pv - v) = 0$$

$$\Rightarrow Pv - v \in \ker(P)$$

$$\Rightarrow Pv - v \perp \text{ran}(P)$$

$$\Rightarrow \langle Pv - v, Pv - v \rangle = 0$$

$$\Rightarrow \|Pv - v\| = 0$$

$$\Rightarrow Pv - v = 0$$

$$\Rightarrow Pv = v$$

2. If

$$v \perp \text{ran}(P)$$

$$\Rightarrow v \in \ker(P)$$

$$\Rightarrow Pv = 0$$

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

$$Qv = v = Pv \quad (\forall v \in \mathcal{S})$$

$$Qv = 0 = Pv$$

$$Qv = Pv \quad \forall v \in \mathcal{S}$$

$$Q = P$$

■

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

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

Then,

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

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 $x, y \in \mathbb{R}^P$

$$\begin{aligned}
\langle x, Ay \rangle &= x^T \Sigma v (v^T \Sigma v)^{-1} v^T \Sigma y \\
&= (v (v^T \Sigma v)^{-1} v^T \Sigma x)^T \Sigma y \\
&= \langle Ax, y \rangle
\end{aligned}$$

3. $\text{ran}(A) = \mathcal{S}$?Let $x \in \mathbb{R}^P$.

$$Ax = v (v^T \Sigma v)^{-1} v^T \Sigma x \in \text{span}(v) = \mathcal{S}$$

So let $x \in \mathcal{S}$,

$$x \in \text{ran}(v)$$

$$x = vy$$

for some $y \in \mathbb{R}^P$

$$= v (v^T \Sigma v)^{-1} v^T \Sigma vy$$

$$\in \text{ran}(A)$$

So, $\mathcal{S} \subseteq \text{ran}(A)$ and then $\mathcal{S} = \text{ran}(A)$. ■

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

In the following, let $I : \mathbb{R}^P \rightarrow \mathbb{R}^P$ be the identity mapping. ($x \mapsto x$)

Let \mathcal{S} be a subspace in \mathbb{R}^P .

Let $Q_{\mathcal{S}}(\Sigma) = I - P_{\mathcal{S}}(\Sigma)$

Proposition 1.2 $Q_{\mathcal{S}}(\Sigma) = P_{\mathcal{S}^\perp}(\Sigma)$

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

1. Idempotent

$$\begin{aligned}
Q_{\mathcal{S}}^2(\Sigma) &= Q_{\mathcal{S}}(\Sigma) Q_{\mathcal{S}}(\Sigma) \\
&= (I - P_{\mathcal{S}}(\Sigma))(I - P_{\mathcal{S}}(\Sigma)) \\
&= I - P_{\mathcal{S}}(\Sigma) - P_{\mathcal{S}}(\Sigma) + P_{\mathcal{S}} P_{\mathcal{S}} \\
&= Q_{\mathcal{S}}(\Sigma)
\end{aligned}$$

2. Self-adjoint

$$x, y \in \mathbb{R}^P$$

$$\begin{aligned}
\langle x, Q_{\mathcal{S}}(\Sigma)y \rangle &= \langle x, (I - P_{\mathcal{S}}(\Sigma))y \rangle \\
&= \langle x, y \rangle - \langle x, P_{\mathcal{S}}(\Sigma)y \rangle \\
&= \langle x, y \rangle - \langle P_{\mathcal{S}}(\Sigma)x, y \rangle \\
&= \langle (I - P_{\mathcal{S}}(\Sigma))x, y \rangle \\
&= \langle Q_{\mathcal{S}}(\Sigma)x, y \rangle
\end{aligned}$$

3. Range

$$\text{ran}(Q_{\mathcal{S}}(\Sigma)) = \mathcal{S}^\perp. \text{ Take } x \perp \mathcal{S} = \text{ran}(P_{\mathcal{S}}(\Sigma))^\perp = \ker(P_{\mathcal{S}}(\Sigma)).$$

$$\Rightarrow P_{\mathcal{S}}(\Sigma) = 0$$

$$\Rightarrow Q_{\mathcal{S}}(\Sigma)x = x - P_{\mathcal{S}}(\Sigma)x = x$$

$$X \in \text{ran}(Q_{\mathcal{S}}(\Sigma))$$

$$\Rightarrow \mathcal{S}^{\perp} \subseteq \text{ran}(Q_{\mathcal{S}}(\Sigma))$$

Take $x \in \text{ran}(Q_{\mathcal{S}}(\Sigma))$, $\forall y \in \mathcal{S} = \text{ran}(P_{\mathcal{S}}(\Sigma))$

$$y = P_{\mathcal{S}}(\Sigma)z \text{ for some } z \in \mathbb{R}^P$$

$$\langle x, y \rangle = \langle x, P_{\mathcal{S}}(\Sigma)z \rangle = \langle P_{\mathcal{S}}(\Sigma)x, z \rangle = 0$$

$$\Rightarrow x \in \mathcal{S}^{\perp}$$

$$\Rightarrow \text{ran}(Q_{\mathcal{S}}(\Sigma)) = \mathcal{S}^{\perp}$$

■

1.2 Cochran's Theorem

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

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

Proof. λ is eigenvalue of A .

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

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

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 $V^T P V \sim \sigma^2 \chi_S^2$ where $\text{df} = \text{rank}(P)$.

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

$$A R A^T$$

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

Then,

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

Let,

$$Z = R A^T V$$

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 distributed as $N(0, 1)$ and the rest are zero and they are independent. So,

$$Z^T Z \sim \chi_S^2 = V^T P V$$

■

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

$$\langle a, b \rangle_{\Sigma^{-1}} = a^T \Sigma^{-1} b$$

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

Then,

$$\|P_{\mathcal{S}}(\Sigma^{-1})X\|_{\Sigma^{-1}}^2 \sim \chi_r^2$$

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

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

$$\begin{aligned} \|P_{\mathcal{S}}(\Sigma^{-1})X\|_{\Sigma^{-1}}^2 &= \langle P_{\mathcal{S}}(\Sigma^{-1})X, P_{\mathcal{S}}(\Sigma^{-1})X \rangle \\ &= X^T P_{\mathcal{S}}(\Sigma^{-1}) \Sigma^{-1} P_{\mathcal{S}}(\Sigma^{-1}) X \\ &= X^T (V(V^T \Sigma^{-1} V)^{-1} V^T \Sigma^{-1})^T \Sigma^{-1} (V(V^T \Sigma^{-1} V)^{-1} V^T \Sigma^{-1}) X \\ &= X^T \Sigma^{-1} V (V^T \Sigma^{-1} V)^{-1} V^T \Sigma^{-1} V (V^T \Sigma^{-1} V)^{-1} V^T \Sigma^{-1} X \\ &= (\Sigma^{-\frac{1}{2}} X)^T [\Sigma^{-\frac{1}{2}} V (V^T \Sigma^{-1} V)^{-1} (\Sigma^{-\frac{1}{2}} V)^T] (\Sigma^{-\frac{1}{2}} X) \end{aligned}$$

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 respect to I_P -inner product (idempotent, self adjoint, YES).

By Lemme 1.1,

$$(*) \sim \chi_r^2$$

■

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

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

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

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

Then,

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

1.3 Gaussian Linear Regresson Model

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

$$X = \begin{pmatrix} x_{11} & \dots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nP} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

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

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

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

$$y \sim N_p(X\beta, \sigma^2 I_n)$$

Multivariate Normal Density

$$y \sim N(\mu, \Sigma)$$

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

In our case,

$$\Sigma = \sigma^2 I_n$$

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

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,

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

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

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

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

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

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,

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

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

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

$$\|Q_X y\|^2 \sim \chi_{(n-p)}^2$$

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

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

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

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

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

$$P_{Xy} \perp\!\!\!\perp Q_{Xy} \quad (\text{due to normality})$$

$$\hat{\beta} \leftrightarrow P_{Xy}$$

$$\hat{\sigma}^2 \text{ is a function of } Q_{Xy}, \text{ so } \hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$$

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

$$H_0 : \beta_1 = \beta_{i0}$$

$$\text{Let } M = (X^T X)^{-1}.$$

Then,

$$\hat{\beta} \sim N(\beta, \sigma^2 M)$$

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

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

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

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

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

Reject H_0 if

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

Recall,

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

$$y \sim \chi_r^2$$

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

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

Power at β_{i1}

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

So,

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

(alternative distribution 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 β_i .

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

1.5 Delete One Prediction

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. Projecting 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 $A \in \mathbb{R}^{P \times P}$ is a symmetric, nonsingular matrix. and $v \in \mathbb{R}^P$.

Then,

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

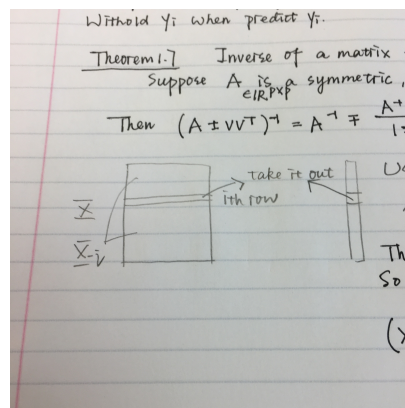


Figure 1.1: Theorem 1.7 Visualization

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

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

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

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

 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.

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

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

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

1.6 Residuals

- Residual

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

- Standardized Residual

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

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

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

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

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

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

- Standardized residual

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

- Prediction Error Sum of Squares (PRESS) Residual

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

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

- Standardized PRESS Error

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

1.7 Influence and Cook's Distance

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

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

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

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

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

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?)

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

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

1.8 Orthogonal Decomposition

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

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

R \leq is subspace
 \subseteq is a subset

For

$$\mathcal{S}_1 \leq \mathcal{S}_1 \mathcal{S}_2 \leq \mathcal{S}$$

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

Suppose $\mathcal{S}_1, \mathcal{S}_2 \leq \mathcal{S}$,

$$\mathcal{S}_1 + \mathcal{S}_2 = \mathcal{S}, \mathcal{S}_1 \perp \mathcal{S}_2$$

then,

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

is called an orthogonal decomposition of \mathcal{S}

In this case,

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

More generally,

Definition 1.8.1 — Orthogonal Decomposition (O.D.). Let $\mathcal{S}_1, \dots, \mathcal{S}_k$ be subspaces of \mathcal{S} such that

$$1. \mathcal{S}_1, \dots, \mathcal{S}_k = \{v_1 + \dots + v_k : v_1 \in \mathcal{S}_1, \dots, v_k \in \mathcal{S}_k\}$$

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

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

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

$$v_1 + \dots + v_k$$

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

Wednesday September 7

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

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

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

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

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

Proposition 1.7 - Orthogonal Decomposition and Projection Consider a Hilbert Space, $\mathcal{H} = \{\mathbb{R}^n, \langle, \rangle_A\}$,

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

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

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

$$P_{\mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_k}(A) = P_{\mathcal{S}_1}(A) + \dots + P_{\mathcal{S}_k}(A)$$

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

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

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

Let $\mathcal{H} = \{\langle, \rangle_{\Sigma^{-1}}\}$. Suppose $\mathcal{S}_1, \dots, \mathcal{S}_k, \mathcal{S} \leq \mathcal{H}$ such that $\mathcal{S} = \mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_k$.

Let

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

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

Then,

$$1. w = w_1 + \dots + w_k$$

$$2. w_1 \perp \dots \perp w_k$$

$$3. w_i \sim \chi_{r_i}^2$$

$$w \sim \chi_r^2$$

where r_i is the $\dim(\mathcal{S}_i)$, r is the $\dim(\mathcal{S})$, and $r = r_1 + \dots + r_k$.

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

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

1.9 Lack of Fit Test

Goodness of Fit

At each x_i you have multiple observations, say y_{i1}, \dots, 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 $y_{i1} \dots y_{jm}$ for each X_i .

In matrix form,

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

So, let N denote a full sample size.

$$N = m_1 + \dots + m_n$$

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

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

So,

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

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

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

$$H_0 : E(y_i) = 1_{m_i}X_i^T \beta$$

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

We are testing whether the arbitrary means, μ_1, \dots, μ_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 ,


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

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

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

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

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

 Above used the fact that $\text{span}(AB) \subseteq \text{span}(A)$

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

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

Go back to lack of fit,

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

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

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

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

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

Now,

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

But,

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \in \mathcal{S}_2$$

and,

$$(1_{m_1} \oplus \cdots \oplus 1_{m_n}) \underline{\mu}$$

$$\text{Var}(P_{\mathcal{S}_3 \ominus \mathcal{S}_2} y) = P_{\mathcal{S}_3 \ominus \mathcal{S}_2} \text{Var}(y) P_{\mathcal{S}_3 \ominus \mathcal{S}_2} = \sigma^2 P_{\mathcal{S}_3 \ominus \mathcal{S}_2}^2 = \sigma^2 P_{\mathcal{S}_3 \ominus \mathcal{S}_2}$$

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

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

Similarly,

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

which under H_0 is,

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

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

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

By Chochran's Theorem:

Under H_0 ,

$$\|P_{\mathcal{S}_3 \ominus \mathcal{S}_2} y\|^2 \sim \chi_{(N-n)}^2$$

$$\|P_{\mathcal{S}_2 \ominus \mathcal{S}_1} y\|^2 \sim \chi_{(n-p)}^2$$

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

So our lack of fit test is:

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

1.10 Explicit Intercept

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

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

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

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

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

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

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

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

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

which is a mean projection. (?)

$$Q_{1_n} = I_n - P_{1_n} \quad (\text{projection on } 1_n^\perp)$$

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

Decompose X:

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

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

So we do least squares of

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

and minimize this over all $\eta^* \in \mathbb{R}^{P \times 1}$

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

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

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

Monday September 12

So

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

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

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

For statistical inference, we want to make a decomposition of \mathbb{R}^n .

Let, $\mathcal{S}_1 = \text{span}(1_n)$, $\mathcal{S}_2 = \text{span}(1_n, X)$, $\mathcal{S}_3 = \mathbb{R}^n$.

Then,

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

Then,

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

Or,

$$SSTotal = SSE_{Error} + SS_{Regression}$$

We may compute these terms,

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

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

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



$$\dim(\mathcal{S}_3) = n$$

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

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

$$H_0 : \beta = 0$$

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

Distributions

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

$$E(\hat{\beta}) = (X^T QX)^{-1} X^T Q(1_n \alpha + X\beta) = (X^T QX)^{-1} X^T QX\beta = \beta$$

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

$$\hat{\alpha} = \hat{\alpha}^* - X^T \hat{\beta}$$

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

$$\text{Var}(\hat{\alpha}) = \text{Var}(\hat{\alpha}^*) + \text{Var}(\bar{X}^T \hat{\beta}) = \text{Var}(\bar{y}) + \text{Var}(\bar{X}^T \hat{\beta}) = \frac{\sigma^2}{n} + \sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X}$$

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

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

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} + \sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} & -\sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} \\ -\sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} & \frac{\sigma^2}{n} + \sigma^2 \bar{X}^T (X^T QX)^{-1} \bar{X} \end{pmatrix}\right)$$

Estimate σ^2

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

So,

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

Thus,

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

Theorem 1.10.1 Under the explicit intercept model,

1. $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ is UMVUE of $(\alpha, \beta, \sigma^2)$ by Lehmann-Sheffe.

2.

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

3. $(n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi_{(n-p-1)}^2$

4. $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \perp \hat{\sigma}^2$

1.11 R^2

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

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

But we know that,

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

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

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

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

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

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

1.12 Multicollinearity

Wednesday September 14

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

$$X = (C_1, \dots, C_p) = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}$$

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

We have C_1, C_2, C_3 .

$$C_1 = a$$

$$C_2 = 2a$$

$$C_3 = b$$

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

β_1 & β_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 $C_1 = aC_2$ 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'.

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

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

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

1.13 Variable Selection

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

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 selection.

Mallow's C_p criterion

The fundamental issue is variable selection, 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 $A \subset \{1, \dots, p\}$,

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

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 $\text{span}(X_A)$, $\text{span}(X_A)^\perp$. 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 $A \subset \{1, \dots, p\}$.

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

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

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

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

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

$$E(U) = e$$

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

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

Also, $W \sim \chi_{(r)}^2(\delta)$ 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

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

and

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

So,

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

So

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

Thus,

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

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 unknown, 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

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

So to recap,

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

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, \dots, X_n , a sample of independent random vectors with joint density $f_\theta(x_1, \dots, x_n)$.

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

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

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

Let Θ_M be the parameter space, corresponding to M .

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

Of course we also have Θ_{M_0} .

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

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

Then,

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

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

Use them

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

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

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

Specialized to Gaussian Linear Regression Model

Here there is no variable selection,

$$\theta = \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}$$

$$\Theta \subset \mathbb{R}^{P+1} \text{ (M is in } \Theta)$$

$$\beta \in B$$

$$\text{subset } \mathbb{R}^P \text{ (A is in B)}$$

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 $X_A = \{C_i : i \in A\}$ 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 $(\hat{\beta}_A, \hat{\sigma}_A^2)$,

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

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

It is equivalent to 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, $(X_1, \dots, X_n) = \mathcal{X}$ (again, generic X, not in regression context). An estimator,

$$\mathcal{X} \rightarrow \Theta$$

Variable selection,

$$\hat{A} : \mathcal{X} \rightarrow 2^{\{1, \dots, P\}}$$

that is to say,

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

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

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

where A_0 is the true action set.

Next, BIC in variable selection consistency.

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 $\frac{a_n}{b_n} \rightarrow 0$ as $n \rightarrow \infty$.

- **Example 1.1**
- $a_n \prec 1 \Leftrightarrow a_n \rightarrow 0$
 - $a_n \prec n \Leftrightarrow \frac{a_n}{n} \rightarrow 0$

$$a_n \succ 1$$

$$\Rightarrow 1 \prec a_n$$

- **Example 1.2**
- $\Rightarrow \frac{1}{a_n} \rightarrow 0$

$$\Rightarrow a_n \rightarrow \infty$$

- $n^{\frac{1}{2}}$

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. $\Theta_{M_0} \subseteq \Theta_M$

$$2(\log L_M - \log L_{M_0}) \rightarrow^{\mathcal{D}} \chi^2_{(\#M - \#M_0)}$$

Here, recall,

$$L_M = \sup_{\theta \in \Theta} f_{\theta}(x_1, \dots, x_n)$$

2. If $\Theta_M \subseteq \Theta_{M_0}$ then,

$$n^{-1} 2(\log L_M - \log L_{M_0}) \rightarrow^P 2(\sup_{\theta \in \Theta} E \log f_{\theta}(x_1, \dots, x_n) - E \log f_{\theta_0}(x_1, \dots, x_n))$$

Moreover, if $M \subset M_0$ then

$$\lim_{n \rightarrow \infty} (2(\sup_{\theta \in \Theta} E \log f_{\theta}(x_1, \dots, x_n) - E \log f_{\theta_0}(x_1, \dots, x_n))) < 0$$

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 $\hat{M} = \arg \min_{M \in \{1, 2, \dots, p\}} BIC(M)$ 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 $c(n)(\#M - \#M_0) \rightarrow \infty$

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

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

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

So,

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

Case 2 $M \subseteq M_0$

$n^{-1}2(\log L_{M_0} - \log L_M) \rightarrow c(n) > 0$

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

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

So again,

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

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

■

1.14 Non iid Linear Regression

Suppose

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

where $\varepsilon \sim N(0, \sigma^2 \Sigma)$ with arbitrary 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 $\|Q_X(\Sigma^{-1})y\|_{\Sigma^{-1}}^2 \sim \sigma^2 \chi_{n-p}^2$, so now we have

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

so the unbiased estimator is,

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

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

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

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 Conf

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 or geometric way.

Let $\mathcal{S} \leq \mathbb{R}^N$.

A general linear model gives,

$$y \sim N(\mu, \sigma^2 I_N)$$

where $\mu \in \mathcal{S}$.

If we take X to be a basis matrix of \mathcal{S} , that is $\text{span}(X) = \mathcal{S}$, then we have

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

the same as before. (because $\mu \in \mathcal{S}, \text{span}(x) = \mathcal{S} \Rightarrow \mu = X\beta$ for some β)

The MLE can be derived in a similar way.

Wednesday September 21

MLE for μ

Likelihood:

$$\frac{1}{(2\pi)^{\frac{n}{2}} [\det(\sigma^2 I_N)]^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2} \|y - \mu\|^2}$$

maximize this over $\mu \in \mathcal{S}, \sigma^2 > 0$.

First we maximize over $\mu \in \mathcal{S}$ equivalent to minimizing $\|y - \mu\|^2$.

$$\begin{aligned}
\|y - \mu\|^2 &= \|y - P_{\mathcal{J}}y + P_{\mathcal{J}}y - \mu\|^2 \\
&= \|y - P_{\mathcal{J}}y\|^2 - 2\langle y - P_{\mathcal{J}}y, P_{\mathcal{J}}y - \mu \rangle + \|P_{\mathcal{J}}y - \mu\|^2 \\
&= \dots 2\langle y - P_{\mathcal{J}}y, P_{\mathcal{J}}y - \mu \rangle = 0 \\
&= \|y - P_{\mathcal{J}}y\|^2 + \|P_{\mathcal{J}}y - \mu\|^2 \\
&\Rightarrow \hat{m}u = P_{\mathcal{J}}y
\end{aligned}$$

By the argument exactly like the coordinate case, we can show that

$$\hat{\sigma}_{MLE}^2 = \frac{y^T Q_{\mathcal{J}} y}{N}$$

Because

$$\frac{Q_{\mathcal{J}} y}{\sigma^2} \sim N(0, Q_{\mathcal{J}})$$

we know that

$$\frac{y^T Q_{\mathcal{J}} y}{\sigma^2} \sim \chi_{N-p}^2$$

$$E\left(\frac{y^T Q_{\mathcal{J}} y}{\sigma^2}\right) = N - p$$

So, an unbiased estimator for σ^2 would be

$$\hat{\sigma}^2 = \frac{y^T Q_{\mathcal{J}} y}{N - p}$$

and an unbiased estimator for μ is

$$E(\hat{\mu}) = P_{\mathcal{J}}\mu = \mu$$

What is the complete and sufficient statistic? We may use results from exponential family.

$$\exp\left(-\frac{1}{2}\|y - \mu\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}(\|P_{\mathcal{J}}y\|^2 + \|Q_{\mathcal{J}}y\|^2) + \frac{1}{2\sigma^2} \langle P_{\mathcal{J}}y, \mu \rangle\right) \exp(\theta_1 t_1 + \theta_2 t_2)$$

So, complete and sufficient statistic would be

$$(\|P_{\mathcal{J}}y\|^2 + \|Q_{\mathcal{J}}y\|^2, P_{\mathcal{J}}y) \leftrightarrow (\|Q_{\mathcal{J}}y\|^2, P_{\mathcal{J}}y)$$

By Lehmann-Scheffe,

Theorem 2.1.1 Under $y \sim N(\mu, \sigma^2 I_n)$, $\mu \in \mathcal{J}$ we have

1. $\hat{\mu} \perp \hat{\sigma}^2$, $\hat{\mu} \sim N(\mu, \sigma^2 P_{\mathcal{J}})$, $(N - p)\hat{\sigma}^2 \sim \sigma^2 \chi_{N-p}^2$
2. $(\hat{\mu}, \hat{\sigma}^2)$ is UMVUE

2.2 Hypothesis Testing

$$y \sim N(\mu, \sigma^2 I_N), \mu \in \mathcal{S}$$

Consider,

$$\mathcal{S}' \leq \mathcal{S} \leq \mathbb{R}^N$$

$$\dim(\mathcal{S}') = k, \dim(\mathcal{S}) = p, \quad k \leq p$$

We have

$$\mathbb{R}^n \ominus \mathcal{S}' = (\mathbb{R}^N \ominus \mathcal{S}) \oplus (\mathcal{S} \ominus \mathcal{S}')$$

So,

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}'} y\|^2 = \|P_{\mathcal{S} \ominus \mathcal{S}'} y\|^2 + \|P_{\mathbb{R}^N \ominus \mathcal{S}} y\|^2$$

Want to get,

$$H_0 : \mu \in \mathcal{S}'$$

So

$$\begin{aligned} \mu \in \mathcal{S}' &\Leftrightarrow \|P_{\mathcal{S} \ominus \mathcal{S}'} \mu\| = 0 \\ &\Leftrightarrow \|P_{\mathcal{S} \ominus \mathcal{S}'} y\| \text{ is small} \end{aligned}$$

Thus we can use,

$$\frac{\|P_{\mathcal{S} \ominus \mathcal{S}'} y\|^2 / (p - k)}{\|P_{\mathbb{R}^N \ominus \mathcal{S}} y\|^2 / (N - p)} \sim F_{p-k, N-p}$$

Both Lack of Fit and explicit intercept can be written in general linear model.

In Lack of Fit,

$$\mathcal{S} = \text{span}(1_{m_1} \oplus \cdots \oplus 1_{m_n})$$

$$\mathcal{S}' = \text{span}((1_{m_1} \oplus \cdots \oplus 1_{m_n})x)$$

In Explicit Intercept,

$$\mathcal{S} = \text{span}(1_{m_1} : X)$$

$$\mathcal{S}' = \text{span}((1_n)$$

Alternative Distribution

When H_0 is not true it means that $\mu \notin \mathcal{S}'$. Then we know that $\|P_{\mathcal{S} \ominus \mathcal{S}'} y\|^2$ is not small. In fact, $E(P_{\mathcal{S} \ominus \mathcal{S}'} y) = P_{\mathcal{S} \ominus \mathcal{S}'} \mu \neq 0$

$$\text{Var}(P_{\mathcal{S} \ominus \mathcal{S}'} y) = \sigma^2 P_{\mathcal{S} \ominus \mathcal{S}'}$$

$$P_{\mathcal{J} \ominus \mathcal{J}'} y \sim N(P_{\mathcal{J} \ominus \mathcal{J}'} \mu, \sigma^2 P_{\mathcal{J} \ominus \mathcal{J}'})$$

$$\|P_{\mathcal{J} \ominus \mathcal{J}'} y\|^2 \sim \chi_{p-k}^2(\|P_{\mathcal{J} \ominus \mathcal{J}'} \mu\|^2)$$

Definition 2.2.1 If

$$X \sim \chi_{r_1}^2(s)$$

$$Y \sim \chi_{r_2}^2$$

$$X \perp\!\!\!\perp Y$$

then,

$$\frac{X/r_1}{Y/r_2} \sim F_{r_1, r_2}(s)$$

We still have that

$$\mathcal{J} \ominus \mathcal{J}' \perp \mathbb{R}^N \ominus \mathcal{J}$$

$$\text{Cov}(P_{\mathcal{J} \ominus \mathcal{J}'} y, P_{\mathbb{R}^N \ominus \mathcal{J}} y) = 0$$

$$P_{\mathcal{J} \ominus \mathcal{J}'} y \perp\!\!\!\perp P_{\mathbb{R}^N \ominus \mathcal{J}} y$$

These are still true even though $\mu \notin \mathcal{J}'$. Why? Because $\mu \in \mathcal{J}$.

$$\|P_{\mathcal{J} \ominus \mathcal{J}'} y\|^2 \perp\!\!\!\perp \|P_{\mathbb{R}^N \ominus \mathcal{J}} y\|^2$$

so to compute power:

$$\frac{\|P_{\mathcal{J} \ominus \mathcal{J}'} y\|^2 (p-k)}{\|P_{\mathbb{R}^N \ominus \mathcal{J}} y\|^2 / (N-p)} \sim F_{p-k, N-p}(\|P_{\mathcal{J} \ominus \mathcal{J}'} \mu\|^2)$$

Friday September 23

2.3 Scheffe's Simultaneous Confidence Intervals

It's conceptually easy to construct individuals C.I.

$$P_\theta(\theta \in C(X)) = 1 - \alpha$$

We want to construct C.I. for several infinite sets of parameters. Then the width of the confidence interval has to be adjusted (wider).

1. Boneroni adjustment
2. Sheffe's approach

Simultaneous C.I.

Say we have a set of parameters.

$$\{\theta_\lambda : \lambda \in \Lambda\}$$

Simultaneous C.I. for $\{\theta_\lambda : \lambda \in \Lambda\}$ is a family of subsets of Θ (the parameter space).

$$\{C_\lambda(x) \subset \Theta : \lambda \in \Lambda\}$$

Where $C_\lambda(x)$ is a set in Θ depending only on x , that is it's a statistic.

$$x \rightarrow 2^\Theta$$

This collection is called **Simultaneous Confidence Region** if

$$P(C_\lambda(x) \text{ covers } \theta_\lambda \forall \lambda \in \Lambda) = 1 - \alpha$$

In general linear model where,

$$y = \mu + \varepsilon, \quad \mu \in \mathcal{S}$$

$$\mathcal{S}' \leq \mathcal{S}$$

$$\mathcal{S} \leq \mathbb{R}^N$$

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

we are interested in constructing S.C.I. for

$$\{C^T P_{\mathcal{S} \ominus \mathcal{S}'} \mu : C \in \mathbb{R}^N\}$$

That is we want,

$$C_c(X) : c \in \mathbb{R}^N$$

such that

$$P(C^T P_{\mathcal{S} \ominus \mathcal{S}'} \mu \in C_c(X) \forall c \in \mathbb{R}^N) = 1 - \alpha$$

or equivalently,

$$P(d^T \mu \in C_d(X) \forall d \in \mathcal{S} \ominus \mathcal{S}') = 1 - \alpha$$

Here, d has a special name.

Definition 2.3.1 — Contrast. Suppose $\mathcal{S} \leq \mathcal{S}' \leq \mathbb{R}^N$. A **contrast** for hypothesis,

$$H_0 : \mu \in \mathcal{S}'$$

$$H_1 : \mu \in \mathcal{S} \ominus \mathcal{S}'$$

is $d^T \mu$ where $d \in \mathcal{S} \ominus \mathcal{S}'$.

Pivotal Quantity

Definition 2.3.2 — Pivotal Quantity.

$$X \sim P_\theta$$

A **pivotal quantity** is a function $T(X, \theta)$ such that its distribution under P_θ is independent of θ . It's almost like an ancillary statistics, except it contains the parameter θ .

Theorem 2.3.1 Suppose $y \sim N(\mu, \sigma^2 I_N)$ and that $\mu \in \mathcal{S}$.

Let

$$\delta = P_{\mathcal{S}^\perp} y$$

Let

$$F(\delta) = \frac{\|P_{\mathcal{S}^\perp} y - \delta\|^2}{(p-k)\hat{\sigma}^2}$$

where $p = \dim(\mathcal{S})$ and $k = \dim(\mathcal{S}^\perp)$

Then,

$$F(\delta) \sim F_{p-k, N-p}$$

This implies that $F(\delta)$ is a pivotal quantity because its distribution doesn't depend on δ .

Proof. We have

$$P_{\mathcal{S}^\perp} y - \delta = P_{\mathcal{S}^\perp} y - P_{\mathcal{S}^\perp} \mu$$

$$P_{\mathcal{S}^\perp} (y - \mu) \sim N(0, P_{\mathcal{S}^\perp})$$

So,

$$\|P_{\mathcal{S}^\perp} y - \delta\|^2 \sim \chi_{p-k}^2$$

But we also know that

$$P_{\mathbb{R}^N} y \perp P_{\mathcal{S}^\perp} y$$

Recall,

$$\|P_{\mathbb{R}^N} y\|^2 \sim \sigma^2 \chi_{N-p}^2$$

Take the ratio and use the definition of $\hat{\sigma}^2$ to complete the Theorem. ■

Equivalence Between Confidence Region and Hypothesis Test

Consider the hypothesis test,

$$H_0 : \{\theta\}$$

$$H_1 : \{\theta\}^C$$

at level α .

A acceptance region is any subset $A_\theta \subseteq \mathcal{X}$ (the sample space) so that

$$P_\theta(X \in A(\theta)) = 1 - \alpha$$

A acceptance region is a mapping from the parameter space to a subset of \mathcal{X} .

$$\Theta \rightarrow 2^{\mathcal{X}}, \theta \mapsto A(\theta)$$

On the other hand, for each $x \in \mathcal{X}$ let

$$C(x) = \{\theta : H_0 \text{ is accepted.}\}$$

$$C(x) = \{\theta : x \in A(\theta)\}$$

By this definition,

$$P_{\theta}(\theta \in C(x)) = P_{\theta}(x \in A(\theta)) = 1 - \alpha$$

As an illustration of this equivalence, let's construct a Confidence Region for $P_{\mathcal{S} \ominus \mathcal{S}'} \mu$.

$$H_0 : P_{\mathcal{S} \ominus \mathcal{S}'} \mu = \delta$$

$$H_1 : P_{\mathcal{S} \ominus \mathcal{S}'} \mu \neq \delta$$

Suppose we use the acceptance rule.

$$F(\delta) < F_{p-k, N-p}(1 - \alpha)$$

Then the $(1 - \alpha) \times 100\%$ Confidence Region for δ is

$$\{\delta : F(\delta) < F_{p-k, N-p}(1 - \alpha)\}$$

we can evaluate if θ in the set by computing this above criteria.

Monday September 26

SCI for contrasts.

We are interested $C_d(x) : d \in \mathbb{R}^N$ such that $P(d^T \mu \in C_d(X, d \in \mathcal{S} \ominus \mathcal{S}'))$.

It turns out we can only do this because we can use Cauchy-Schwarz Inequality for a uniform bound.

As before, $\hat{\delta} = P_{\mathcal{S} \ominus \mathcal{S}'} y$, $\delta = P_{\mathcal{S} \ominus \mathcal{S}'} \mu$. By CS,

$$|d^T (\hat{\delta} - \delta)|^2 \leq \|d\|^2 - \|\hat{\delta} - \delta\|^2$$

But we know (from last lecture) that

$$\|\hat{\delta} - \delta\|^2 \sim \sigma^2 \chi_{p-q}^2$$

$$\frac{\|P_{\mathbb{R}^N \ominus \mathcal{S}} y\|^2}{\sigma^2} \sim \chi_{N-p}^2$$

and also that they are independent.

$$\hat{\sigma}^2 = \frac{\|P_{\mathbb{R}^N \ominus \mathcal{S}} y\|^2}{N - p}$$

$$\frac{\|P_{\mathcal{S} \ominus \mathcal{S}'} y - \delta\|^2}{\sigma^2(p - q)} \sim F_{p-q, N-p}$$

$$P\left(\|\hat{\delta} - \delta\|^2 \leq \hat{\sigma}^2(p-q)F_{p-q, N-p}(1-\alpha)\right) = 1 - \alpha$$

Using CS \neq ,

$$P\left(\frac{d^T(\hat{\delta} - \delta)^2}{\|d\|^2} \leq \hat{\sigma}^2(p-q)F_{p-q, N-p}(1-\alpha)\right) \geq 1 - \alpha$$

$$\Leftrightarrow P(d^T \delta \in d^T \hat{\sigma}^2 \pm \|d\| \hat{\sigma} \sqrt{(p-q)F_{p-q, N-p}(1-\alpha)}) \geq 1 - \alpha$$

In geometric terms,

$$d^T P_{\mathcal{S} \ominus \mathcal{S}'} y \pm \|d\| \|P_{\mathbb{R}^N \ominus \mathcal{S}} y\| \sqrt{\frac{\dim \mathcal{S} - \dim \mathcal{S}'}{N - \mathcal{S}} F_{\dim \mathcal{S} - \dim \mathcal{S}', N - \mathcal{S}}(1-\alpha)} (***)$$

2.4 Coordinate Version of SCI

Instead of $y \sim N(\mu, \sigma^2 I_n)$, $\mu \in \mathcal{S}$, we can see that $y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2, I_n)$.

Typically we want to construct simultaneous SC for β_1, \dots, β_p , that is,

$$P(\beta_1 \in C_1(x), \dots, \beta_p \in C_p(x)) \geq 1 - \alpha$$

If we can construct SCI function for all $S^T \beta$, $S \in \mathbb{R}^p$ then we can solve the problem because,

$$S = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

More generally, we may want to construct SCI for β_1, \dots, β_q for some q less than p . In this case we need SCI in the form

$$\left\{ S^T (I_q; 0) \beta : S \in \mathbb{R}^q \right\} (**)$$

instead of $S^T \beta$.

So we construct general SCI,

$$S^T A^T \beta$$

where $A \in \mathbb{R}^{p \times q}$. This would accommodate both (*), (**). This can be cast into the general SCI, then (***). Need \mathcal{S}' , \mathcal{S} , \mathbb{R}^N .

Consider $H_0 : A^T \beta = 0 \Leftrightarrow A^T (X^T X)^{-1} X^T \mu = 0 \Leftrightarrow C^T \mu = 0 \Leftrightarrow \mu \in \text{span}(x) \ominus \text{span}(C)$.

So

$$\mathcal{S}' = \text{span}(x) \ominus \text{span}(C)$$

$$\mathcal{S} = \text{span}(x)$$

$$\mathcal{S} \ominus \mathcal{S}' = \text{span}(C)$$

We are now testing $H_0 : \mu \in \mathcal{S}'$ against $H_1 : \mu \in \mathcal{S}$.

Compute specific expressions in (***) (general SCI form),

$$\begin{aligned} P_{\mathcal{S} \ominus \mathcal{S}'} y &= P_{\text{span}(C)} y = P_C y \\ &= C(C^T C)^{-1} C^T y \\ &= X(X^T X)^{-1} A [A^T (X^T X)^{-1} A]^{-1} A^T (X^T X)^{-1} X^T y \\ d \in \mathcal{S} \ominus \mathcal{S}' &= \text{span}(c) = Cs = X(X^T X)^{-1} As \end{aligned}$$

$$d^T P_{\mathcal{S} \ominus \mathcal{S}'} y = d^T P_C y = S^T A^T (X^T X)^{-1} X^T y = S^T A^T (X^T X)^{-1} X^T y = S^T A^T \hat{\beta}$$

Recall that,

$$\dim(\mathcal{S}) = p, \dim(\mathcal{S}') = p - q$$

so plug everything into (***) to get, $(1 - \alpha)$ -level SCI (conservative: Prob $\geq (1 - \alpha)$),

$$S^T A^T \hat{\beta} \pm \|X(X^T X)^{-1} As\| \hat{\sigma} \sqrt{q F_{q, N-p}(1 - \alpha)}$$

To summarize, the whole procedure, suppose we wanted to construct SCI for β_1, \dots, β_p or β_1, \dots, β_q or $\beta_1 - \beta_2, \beta_2 - \beta_3, \dots$.

Then, we let A be a matrix such that $\text{span}(A)$ encloses (minimally) the above ranges of β s, so that

$$\beta_j \text{ or } \beta_1 - \beta_2 = S^T A^T \beta$$

For example for β_1, \dots, β_p ,

$$A = I_p$$

$$A^T = (I_q; 0)$$

For $\beta_1 - \beta_2, \beta_2 - \beta_3, \dots$,

$$A = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \\ \vdots & 0 \\ 0 & \vdots \\ 0 & 0 \end{pmatrix}$$

The idea of Scheffe SCI is to enlarge the set to linear space. That is, even though you only want

$$e_1^T \beta, \dots, e_q^T \beta$$

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

you still construct sCI for more parameters than you want.

$$\{s^T(I_q; 0)\beta : A \in \mathcal{R}^q\}$$

The disadvantage is that for smaller number of contrasts, this tends to be conservative. Here, you may use Bonferroni's Method. The advantage is that the width is fixed regardless of number of contrasts, as long as they are the same subspace.

Wednesday September 28

Last time, we covered Scheffe's SCI; one feature is that SCI for infinitely many linear combinations. If you just want to SCI for a few linear combinations then conservative approach is needed. In this case, Bonferroni SCI is preferred, but Bonferroni SCI gets wider and wider as the number of parameters increases. So Scheffe's is preferred for large number of parameters.

2.5 Bonferroni's SCI

Suppose we want to SCI for $\theta_1, \dots, \theta_k$ (that is we want $C_1(X), \dots, C_k(X)$) such that

$$P(\theta_1 \in C_1(X), \dots, \theta_k \in C_k(X)) \geq 1 - \alpha$$

Let

So if you let $P(A_i) = 1 - \frac{\alpha}{k}$ then

$$P\left(\bigcap_{i=1}^n A_i\right) \geq 1 - k\left(1 - \left(1 - \frac{\alpha}{k}\right)\right) = 1 - \alpha$$

So the $(1 - \alpha)$ -level SCI is simply $(1 - \frac{\alpha}{k})$ -level ICI. (Recall, S - Simultaneous, I - Individual).

As $k \rightarrow 0$, $1 - \frac{\alpha}{k} \rightarrow 1$ (which is disadvantageous for large k). Specialize to linear regression, where we want Bonferroni SCI for $\alpha_1^T \beta, \dots, \alpha_q^T \beta$ where β is as in,

$$y = X\beta - \varepsilon$$

and

$$\alpha_1, \dots, \alpha_q \in \mathbb{R}^p$$

We know that

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

and

$$(N - p)\hat{\sigma}^2 \sim \sigma^2 \chi_{N-p}^2$$

where note we are using the biased $\hat{\sigma}^2$. Finally we have

$$\frac{(N - p)\hat{\sigma}^2}{\sigma^2} \sim \chi_{N-p}^2 \quad (*)$$

Ultimately, with the α we obtain,

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

$$\frac{\alpha_k^T \hat{\beta} - \alpha_k^T \beta}{\sigma \sqrt{\alpha_k^T (X^T X)^{-1} \alpha_k}} \sim N(0, 1) \quad (**)$$

Note that (*) and (**) are independent.

Studentize:

3. One-Way ANOVA

3.1 ANOVA Model and Test Statistic

This is a special case of general linear model.

$$y_{ij} \sim N(\mu_i, \sigma^2) \quad j = 1, \dots, n_i,$$

All y_{ij} are independent.

In matrix form we get,

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_i} \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_i} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \vdots \\ \mu_p \\ \vdots \\ \mu_p \end{pmatrix}, \sigma^2 I_n \right)$$

$$\mu = \begin{pmatrix} 1_{n_1} & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & 1_{n_p} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

In this case,

$$Y \sim N(\mu, \sigma^2 I), \mu \in \mathcal{S}$$

$$\mathcal{S} = \text{span} \begin{pmatrix} 1_{n_1} & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & 1_{n_p} \end{pmatrix}$$

Because, μ is defined as above.

So we have a special case of General linear model.

Once you know this form, we know everything: decomposition of Sum of Squares, F-Statistics, Scheffe's, Bonferroni's, etc. We just need to specialize the formulae using a specific model. The same general principle applies to all the linear models yet to come. Here, we want to test

$$H_0 : \mu_1 = \dots = \mu_p$$

or

$$H_0 : \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \end{pmatrix}$$

$$\mathcal{S}' = \text{span}(1_N)$$

So our hypotheses are now

$$H_0 : \mu \in \mathcal{S}'$$

$$H_1 : \mu \notin \mathcal{S}'$$

F-Statistic

$$F = \frac{\|P_{\mathcal{S} \ominus \mathcal{S}'} Y\|^2 / \dim(\mathcal{S} \ominus \mathcal{S}')}{\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 / \dim(\mathbb{R}^N \ominus \mathcal{S})} \sim F_{\dim(\mathcal{S} \ominus \mathcal{S}'), \dim(\mathbb{R}^N \ominus \mathcal{S})}$$

Here we'd reject if

$$F > F_{\dim(\mathcal{S} \ominus \mathcal{S}'), \dim(\mathbb{R}^N \ominus \mathcal{S})}(1 - \alpha)$$

In one-way ANOVA we have some special names.

$$\|P_{\mathcal{S} \ominus \mathcal{S}'} Y\|^2 \leftarrow \text{SSH}$$

$$\dim(\mathcal{S} \ominus \mathcal{S}') = p - 1$$

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 \leftarrow \text{SSE}$$

$$\dim(\mathbb{R}^N \ominus \mathcal{S}) = N - p$$

$$\frac{\text{SSH}}{p - 1} = \text{MSH}, \frac{\text{SSE}}{N - p} = \text{MSE}$$

$$F = \frac{\text{MSH}}{\text{MSE}}$$

Due to the simple structure of $X = \begin{pmatrix} 1_{n_1} & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & 1_{n_p} \end{pmatrix}$ we can see (in HW) that SSH and SSE have special forms.

$$SSH = \sum_{i=1}^P n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$\text{where } \bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \text{ and } \bar{Y}_{..}^2 = \frac{1}{N} \sum_{i=1}^P \sum_{j=1}^{n_i} Y_{ij}.$$

$$SSE = \sum_{i=1}^P \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$$

$$SST = \|P_{\mathbb{R}^N \ominus \mathcal{S}'} Y\|^2$$

Friday September 30

3.2 Scheffe's SCI

In this case recall,

$$\mathcal{S} = \text{span}\{1_{n_1} \oplus \cdots \oplus 1_{n_p}\}$$

$$\mathcal{S}' = \text{span}\{1_N\}$$

The general form from Scheffe's for GLH (General Linear Hypothesis).

$$H_0 : \mu \in \mathcal{S}'$$

$$H_1 : \mu \in \mathcal{S}$$

$$d^T \hat{\mu} \pm \hat{\sigma} \|d\| \sqrt{(p-1)F_{p-1, N-p}(1-\alpha)}$$

We want $d \in \mathcal{S} \ominus \mathcal{S}'$.

$$X = \begin{pmatrix} 1_{n_1} & & \\ & \ddots & \\ & & 1_{n_p} \end{pmatrix} = 1_{n_1} \oplus \cdots \oplus 1_{n_p}$$

$$d \in \mathcal{S} = \text{span} X$$

$$d = XC$$

$$d \perp \mathcal{S}'$$

This means that $d^T 1_N = 0$ and $C^T X^T 1_N = 0$.

$$(C_1, \dots, C_p) \begin{pmatrix} 1_{n_1}^T & & \\ & \ddots & \\ & & 1_{n_p}^T \end{pmatrix} \begin{pmatrix} 1_{n_1} \\ \vdots \\ 1_{n_p} \end{pmatrix} = C_1 n_1 + \dots + C_p n_p$$

So d is of the form XC where $n_1 C_1 + \dots + n_p C_p = 0$.

$$\hat{\mu} = \begin{pmatrix} 1_{n_1} \bar{Y}_{1\cdot} \\ \vdots \\ 1_{n_p} \bar{Y}_{p\cdot} \end{pmatrix} = X \begin{pmatrix} \bar{Y}_{1\cdot} \\ \vdots \\ \bar{Y}_{p\cdot} \end{pmatrix} = P_{1_{n_1} \oplus \dots \oplus 1_{n_p}} Y = P_{\mathcal{J}} Y$$

$$d^T \hat{\mu} = C^T (X^T X) \begin{pmatrix} \bar{Y}_{1\cdot} \\ \vdots \\ \bar{Y}_{p\cdot} \end{pmatrix} = C^T \begin{pmatrix} n_1 & & \\ & \ddots & \\ & & n_p \end{pmatrix} \begin{pmatrix} \bar{Y}_{1\cdot} \\ \vdots \\ \bar{Y}_{p\cdot} \end{pmatrix} = c_1 n_1 \bar{Y}_{1\cdot} + \dots + C_p n_p \bar{Y}_{p\cdot}$$

$$\|d\| = \sqrt{d^T d} = \sqrt{c^T X^T X C} = \sqrt{\sum_{i=1}^p C_i^2 n_i}$$

We have $\hat{\sigma}^2$ as before.

$$\frac{\|P_{\mathbb{R}^N \ominus \mathcal{J}} Y\|^2}{N-p} = \frac{\sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2}{N-p} = MSE$$

A alternative parameterization $t_i = n_i c_i$.

In this from, let $\hat{v} = \begin{pmatrix} \bar{Y}_{1\cdot} \\ \vdots \\ \bar{Y}_{p\cdot} \end{pmatrix}$

The SCI is

$$t^T \hat{v} \pm \hat{\sigma} \sqrt{\sigma(t_i^2/n_i)(p-1)F_{p-1, N-p}(1-\alpha)}$$

Commonly need contrasts are $\mu_i - \mu_i', \mu_1 - 3\mu_2 + 2\mu_3$.

The usually mention hypothesis,

$$\mu_1 = \dots = \mu_p, \quad \mu_i = \mu_i' \quad \forall i \neq i'$$

Test equality of subsets of mean (as given before) just use

$$t = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Where 1 is the i^{th} entry and -1 is the i'^{th} entry.

We can do several of these.

$$t^T \hat{v} = \bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}$$

$$\sum \frac{t_i^2}{n_i} = \frac{1}{n_i} - \frac{1}{n_{i'}}$$

3.3 Bonferonni SCI

Say we want to construct SCI.

$$\{t_1^T v, \dots, t_q^T v\}$$

As discussed before, SCI,

$$t_k^T \hat{v} \pm t_{N-p} \left(1 - \frac{\alpha}{2q}\right) \hat{\sigma} \sqrt{t_k^T (X^T X)^{-1} t_k}$$

Here,

$$X^T X = \begin{pmatrix} n_1 & & \\ & \ddots & \\ & & n_p \end{pmatrix}$$

and,

$$t_k^T (X^T X)^{-1} t_k = \sum_{i=1}^p \frac{1}{n_i} t_{ki}^2$$

4. Mutiway ANOVA

4.1 Orthogonal Design

Recall, the General Linear Model: $Y \sim N(\mu, \sigma^2 I), \mu \in \mathcal{S}, \mathcal{S} \leq \mathbb{R}^N$.

Orthogonal esigns mean that \mathcal{S} can be decomposed in to $\mathcal{S} = \mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_v$. Later on:

\mathcal{S}_1 factor A

\mathcal{S}_2 factor B

\mathcal{S}_3 interacts

In this case,

$$\hat{\mu} = P_{\mathcal{S}} Y = P_{\mathcal{S}_1} Y + \cdots + P_{\mathcal{S}_v} Y$$

$$\mu = P_{\mathcal{S}} \mu = P_{\mathcal{S}_1} \mu + \cdots + P_{\mathcal{S}_v} \mu = v_1 \in \mathcal{S}_1 + \cdots + v_v \in \mathcal{S}_v$$

Unique Decomposition is covered in Chapter 1.

Suppose we want to test that there is no interaction:

$$H_0 : v_i = 0$$

$$H_1 : v_i \neq 0$$

This is equivalent to

$$H_0 : \mu \in \oplus_{j \neq i} \mathcal{S}_j$$

$$H_1 : \mu \in \mathcal{S}$$

In this case, $\mathcal{S}' = \oplus_{j \neq i} \mathcal{S}_j, \mathcal{S} = \oplus_{j=1}^v \mathcal{S}_j$.

So $\mathcal{S} \ominus \mathcal{S}' = \mathcal{S}_i$. Thus the F-statistics for GLH is

$$\frac{||P_{\mathcal{S}_i} Y||^2 / d_i}{||P_{\mathbb{R}^N \ominus \mathcal{S}} Y||^2 / (N - p)} \stackrel{H_0}{\sim} F_{d_i, N-p}$$

where $d_i = \dim(\mathcal{S}_i)$

If we don't have orthogonality, suppose we have GLM,

$$Y = X_1 \beta_1 + \dots + X_p \beta_p + \varepsilon$$

Letting $\beta_i = 0$, simply means

$$\mu \in \text{span}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p) (= \mathcal{S}^i)$$

Since X_1, \dots, X_p are not ortogonal, this is not $\text{span}(X) \ominus \text{span}(X_i)$ so we have that $\mathcal{S} \ominus \mathcal{S}' \neq \text{span}(X_i)$.

Moreover, in the orthogonal case, the point estimation of β_i relies entirely on (Y, X_i) .

Screening (?) on Variable Selection

In the orthogonal case, they are the same. We must demonstrate this.

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

If $X_1 \perp \dots \perp X_p$,

$$X^T X = \begin{pmatrix} X_1^T X_1 & & 0 \\ & \ddots & \\ 0 & & X_p^T X_p \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \frac{X_1^T Y}{X_1^T X_1} \\ \vdots \\ \frac{X_p^T Y}{X_p^T X_p} \end{pmatrix}$$

$$\hat{\beta}_i = \frac{X_i^T Y}{X_i^T X_i}$$

So to get β_i you simply regress Y on X_i which doesn't involve the other column.

Another effect of orthogonlity is you can decompose sum of squares additively.

$$||P_{\mathcal{S}} Y||^2 = ||P_{\mathcal{S}_1} Y||^2 + \dots ||P_{\mathcal{S}_p} Y||^2$$

You can tabulate this nicely in ANOVA table.

If no orthogonality then you don't report $||P_{\mathcal{S}_i} Y||^2$ as the sum of squares associated with β_i .

The correct sum of squares,

$$||P_{\text{span}(X) \ominus \text{span}(X_{-i})} Y||^2$$

where $X_{-1} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p)$.

So even though you can still tabulate these substitutes they don't sum up to $\|P_{\mathcal{S}}Y\|^2$.

This is less meaningful than ANOVA table. In generalized linear models we have to be content with this imperfect ANOVA.

Monday October 2

4.2 Two-Way ANOVA (without Interactions)

Model - special case of general linear model

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2 I_N)$$

$$k = 1, \dots, n_{ij}$$

$$j = 1, \dots, c$$

$$i = 1, \dots, r$$

Assume (for now),

$$\mu_{ij} = \gamma_i + \tau_j$$

Orthogonal Design

To ensure orthogonal design, $n_{ij} = p_i q_j$, where p_i, q_j are positive integers. We will show that this condition ensures orthogonality.

Notation 4.1 (Dot Notation). n_i indicates that the sec

Apply this notation to both numbers and matrix/vector notation.

$$n_{ij} = p_i q_j$$

$$n_{i\cdot} = p_i q_{\cdot}$$

$$n_{\cdot j} = p_{\cdot} q_j$$

$$n_{\cdot\cdot} = p_{\cdot} q_{\cdot}$$

Orthogonal design means

$$\frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} = n_{ij}$$

Matrix Notation

$$\mu = \begin{pmatrix} \mu_{11} \\ \vdots \\ \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{12} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1_{n_{11}}(\gamma_1 + \tau_1) \\ \vdots \\ 1_{n_{ij}}(\gamma_i + \tau_j) \\ \vdots \\ 1_{n_{rc}}(\gamma_r + \tau_c) \end{pmatrix} \in \mathbb{R}^{n..}$$

$$\{1_{n_{ij}}(\gamma_i + \tau_j), i = 1, \dots, r; j = 1, \dots, c\}$$

Note that the latter index varies first.

Now, we introduce a systematic δ notation that will also be useful later.

Notation 4.2. For

$$(i, j) = \{1, \dots, r\} \times \{1, \dots, c\}$$

$$(u, v) = \{1, \dots, r\} \times \{1, \dots, c\}$$

$$w = 1, \dots, n_{uv}$$

We have

$$\delta_{uvw}^{ij} = \begin{cases} 1 & \text{if } (u, v) = (i, j) \\ 0 & \text{else} \end{cases}$$

That means there are n_{uv} 1's in this vector and $n.. - n_{uv}$ 0's.

Let $E_{ij} = \{\delta_{uvw}^{ij} : u = 1, \dots, r, v = 1, \dots, c, w = 1, \dots, n_{uv}\}$ where the last index runs first.

Let

$$E_{i.} = \sum_{j=1}^n E_{ij} \tag{4.1}$$

$$E_{.j} = \sum_{i=1}^n E_{ij} \tag{4.2}$$

$$\tag{4.3}$$

$$R = (E_{i.}, \dots, E_{r.})$$

$$C = (E_{.j}, \dots, E_{.c})$$

In this notation, $\mu = \{\mu_{ij} : k = 1, \dots, n_i, j = 1, \dots, c, j = 1, \dots, r\}$.

$$\mu = E_{1.}\gamma_1 + \dots + E_{r.}\gamma_r + E_{.1}\tau_1 + \dots + E_{.c}\tau_c = R\gamma + C\tau$$

In the above, both γ and τ are column vectors.

Overall Mean

$$\mu = P_{1_n}\mu + Q_{1_n}\mu$$

Where,

$$P_{1_n} = \frac{1_N 1_N^T}{1_N^T 1_N}$$

$$Q_{1_n} = I_N - P_{1_n}$$

So we may write μ as,

$$\begin{aligned}\mu &= P_{1_n}\mu + Q_{1_n}(R\gamma + C\tau) \\ &= P_{1_n}\mu + \alpha + \beta\end{aligned}$$

It turns out that $\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}} = n_{ij}$ ensures that $\alpha^T \beta = 0$. We must show this.

$$\begin{aligned}\alpha^T \beta &= (Q_{1_n}R\gamma)^T (Q_{1_n}C\tau) \\ &= \gamma^T (R^T \frac{Q_{1_n}Q_{1_n}}{Q_{1_n}} C) \tau \\ R^T Q_{1_n} C &= R^T (I_N - \frac{1_N 1_N^T}{1_N^T 1_N}) C \\ &= R^T C - \frac{R^T 1_N 1_N^T C}{N} \\ R^T C &= \begin{pmatrix} E_{1\cdot}^T \\ \vdots \\ E_{r\cdot}^T \end{pmatrix} (E_{1\cdot}, \vdots, E_{r\cdot})\end{aligned}$$

Now look at,

$$\begin{aligned}E_{i\cdot}^T E_{\cdot j} &= \left(\sum_{s=1}^c E_{is} \right)^T \left(\sum_{t=1}^r E_{tj} \right) \\ &= \sum_{s=1}^c \sum_{t=1}^r E_{is}^T E_{tj}\end{aligned}$$

Hence,

$$\begin{aligned}R^T C &= \begin{pmatrix} n_{11} & \dots & n_{1c} \\ \vdots & & \vdots \\ n_{r1} & \dots & n_{rc} \end{pmatrix} \\ R^T 1_N &= \begin{pmatrix} E_{1\cdot}^T \\ \vdots \\ E_{r\cdot}^T \end{pmatrix} 1_N = \begin{pmatrix} n_{1\cdot} \\ \vdots \\ n_{r\cdot} \end{pmatrix}\end{aligned}$$

$$1_N^T C = (n_{\cdot 1}, \dots, n_{\cdot c})$$

$$\frac{R^T 1_N 1_N^T C}{N} = \begin{pmatrix} \frac{n_{1\cdot} n_{\cdot 1}}{n_{\cdot\cdot}} & \dots & \dots \\ & \ddots & \\ \vdots & \dots & \ddots \end{pmatrix}$$

So by orthogonal density,

$$R^T C - \frac{R^T 1_N 1_N^T C}{N} = 0, \quad \alpha^T \beta = 0$$

Wednesday October 5

Geometric Representation

$$\mathcal{S}_1 = \text{span}(1_N)$$

$$\mathcal{S}_2 = \text{span}(R) \ominus \text{span}(1_N)$$

$$\mathcal{S}_3 = \text{span}(C) \ominus \text{span}(1_N)$$

By construction, $\mathcal{S}_1 \perp \mathcal{S}_2, \mathcal{S}_1 \perp \mathcal{S}_3$.

Since,

$$\mathcal{S}_2 = Q_{1_N} \text{span}(R) = \text{span}(Q_{1_N} R)$$

$$\mathcal{S}_3 = \text{span}(1_N C)$$

we know that $R^T Q_{1_N} C = 0$ by orthogonal design.

So $\mathcal{S}_2 \perp \mathcal{S}_3$.

So it is justified to write

$$\mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3$$

which means that $\mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3$ and $\mathcal{S}_1 \perp \mathcal{S}_2 \perp \mathcal{S}_3$

4.3 Testing Hypotheses

$$\mu_{ij} = \gamma_i + \tau_j = (\bar{\gamma} + \bar{\tau}) + (\gamma_i - \bar{\gamma}) + (\tau_j - \bar{\tau})$$

Where $(\bar{\gamma} + \bar{\tau}) \in \mathcal{S}_1, (\gamma_i - \bar{\gamma}) \in \mathcal{S}_2, (\tau_j - \bar{\tau}) \in \mathcal{S}_3$.

$$\mu_{ij} = w + \alpha_i + \beta_j$$

$$\begin{aligned} \sum \alpha_i = 0 &\Leftrightarrow \mathcal{S}_2 \perp \mathcal{S}_1 \\ \sum \beta_j = 0 &\Leftrightarrow \mathcal{S}_3 \perp \mathcal{S}_1 \end{aligned}$$

We can test these hypotheses (among many other hypotheses),

I $H_0 : \alpha_1 = \dots = \alpha_r = 0$ (no row effect)

II $H_1 : \beta_1 = \dots = \beta_c = 0$ (no column effect)

For Hypothesis I,

$$\mu \perp \mathcal{S} \Leftrightarrow \mu \in \mathcal{S}_1 \oplus \mathcal{S}_3$$

$$\mathcal{S}' = \mathcal{S}_1 \oplus \mathcal{S}_3, \mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3$$

So specialized on GLM,

F-Statistic:

$$\frac{\|P_{\mathcal{S} \ominus \mathcal{S}'} Y\|^2 / \dim(\mathcal{S} \ominus \mathcal{S}')}{\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 / \dim(\mathbb{R}^N \ominus \mathcal{S})} \sim F_{\dim(\mathcal{S} \ominus \mathcal{S}'), \dim(\mathbb{R}^N \ominus \mathcal{S})}$$

Specialize to our own context:

$$P_{\mathcal{S} \ominus \mathcal{S}'} Y = \{\bar{Y}_{..} - Y_{...}; k = 1, \dots, n_{ij}; i = 1, \dots, r; j = 1, \dots, c\}$$

$$\|P_{\mathcal{S} \ominus \mathcal{S}'} Y\|^2 = \sum_{i=1}^r n_{i.} (\bar{Y}_{..} - Y_{...})^2$$

$$\dim(\mathcal{S} \ominus \mathcal{S}') = \dim(\mathcal{S}_2) = r - 1$$

$$P_{\mathbb{R}^N \ominus (\mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3)}$$

Now let's take just the subscript and apply results from HW problem (note the \mathcal{S} are not the exact same),

$$\mathbb{R} \ominus (\mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3) = (\mathbb{R}^N \ominus \mathcal{S}_1) \ominus (\mathcal{S}_2 \oplus \mathcal{S}_3)$$

So we may rewrite above and again apply result from homework to get,

$$P_{(\mathbb{R}^N \ominus \mathcal{S}_1) \ominus (\mathcal{S}_2 \oplus \mathcal{S}_3)} = P_{(\mathbb{R}^N \ominus \mathcal{S}_1)} - P_{(\mathcal{S}_2 \oplus \mathcal{S}_3)} = P_{(\mathbb{R}^N \ominus \mathcal{S}_1)} - P_{(\mathcal{S}_2)} - P_{(\mathcal{S}_3)}$$

So,

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 = \|P_{\mathbb{R}^N \ominus \mathcal{S}_1} Y\|^2 - \|P_{\mathcal{S}_2} Y\|^2 - \|P_{\mathcal{S}_3} Y\|^2$$

This is left as HW.

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 = \{Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}, \dots\}$$

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}_1} Y\|^2 = \{Y_{ijk} - \bar{Y}_{...}, k = \dots\}$$

$$\|P_{\mathcal{S}_2} Y\|^2 = \{\bar{Y}_{i..} - \bar{Y}_{...}, \dots\}$$

$$\|P_{\mathcal{S}_3} Y\|^2 = \{\bar{Y}_{.j.} - \bar{Y}_{...}, \dots\}$$

So F_I becomes the familiar form,

$$F_I = \frac{MSR}{MSE}$$

$$\begin{aligned}
MSR &= SSR/(r-1) \\
SSR &= \sum_{i=1}^r n_i (\bar{Y}_{j..} - \bar{Y}_{...}) \\
MSE &= SSE/(N-r-c+1) \\
SSE &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})
\end{aligned}$$

$$F_I \sim F_{r-1, N-r-c+1}$$

Similarly, for testing Hypothesis II, still using GHL,

$$\beta_1 = \cdots = \beta_c = 0$$

$$\mu \perp \mathcal{S}_3$$

$$\mu \in \mathcal{S}_1 \oplus \mathcal{S}_2$$

$$\mathcal{S}' = \mathcal{S}_1 \oplus \mathcal{S}_2$$

$$\mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3$$

$$\mathcal{S} \ominus \mathcal{S}' = \mathcal{S}_3$$

$$\|P_{\mathcal{S} \ominus \mathcal{S}'}\|^2 = \|P_{\mathcal{S}_3}\|^2 = \sum_{j=1}^c n_{ij} (\bar{Y}_{.j.} - \bar{Y}_{...})^2 = SSC$$

$$\dim(\mathcal{S}_3) = c-1$$

$$F_{II} = \frac{MSC}{MSE}$$

$MSC = SSC/(c-1)$
 SSC is as above.

$$F_{II} \sim F_{c-1, N-r-c+1}$$

Now we may summarize everything into the ANOVA table.

Friday October 7

4.4 Scheffe's SCI

Recall, the general hypothesis,

$$\begin{aligned} H_0 : \mu &\in \mathcal{S}', \\ H_1 : \mu &\in \mathcal{S} \end{aligned}$$

We want SCI for contrasts,

$$\{c^T \mu : c \in \mathcal{S} \ominus \mathcal{S}'\} = \{c^T \delta : c \in \mathcal{S} \ominus \mathcal{S}'\}$$

where $\delta = P_{\mathcal{S} \ominus \mathcal{S}'} \mu$.

This because

$$c = P_{\mathcal{S} \ominus \mathcal{S}'} c$$

so,

$$C^T \mu = C^T P_{\mathcal{S} \ominus \mathcal{S}'} \mu = C^T \delta$$

We have

$$\begin{aligned} &C^T P_{\mathcal{S} \ominus \mathcal{S}'} Y \pm \|C\| \|P_{\mathcal{S} \ominus \mathcal{S}'} Y\| \\ &\sqrt{\frac{\dim(\mathcal{S} \ominus \mathcal{S}')}{\dim(\mathbb{R}^N \ominus \mathcal{S})} F_{\dim(\mathcal{S} \ominus \mathcal{S}'), \dim(\mathbb{R}^N \ominus \mathcal{S})}} \end{aligned}$$

Here we have two H_0 of interest.

I

$$\alpha_1 = \dots = \alpha_r = 0 \Leftrightarrow \mu \in \mathcal{S}_1 \oplus \mathcal{S}_3 = \mathcal{S}'$$

$$\mathcal{S} \ominus \mathcal{S}' = \mathcal{S}_2$$

So

$$\begin{aligned} \delta &= P_{\mathcal{S}_2} \mu \\ &= \alpha \\ &= \{\alpha_i : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\} \end{aligned}$$

Note that

$$C \in \mathcal{S} \ominus \mathcal{S}' = \mathcal{S}_2$$

So C is of the form

$$C = \{C : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

But we also know that $C \perp 1_N$. Therefore,

$$\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} C_i = \sum_{i=1}^r C_i \sum_{j=1}^c \sum_{k=1}^{n_{ij}} 1 = \sum_{i=1}^r C_i * n_{i.} = 0$$

$$P_{\mathcal{S} \ominus \mathcal{S}'} Y = P_{\mathcal{S}_2} Y = \{\bar{Y}_{i..} - \bar{Y}_{...} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

$$C^T P_{\mathcal{S}_2} Y = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} C_i (\bar{Y}_{i..} - \bar{Y}_{...}) = \sum_{i=1}^r C_i * n_{i.} (\bar{Y}_{i..} - \bar{Y}_{...})$$

So usually we use this alternative parameterization,

$$t_i = n_{i.} C_i$$

$$\text{So } \sum_{i=1}^r t_i = 0.$$

$$C^T P_{\mathcal{S}_2} Y = \sum_{i=1}^r t_i (\bar{Y}_{i..} - \bar{Y}_{...})$$

$$\dim(\mathcal{S} \ominus \mathcal{S}') = \dim(\mathcal{S}_2) = r - 1$$

$$\dim(\mathbb{R}^N \ominus \mathcal{S}) = N - r - c + 1$$

$$\begin{aligned} \|c\|^2 &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} C_i^2 \\ &= \sum_{i=1}^r n_{i.} C_i^2 \\ &= \sum_{i=1}^r \frac{t_i^2}{n_{i.}} \end{aligned}$$

$$P_{\mathbb{R}^N \ominus \mathcal{S}} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$$

So to summarize Scheffe's SCI for H_0 is

$$\sum_i t_i (\bar{Y}_{i..} - \bar{Y}_{...}) \pm \sqrt{\sum_{i=1}^r \frac{t_i^2}{n_{i.}}} \sqrt{\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2} \sqrt{\frac{r-1}{N-r-c+1} F_{r-1, N-r-c+1}(1-\alpha)}$$

$$\text{II } H_0 : \beta_1 = \dots = \beta_c = 0$$

$$\sum_j u_j (\bar{Y}_{.j.} - \bar{Y}_{...}) \pm \sqrt{\sum_{j=1}^c \frac{u_j^2}{n_{.j}}} \sqrt{\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2} \sqrt{\frac{c-1}{N-r-c+1} F_{c-1, N-r-c+1}(1-\alpha)}$$

4.5 Non-additive 2-way ANOVA (with Interactions)

"The whole is greater than the sum of its parts."

In this case, we have,

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$$

where μ_{ij} cannot be decomposed.

So

$$\mu_{ij} = \theta + \alpha_i + \beta_j + \gamma_{ij}$$

$$\begin{aligned} k &= 1, \dots, n_{ij} \\ j &= 1, \dots, c \\ i &= 1, \dots, r \end{aligned}$$

$$\mathcal{S}_1 = \text{span}(\theta : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r)$$

$$\mathcal{S}_2 = \text{span}(\alpha_i : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r)$$

$$\mathcal{S}_3 = \text{span}(\beta_j : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r)$$

$$\mathcal{S}_4 = \text{span}(\gamma_{ij} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r) \ominus (\mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3)$$

with $n_{ij} = \phi_i \varepsilon_j$ (orthogonal design)

$$\mathcal{S}_1 \perp \mathcal{S}_2 \perp \mathcal{S}_3 \perp \mathcal{S}_4$$

So that it is justified to write

$$\mathcal{S}_1 + \dots \mathcal{S}_4 = \mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_4$$

Using matrix notation as before,

$$\mathcal{S}_1 = \text{span}(1_N)$$

$$\mathcal{S}_2 = \text{span}\{E_1 \dots E_r\} \ominus \mathcal{S}_1$$

$$\mathcal{S}_3 = \text{span}\{E_1 \dots E_c\} \ominus \mathcal{S}_1$$

$$\mathcal{S}_4 = \text{span}\{E_{ij} : j = 1, \dots, c; i = 1, \dots, r\} \ominus (\mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3)$$

$$E_{ij} = \{d_{uvw}^{ij} : w = 1, \dots, n_{uv}; v = 1, \dots, c; u = 1, \dots, r\}$$

$$\mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3 \oplus \mathcal{S}_4 = \text{span}\{E_{ij} : j = 1, \dots, c; i = 1, \dots, r\}$$

The projections, $P_{\mathcal{S}_i} Y, i = 1, \dots, 4$ and $P_{\mathbb{R}^N \ominus \mathcal{S}}$ are derived similarly.

Monday October 10

$\mathbb{R}^N \ominus \mathcal{S}$ is the garbage, but it's very useful for testing.

Explicit expression of projections: (check yourself in HW)

$$P_{\mathcal{S}_1} Y = \{\bar{Y}_{\dots} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

$$P_{\mathcal{S}_2}Y = \{\bar{Y}_{i..} - \bar{Y}_{...} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

$$P_{\mathcal{S}_3}Y = \{\bar{Y}_{.j.} - \bar{Y}_{...} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

$$P_{\mathcal{S}_4}Y = \{\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

$$P_{\mathbb{R}^N \ominus \mathcal{S}}Y = \{Y_{ijk} - \{\bar{Y}_{ij.} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

We may extend this to,

$$\|P_{\mathcal{S}_1}Y\|^2 = N\bar{Y}_{...}^2 =$$

$$\|P_{\mathcal{S}_2}Y\|^2 = \sum_{i=1}^r n_{i.}(\{\bar{Y}_{i..} - \bar{Y}_{...}\})^2 = SSR$$

$$\|P_{\mathcal{S}_3}Y\|^2 = \sum_{j=1}^c n_{.j}(\{\bar{Y}_{.j.} - \bar{Y}_{...}\})^2 = SSC$$

$$\|P_{\mathcal{S}_4}Y\|^2 = \sum_{i=1}^r \sum_{j=1}^c n_{ij}(\{\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}\})^2 = SSRC$$

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}}Y\|^2 = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} n_{ij}(Y_{ijk} - \{\bar{Y}_{ij.}\})^2 = SSE$$

$$df(R) = r - 1$$

$$df(C) = c - 1$$

$$df(RC) = rc - r - c + 1$$

How do we get this last degrees of freedom?

$$\dim(\text{span}(E_{ij})) - \dim(\mathcal{S}_1) - \dim(\mathcal{S}_2) - \dim(\mathcal{S}_3)$$

$$df(E) = N - rc$$

We test the following hypotheses.

R effect:

$$H_R : \alpha_1 = \dots = \alpha_r = 0$$

$$H_C : \beta_1 = \dots = \beta_c = 0$$

$$H_{RC} : \gamma_{ij} = 0$$

SSH (where H is null hypothesis),

$$H = R, C, RC$$

We use

$$\frac{MSH}{MSE} = \frac{SSH/df(SSH)}{SSE/df(SSE)} \sim F_{df(SSH), df(SSE)}$$

For example, if $H = RC$,

$$SSH = SSSRC$$

$$df(SSH) = rc - r - c + 1$$

$$\frac{MSRD}{MSE} \sim F_{rc-r-c+1, N-rc}$$

4.6 Scheffe SCI for 2-Way ANOVA with Interactions

Again this depends on which hypothesis (R, C, RC) you are interested in.

For H_{RC}

$$c^T P_{\mathcal{S}_4} Y \pm \|C\| \|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\| - \sqrt{\frac{\dim(\mathcal{S}_4)}{\dim(\mathbb{R}^N \ominus \mathcal{S})} F_{\dim(\mathcal{S}_4), \dim(\mathbb{R}^N \ominus \mathcal{S})}(1-\alpha)}$$

First, figure out specific form of C .

$$C \in \mathcal{S} \ominus \mathcal{S}'$$

$$C \in \mathcal{S} \Rightarrow C = \{C_{ij} : k = 1, \dots, n_{ij}; j = 1, \dots, c; i = 1, \dots, r\}$$

Also,

$$C^T 1_N = 0, C^T E_{i\cdot} = 0, C^T E_{\cdot j} = 0,$$

To show this,

$$\begin{aligned} \sum_i \sum_j \sum_k C_{ij} 1_N &= \sum_i \sum_j C_{ij} \sum_k 1_N \\ &= \sum_i \sum_j C_{ij} n_{ij} \\ &= 0 \end{aligned}$$

An similarly for the other two equations equal to zero.

So if we let $t_{ij} = C_{ij} n_{ij}$ we can see again that summing it over i, j or i , or j all give zero.

Scheffe's SCI for Interactions

$$\sum_{i=1}^r \sum_{j=1}^c (\{\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot}\}) \pm \sqrt{\sum_{i=1}^r \sum_{j=1}^c \frac{t_{ij}^2}{n_{ij}}} \sqrt{\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij\cdot})^2} \sqrt{\frac{rc-r-c+1}{N-rc} F_{rc-r-c+1, N-rc}(1-\alpha)}$$

4.7 Latin Squares

Suppose we have three main effects. In general, if you have three effects you need to make a cube for the design and observations.

Sometimes experiments over the entire cube would require too much time/money/etc. Can we test three effects using two-way table? Intuitively you have to avoid entangling the third effect with the first two effects.

Let A be a finite set, $A = \{a_1, \dots, a_m\}$.

A latin square is a $m \times m$ matrix,

$$L = \begin{pmatrix} b_{11} & \dots & b_{1m} \\ \vdots & & \vdots \\ b_{m1} & \dots & b_{mm} \end{pmatrix}$$

such that

1. Each row is a permutation of $(1, \dots, m)$.
2. Each column is a permutation of $(1, \dots, m)$.

■ **Example 4.1** $m = 3$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{pmatrix}$$

(You may shift each row to the left each time (or right).) ■

Let $k(i, j)$ be the element at the (i, j) th entry.

So, mathematically, a latin square is a mapping

$$K : A \times A \rightarrow A, (i, j) \mapsto k(i, j)$$

such that (1) and (2) are satisfied. The following is the property of Latin Square.

Theorem 4.7.1 Let $K : A \times A \rightarrow A$ be a Latin Square. Let $\eta : A \rightarrow \mathbb{R}, k \mapsto \eta(k)$ be any function. Then

$$\begin{aligned} \sum_i \sum_j \eta(k(i, j)) &= \sum_i \eta(k(i, j)) \quad \forall j \in A \\ &= \sum_j \eta(k(i, j)) \quad \forall i \in A \end{aligned}$$

That is, all row and column totals are the same.

Proof. This is simple because for each i ,

$$k(i, 1), \dots, k(i, m)$$

is a permutation of $1, \dots, m$. Therefore,

$$\eta(k(i, 1)), \dots, \eta(k(i, m))$$

is also a permutation of $\eta(1), \dots, \eta(m)$.

So they sum to the same values, regardless of i .

Thus $\sum_j \eta(k(i, j)) = \text{a constant not dependant on } i$.

Similarly, for $\sum_i \eta(k(i, j))$.

■

Linear Model with Latin Square Design

$$\frac{1}{ij} \sim N(\mu_{ij}, \sigma^2) \quad (\text{independent})$$

where $\mu_{ij} = \delta_i + \varepsilon_j + \eta_{k(i,j)}$ where the η values is the Latin Effect.

Orthogonal Decomposition

Notation 4.3. *Dot Notation. For a latin square,*

$$\{a_{ij} : j = 1, \dots, m; i = 1, \dots, m\}$$

Let,

$$a_{i\cdot} = \sum_j a_{ij}, a_{\cdot j} = \sum_i a_{ij}$$

Let

$$a_{\cdot k} = \sum_{k(i,j)=k} a_{ij}$$

be the sum over all cells whose latin letter is k .

Notation 4.4.

$$d_{uv}^{ij} = \begin{cases} 1 & (u, v) = (i, j) \\ 0 & \text{else} \end{cases}$$

$$E_{ij} = \{d_{uv}^{ij} : v = 1, \dots, m; u = 1, \dots, m\}$$

$$E_{i\cdot} = \sum_{j=1}^m E_{ij}$$

$$E_{\cdot j} = \sum_{i=1}^m E_{ij}$$

$$E_{\cdot k} = \sum_{k(i,j)=k} E_{ij}$$

$$\mathcal{S}_1 = \text{span}(1_N), N = m^2$$

$$\mathcal{S}_2 = \text{span}\{E_{i\cdot}\} \ominus \mathcal{S}_1$$

$$\mathcal{S}_3 = \text{span}(E_{\cdot j}) \ominus \mathcal{S}_1$$

$$\mathcal{S}_4 = \text{span}(E_{\cdot k}) \ominus \mathcal{S}_1$$

$$\mathcal{S}_5 = \mathbb{R}^N \ominus (\mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_4)$$

By construction,

$$\mathcal{S}_2 \perp \mathcal{S}_1, \mathcal{S}_3 \perp \mathcal{S}_1, \mathcal{S}_4 \perp \mathcal{S}_1$$

By Latin Square design we can show that $\mathcal{S}_2 \perp \mathcal{S}_3$ and $\mathcal{S}_2 \perp \mathcal{S}_3 \perp \mathcal{S}_4$.

Only need to check that $\mathcal{S}_2 \perp \mathcal{S}_4$.

Proof in PHOTO

Point Estimation

Let

$$\mathcal{S} = \oplus_{i=1}^4 \mathcal{S}_i$$

$$\mathcal{S}_5 = \mathbb{R}^N \ominus \mathcal{S}$$

$$\mathcal{S}_1 \perp \cdots \perp \mathcal{S}_5$$

$$Y = P_{\mathcal{S}_1} Y + \cdots + P_{\mathcal{S}_5} Y$$

$$\mu = P_{\mathcal{S}_1} \mu + \cdots + P_{\mathcal{S}_5} \mu$$

As before,

$$\begin{aligned} \mathcal{S}_5 &= \mathbb{R}^N \ominus (\mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_4) \\ &= (\mathbb{R}^N \ominus \mathcal{S}_1) \ominus (\mathcal{S}_2 \oplus \cdots \oplus \mathcal{S}_4) \end{aligned}$$

For any vector,

$$a = \{a_{ij} : i = 1, \dots, m; j = 1, \dots, m\}$$

$$P_{\mathcal{S}_1} a = \{\bar{a}_{..} : i = 1, \dots, m; j = 1, \dots, m\}$$

$$P_{\mathcal{S}_2} a = \{\bar{a}_{i.} - \bar{a}_{..} : i = 1, \dots, m; j = 1, \dots, m\}$$

$$P_{\mathcal{S}_3} a = \{\bar{a}_{.j} - \bar{a}_{..} : i = 1, \dots, m; j = 1, \dots, m\}$$

$$P_{\mathcal{S}_4} a = \{\bar{a}_{.k} - \bar{a}_{..} : i = 1, \dots, m; j = 1, \dots, m\}$$

$$P_{\mathcal{S}_5} a = \{a_{ij} - \bar{a}_{..} - (\bar{a}_{i.} - \bar{a}_{..}) - (\bar{a}_{.j} - \bar{a}_{..}) - (\bar{a}_{.k} - \bar{a}_{..}) : i = 1, \dots, m; j = 1, \dots, m\} = \{a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} - \bar{a}_{.k} + 2\bar{a}_{..} : i = 1, \dots,$$

Friday October 14

Clarification of last lecture:

$$P_{\mathcal{S}_4}a = \{\bar{a}_{k.} - a_{..} : i = 1, \dots, m; j = 1, \dots, m\}$$

By $\bar{a}_{k.}$ we mean,

$$\bar{a}_{k.} = \frac{1}{m.} \sum_{\{(i,j):k(i,j)=k\}} a)ij$$

$$(\bar{a}_{k.})_{k=k(i,j)=\bar{a}_{k(i,j)}}.$$

Using the above results and projection,

$$\mu_{ij} = \theta + \alpha_i + \beta_j + \gamma_{k(i,j)}$$

where,

$$\begin{aligned}\theta &= \bar{\mu}_{..} \\ \alpha_i &= \bar{\mu}_{i.} - \bar{\mu}_{..} \\ \beta_j &= \bar{\mu}_{.j} - \bar{\mu}_{..} \\ \gamma_{k(i,j)} &= \bar{\mu}_{k.(i,j)} - \bar{\mu}_{..}\end{aligned}$$

We may estimate these by $P_{\mathcal{S}_i}$ for $i = 1, 2, 3$.

So,

$$\begin{aligned}\bar{\mu}_{..} &\leftarrow \bar{Y}_{..} \\ \bar{\mu}_{i.} &\leftarrow \bar{Y}_{i.} \\ \bar{\mu}_{.j} &\leftarrow \bar{Y}_{.j} \\ \bar{\mu}_{k.(i,j)} &\leftarrow \bar{Y}_{k.(i,j)}\end{aligned}$$

Decomposition of Sum of Squares

$$||P_{\mathcal{S}_1}Y||^2 = m^2 \bar{Y}_{..}^2$$

$$||P_{\mathcal{S}_2}Y||^2 = m \sum_{i=1}^m (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$||P_{\mathcal{S}_3}Y||^2 = m \sum_{j=1}^m (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$||P_{\mathcal{S}_4}Y||^2 = m \sum_{k=1}^m (\bar{Y}_{k.} - \bar{Y}_{..})^2$$

Test Hypothesis

For example, we would want to test H_0 that there is no Latin effect.

$$H_0 : \mu \in \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3 \leftarrow \mathcal{S}'$$

$$H_1 : \mu \text{ in } \mathcal{S} = \mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_4$$

So our F-statistic is:

$$\frac{\|P_{\mathcal{S} \ominus \mathcal{S}'} Y\|^2 / \dim(\mathcal{S} \ominus \mathcal{S}')}{\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 / \dim(\mathbb{R}^N \ominus \mathcal{S})}$$

Note that $\mathcal{S} \ominus \mathcal{S}' = \mathcal{S}_4$ and its dimension is $m - 1$.

Also that,

$$P_{\mathbb{R}^N \ominus \mathcal{S}} Y = \{\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} - \bar{Y}_{k.} + 2\bar{Y}_{..} : 1 = 1, \dots, m; j = 1, \dots, m\}$$

and its dimension is equal to $(m - 1)(m - 2)$.

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 = \sum_{i=1}^m \sum_{j=1}^m (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} - \bar{Y}_{k.} + 2\bar{Y}_{..})^2$$

Simultaneous Confidence Interval

Again, we use Latin as example,

$$H_0 : \mu \in \mathcal{S}'$$

$$H_1 : \mu \in \mathcal{S}$$

$$\mathcal{S} \ominus \mathcal{S}' = \mathcal{S}_4$$

Set of contrats:

$$\{C^T \mu : C \in \mathcal{S} \ominus \mathcal{S}' = \mathcal{S}_4\}$$

$$C \in \mathcal{S}_4$$

$$C^T 1_N = \sum_{i=1}^m \sum_{j=1}^m C_{k(i,j)} 1 \tag{4.4}$$

$$= m \sum_{k=1}^m C_k \tag{4.5}$$

$$= 0 \tag{4.6}$$

SCI

$$C^T P_{\mathcal{S}_4} Y \pm \|C\| \|P_{\mathcal{S}_5} Y\| \sqrt{\frac{\dim(\mathcal{S}_4)}{\dim(\mathcal{S}_5)} F_{\dim(\mathcal{S}_4), \dim(\mathcal{S}_5)}(1 - \alpha)}$$

where,

$$\|C\|^2 = \sum_{i=1}^m \sum_{j=1}^m C_{k(i,j)}^2 = m \sum_{k=1}^m C_k^2$$

4.8 Orthogonal Nested Model

This is somewhat like interaction with one main effect, but this is not interpreted as interaction. So the application background is not interaction but nested design.

Comparison between a crossed design and nested design.

Mathematically,

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$$

$$\begin{aligned} i &= 1, \dots, r \\ j &= 1, \dots, c_i \\ k &= 1, \dots, n_{ij} \end{aligned}$$

To decompose μ ,

$$\mathcal{S}_1 = \text{span}(1_N)$$

$$\mathcal{S}_2 = \{\mu_i : i = 1, \dots, r; j = 1, \dots, c_i\} \ominus \mathcal{S}_1$$

$$\mathcal{S}_3 = \{d_{ij} : k = 1, \dots, n_{ij}; i = 1, \dots, r; j = 1, \dots, c_i\} \ominus (\mathcal{S}_1 \oplus \mathcal{S}_2)$$

$$\mathcal{S}_4 = \mathbb{R}^n \ominus (\mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3)$$

Again,

$$E_{ij} = \{\delta_{uvw}^{ij} : u = 1, \dots, r, v = 1, \dots, c, w = 1, \dots, n_{ij}\}$$

$$\mathcal{S}_2 = \text{span}(E_i : i = 1, \dots, r) \ominus \mathcal{S}_1$$

$$\mathcal{S}_3 = \text{span}(E_{ij}) \ominus \mathcal{S}_1 \ominus \mathcal{S}_2$$

Explicitly,

$$P_{\mathcal{S}_1} \mu = \{\bar{\mu}_{..} : \dots\}$$

Monday October 17

$$P_{\mathcal{S}_2} \mu = \{\bar{\mu}_{i.} - \bar{\mu}_{..} : \dots\}$$

$$P_{\mathcal{S}_3} \mu = \{\bar{\mu}_{ij} - \bar{\mu}_{i.} : \dots\}$$

$$\bar{\mu}_{i.} = \theta$$

$$\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..}$$

$$\delta_{ij} = \mu_{ij} - \bar{\mu}_{..}$$

Then we have our parameterization according to orthogonal decomposition,

$$\mu_{ij} = \theta + \alpha_i + \delta_{ij}$$

UMVUE Estimation:

$$\hat{\theta} = P_{\mathcal{S}_1} Y = \{\bar{Y}_{...}\}$$

$$\hat{\alpha}_i = P_{\mathcal{S}_2} Y = \{\bar{Y}_{i..} - \bar{Y}_{...}\}$$

$$\hat{\delta}_{ij} = P_{\mathcal{S}_2} Y = \{\bar{Y}_{ij.} - \bar{Y}_{i..}\}$$

Hypothesis Testing

$$||P_{\mathcal{S}_1} Y||^2 = N\bar{Y}_{...}^2$$

$$||P_{\mathcal{S}_2} Y||^2 = \sum_i n_{i.} (\bar{Y}_{i..} - \bar{Y}_{...})^2 = SSA$$

$$||P_{\mathcal{S}_3} Y||^2 = \sum_i \sum_j n_{ij} (\bar{Y}_{ij.} - \bar{Y}_{i..})^2 = SSB(A)$$

$$||P_{\mathcal{S}_4} Y||^2 = \sum_i \sum_j \sum_k (\bar{Y}_{ijk} - \bar{Y}_{ij.})^2 = SSE$$

$$n_{i.} = \sum_j \sum_k 1 = \sum_j^{C_i} n_{ij}$$

We may show that these are independent (look at Covariance.)

$$\dim(\mathcal{S}_2) = r - 1$$

$$\dim(\mathcal{S}_3) = \sum_i^r C_i - r$$

$$\dim(\mathcal{S}_4) = N - \sum_i^r C_i$$

Scheffe's SCI

$$H_0 : \alpha = 0$$

$$\begin{aligned} \mathcal{S}' &= \mathcal{S}_1 \oplus \mathcal{S}_3 \\ \mathcal{S} &= \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3 \end{aligned}$$

$$\mathcal{S} - \mathcal{S}' = \mathcal{S}_2$$

$$\{C^T \mu : c \in \mathcal{S}_2\} = \{c^T P_{\mathcal{S}_2} \mu : C \in \mathcal{S}_2\}$$

$$C^T P_{\mathcal{S}_2} Y \pm ||c|| ||P_{\mathcal{S}_4} Y|| \sqrt{\frac{dim(\mathcal{S}_2)}{dim(\mathcal{S}_4)} F_{dim(\mathcal{S}_2), dim(\mathcal{S}_4)}(1-\alpha)}$$

$$c = \{f_i : \dots\}$$

$$C \perp 1_N$$

$$C^T 1_N = 0$$

$$\Sigma \Sigma \Sigma f_i = 0$$

$$\Sigma_i n_i . f_i = 0$$

$$||C||^2 = \sum_{i=1}^r n_i . f_i^2$$

$$t_i = n_i . f_i$$

$$||C||^2 = \sum_{i=1}^r \frac{t_i^2}{n_i .}$$

Part Two

5	Random Effects Model	85
5.1	Introduction to Random Effects	
5.2	Sampling Distributions	
5.3	Restricted MLE - REMLE	
5.4	Unbalanced Case of One Way Random Effect ANOVA	
5.5	Balance, Nested Random Effect Model	
5.6	Nested Mixed Effect Model	
5.7	Mean Parameterization of Link Function	
5.8	Bringing in the Predictors	
5.9	Dispersion Parameters	
6	Estimation in GLM	105
6.1	Overview	
6.2	Estimation of ϕ	
6.3	Statistical Inference for Generalized Linear Model	
7	Omitted	113
8	Omitted	115
9	Statistically Inference for GLM	117
9.1	Asymptotic Distribution	
9.2	Estimation of Asymptotic Variance of $(\hat{\beta})$ and its Confidence Interval	
9.3	Deviance Function	
9.4	Residuals	
9.5	Deviance Residual	
10	Omitted - adding to make next chapter	11
		123
11	Nature of Predictors	125
11.1	Link to ANOVA	
11.2	Analysis of Deviance (ANODEV)	
11.3	Numerical Prediction	
11.4	Testing Hypothesis	
12	Two Special Cases of GLM	131
12.1	Logistic Regression	

5. Random Effects Model

5.1 Introduction to Random Effects

ANOVA fixed effect model is based on decomposition of means. It could conceivably be called Analysis of Means (ANOME!). But random effect models is more strictly about variance (ANOVA).

In fixed effect model, the attention is focused on means, for example One-Way ANOVA where we explore μ_1, \dots, μ_p and check whether they are different through estimation. But in random effect model, we just want to know if they treatment makes a difference at all.

In this case, we can assume μ_1, \dots, μ_p are latent random variables (getting into a more Bayesian mindset) sampled from the same distribution.

Othertimes we have a case where p is large relative to n ($n = 200, p = 100$). We essentially only have two observations for each parameter! It's not a good strategy to estimate $\alpha_1, \dots, \alpha_{100}$, so we assume they are all independent and identically distributed at some distribution, say Normal with mean 0 and variance σ_a^2 . Then we only have one parameter! A very general idea, to deal with a lot of parameters.

Monday October 24

One Way Random ANOVA Model

$$Y_{ij} = \theta + \alpha_i + \varepsilon_{ij}$$

where $\theta \in \mathbb{R}$, fixed, and $i = 1, \dots, \phi; j = 1, \dots, n$.

$$\alpha_i \sim^{iid} N(0, \sigma_a^2)$$

$$\varepsilon_{ij} \sim^{iid} N(0, \sigma_e^2)$$

Moreover,

$$\{\alpha_i\} \perp\!\!\!\perp \{\varepsilon_{ij}\}$$

Recall that with fixed effect our hypothesis was that all α are the same and equal to zero. Our goal with random effect is to test the variance

$$H_0 : \sigma_a^2 = 0$$

5.2 Sampling Distributions

Theorem 5.2.1 Under the one way random effect model above, we have

$$E(Y_{ij}) = \theta$$

$$\text{Cov}(Y_{ij}) = \begin{cases} \sigma_a^2 + \sigma_e^2 & (i, j) = (i', j') \\ \sigma_a^2 & i = i', j \neq j' \\ 0 & i \neq i', j \neq j' \end{cases} \quad [ll]$$

$$E(Y_{ij}) = E(\theta + \alpha_i + \varepsilon_{ij})$$

$$\begin{aligned} \text{Proof.} \quad &= \theta + E(\alpha_i) + E(\varepsilon_{ij}) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{i'j'}) &= \text{Cov}(\alpha_i + \varepsilon_{ij}, \alpha_{i'} + \varepsilon_{i'j'}) \\ &= \text{Cov}(\alpha_i, \alpha_{i'}) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) \end{aligned}$$

$$\text{Cov}(\alpha_i, \alpha_{i'}) = \sigma_a^2 \quad (i = i')$$

$$\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = \sigma_e^2 \quad (i = i', j = j')$$

■

In matrix notation we may write the block of variances as

$$\sigma_e^2(I_n) + \sigma_a^2(1_n 1_n^T) = \sigma_e^2(I_n) + n\sigma_a^2 P_0$$

where $P_0 = P_{\text{span}(1_n)}$

We may rewrite as

$$(\sigma_a^2 + \sigma_e^2) \left(\frac{\sigma_e^2}{\sigma_a^2 + \sigma_e^2} (I_n) + n \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} P_0 \right)$$

where the fractions may be denoted by $1 - \rho$ and ρ , respectfully.

Above is for each block of variances, all together we have,

$$\text{Var}(Y) = (\sigma_a^2 + \sigma_e^2) \begin{pmatrix} (1 - \rho)I_n + n\rho(P_0) & & 0 \\ & \ddots & \\ 0 & & (1 - \rho)I_n + n\rho(P_0) \end{pmatrix}$$

Orthogonal Decomposition

$$\delta_{uvw}^{ij} = \begin{cases} 1 & \text{if } (u, v) = (i, j) \\ 0 & \text{else} \end{cases}$$

Let $E_{ij} = \{\delta_{uvw}^{ij} : u = 1, \dots, r, v = 1, \dots, c, w = 1, \dots, n_{uv}\}$ where the last index runs first.

$$E_{i.} = \sum_{j=1}^n E_{ij}$$

Let

$$\mathcal{S}_1 = \text{span}(1_N)$$

$$\mathcal{S}_2 = \text{span}(E_{1.}, \dots, E_{p.}) \ominus \mathcal{S}_1$$

So $\mathcal{S}_1 \perp \mathcal{S}_2$.

Let $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$

And $P = P_{\mathcal{S}} = \frac{1}{N} 1_N 1_N^T$ where $N = np$.

Lemma 6.2

We have the following identities about P_0 and P .

1. $\begin{pmatrix} P_0 & & \\ & \ddots & \\ & & P_0 \end{pmatrix} * P = P$
2. For any $C_1, C_2 \in \mathbb{R}$,

$$\begin{pmatrix} C_1 I_n + C_2 P_0 & & \\ & \ddots & \\ & & C_1 I_n + C_2 P_0 \end{pmatrix} P = (C_1 + C_2)P$$

This identity may be rewritten as

$$I_P \otimes (C_1 I_n + C_2 P_0)P = (C_1 + C_2)P$$

3. For any $C_1, C_2 \in \mathbb{R}$,

$$I_P \otimes (C_1 I_n + C_2 P_0)(I_P \otimes P_0) = (C_1 + C_2)(I_P \otimes P_0)$$

Proof. See in someone else's notes. Lots of matrices. ■

Wednesday October 26

Theorem 5.2.2 Under the one way random effect ANOVA model, we have

1. $P_{\mathcal{S}_2} Y \perp\!\!\!\perp P_{\mathbb{R}^N \ominus \mathcal{S}} Y$
2. $P_{\mathcal{S}_2} Y \sim N(0, (\sigma_e^2 + n\sigma_a^2)(\begin{pmatrix} P_0 & & \\ & \ddots & \\ & & P_0 \end{pmatrix} - P))$
3. $P_{\mathbb{R}^N \ominus \mathcal{S}} Y \sim N(0, (\sigma_e^2)(I_N - \begin{pmatrix} P_0 & & \\ & \ddots & \\ & & P_0 \end{pmatrix}))$
4. $\|P_{\mathcal{S}_2} Y\|^2 \sim (\sigma_a^2 + \sigma_e^2)\chi_{(p-1)}^2$
5. $\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 \sim (\sigma_e^2)\chi_{(np-p)}^2$

Proof. 1. $\text{Cov}(P_{\mathcal{J}_2}Y, P_{\mathbb{R}^N \ominus \mathcal{J}_2}Y) = P_{\mathcal{J}_2}\text{Var}(Y)P_{\mathbb{R}^N \ominus \mathcal{J}_2}Y$
See last lecture for $\text{Var}(Y)$.

Let

$$\begin{aligned}\mathcal{T}_1 &= \mathcal{J}_1 = \text{span}(1_N) \\ \mathcal{T}_2 &= \text{span}(E_1, \dots, E_p, \cdot)\end{aligned}$$

Then,

$$\mathcal{J}_1 = \mathcal{T}_1, \mathcal{J}_2 = \mathcal{T}_2 \ominus \mathcal{T}_1, \mathcal{J} = \mathcal{T}_2$$

So,

$$P_{\mathcal{J}_2} = P_{\mathcal{T}_2 \ominus \mathcal{T}_1} = P_{\mathcal{T}_2} - P_{\mathcal{T}_1}$$

Ultimately, because the two are multivariate normal with covariance 0, they must be independent.

2. Only need to calculate the mean and variance (we know that $P_{\mathcal{J}_2}Y$ is MVN).

$$\begin{aligned}E(P_{\mathcal{J}_2}Y) &= P_{\mathcal{J}_2}E(Y) \\ &= P_{\mathcal{J}_2}(\theta 1_N) \\ &= \theta P_{\mathcal{J}_2}1_N \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Var}(P_{\mathcal{J}_2}Y) &= P_{\mathcal{J}_2}\text{Var}(Y)P_{\mathcal{J}_2} \\ &= P_{\mathcal{J}_2}(\sigma_a^2 + \sigma_e^2) \begin{pmatrix} (1-\rho)I_n + n\rho(P_0) & & 0 \\ & \ddots & \\ 0 & & (1-\rho)I_n + n\rho(P_0) \end{pmatrix} P_{\mathcal{J}_2} \\ &= \\ &= (\sigma_a^2 + \sigma_e^2) \left(\begin{pmatrix} P_0 & & \\ & \ddots & \\ & & P_0 \end{pmatrix} - P \right)\end{aligned}$$

3. Again, we only need to check moments.
4. Follows from Cochran's Theorem.
5. Follows from Cochran's Theorem.

■

Hypothesis Test

$$H_0 : \sigma_a = 0$$

$$H_1 : \sigma_a > 0$$

By previous Theorem, under H_0 we have the following (independent) distributions:

$$||P_{\mathcal{J}_2}Y||^2 \sim (\sigma_a^2 + \sigma_e^2)\chi_{(p-1)}^2$$

$$||P_{\mathbb{R}^N \ominus \mathcal{J}}Y||^2 \sim (\sigma_e^2)\chi_{(np-p)}^2$$

Again we may calculate the F statistic under the null,

$$F = \frac{||P_{\mathcal{J}_2}Y||^2/(p-1)}{||P_{\mathbb{R}^N \ominus \mathcal{J}}Y||^2/(np-p)} \sim F_{p-1, np-p}$$

The same again (as the fixed effect model)! But we will find that under the alternative hypothesis, the statistic will be different.

Friday October 28

Alternative Hypothesis

If H_0 is not true,

$$||P_{\mathcal{J}_2}Y||^2 \sim (\sigma_e^2 + \sigma_a^2)\chi_{p-1}^2$$

$$||P_{\mathbb{R}^N \ominus \mathcal{J}}Y||^2 \sim \sigma_e^2\chi_{np-p}^2$$

$$\frac{\frac{||P_{\mathcal{J}_2}Y||^2/(p-1)}{(\sigma_e^2 + \sigma_a^2)}}{\frac{||P_{\mathbb{R}^N \ominus \mathcal{J}}Y||^2/\sigma_e^2}{(np-p)}} \sim F_{p-1, np-p}$$

Which gives us a scaled statistic,

$$\frac{\sigma_e^2}{(\sigma_e^2 + \sigma_a^2)}F \sim F_{p-1, np-p}$$

which is equivalent to a scaled F distribution,

$$F \sim \frac{(\sigma_e^2 + \sigma_a^2)}{\sigma_e^2}F_{p-1, np-p}$$

5.3 Restricted MLE - REMLE

For estimations of variance component, the idea of REMLE is that we use the density

$$P_{\mathbb{R}^N \ominus \mathcal{J}_1}Y$$

as the object function to be maximized instead of the full likelihood, which is the density of Y .

So when you want to estimate the variance you use the part of the likelihood that depends only on the variance if the likelihood can be factorized.

To motivate REMLE, we will consider a simple case.

$$Y_1, \dots, Y_n \sim^{iid} N(\mu, \sigma^2)$$

Note that μ is our fixed effect and σ^2 is the random effect.

MLE for σ^2 : $\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y})^2$

But this estimate is biased. But if we use distribution,

$$\|(I - P)Y\|^2 \sim \sigma^2 \chi_{n-1}^2$$

we have the following density to use,

$$f(s) \approx \left(\frac{s}{\sigma^2}\right)^{\frac{n-1}{2}-1} \exp \frac{s/\sigma^2}{2} \frac{1}{\sigma^2}$$

where $s = \|(I - P)Y\|^2$.

Note that we no longer have μ in the density.

We can show that now we have an unbiased estimator of σ^2 ,

$$\tilde{\sigma}^2 = \frac{1}{n-1} s$$

This is REMLE, which is just a systematic way of doing this in more complicated situations.

Now we may consider one way ANOVA for random effect. We are going to use

$$P_{\mathbb{R}^N \ominus \mathcal{S}_1} Y$$

as our restricted likelihood. But recall that we may rewrite the following

$$\mathbb{R}^N \ominus \mathcal{S}_1 = \mathcal{S}_2 \oplus (\mathbb{R}^N \ominus \mathcal{S})$$

also that there is one to one correspondence between

$$P_{\mathbb{R}^N \ominus \mathcal{S}_1} Y \leftrightarrow (P_{\mathcal{S}_2}, P_{(\mathbb{R}^N \ominus \mathcal{S})})$$

So it suffices to use the following distribution of

$$(P_{\mathcal{S}_2}, P_{(\mathbb{R}^N \ominus \mathcal{S})}) = (U, V)$$

which, it may be noted, are independent.

$$U \sim N \left(0, (\sigma_e^2 + \sigma_a^2) \left(\begin{pmatrix} P_0 & & \\ & \ddots & \\ & & P_0 \end{pmatrix} - P \right) \right)$$

$$V \sim N \left(0, \sigma_e^2 \left(I_N - \begin{pmatrix} P_0 & & \\ & \ddots & \\ & & P_0 \end{pmatrix} \right) \right)$$

In the following we may write $(\sigma_e^2 + \sigma_a^2)$ as γ .

First, let's understand what it means for a random vector, X to have $MVN(0, P)$, with P being a projection. It means that some projection of X is degenerate at 0.

Suppose $\text{rank}(P) = r$, then by spectral decomposition,

$$P = AA^T$$

where $A^T A = I_r$.

The distribution,

$$X \sim N(\mu, \sigma^2 P)$$

gives us two things,

1. $(I_p - P)X = 0$ ($\text{Var}(\dots) = 0$)
2. $A^T X \sim N(A^T \mu, \sigma^2 I_p)$ and this is the distribution of X .

So the distribution of

$$U = P_{\mathcal{S}_2}$$

gives us,

$$f(u) = f(A^T u) = \frac{1}{(2\pi)^{\frac{p-1}{2}} \gamma^{p-1}} e^{\frac{-1}{2\gamma^2} \|U\|^2}$$

Similarly,

$$V = P_{(\mathbb{R}^N \ominus \mathcal{S})}$$

gives us,

$$f(v) \approx \sigma_e^{\frac{np-p}{2}} e^{\frac{-1}{2\sigma_e^2} \|V\|^2}$$

We know that $U \perp V$ so we may find the joint density by multiplying,

$$f(u, v) \approx \gamma^{\frac{-(p-1)}{2}} \sigma_e^{\frac{-(np-p)}{2}} e^{-\frac{\|U\|^2}{2\gamma} - \frac{\|V\|^2}{2\sigma_e^2}} = \ell_R(\gamma, \sigma_e^2)$$

Now we want to maximize $\ell_R(\gamma, \sigma_e^2)$ over $\gamma \geq \sigma_e^2 \geq 0$.

Monday October 31

Let $(\tilde{\sigma}_e^2, \tilde{\gamma})$ be the unconstrained maximizer of ℓ_R . And let $(\hat{\sigma}_e^2, \hat{\gamma})$ be the constrained maximizer of ℓ_R .

We have two scenarios.

Scenario 1: $(\tilde{\sigma}_e^2, \tilde{\gamma}) = (\hat{\sigma}_e^2, \hat{\gamma})$

Scenario 2: $(\hat{\sigma}_e^2, \hat{\gamma})$

In other words, we should first maximize $\ell_R(\gamma, \sigma_e^2)$ to get $(\tilde{\sigma}_e^2, \tilde{\gamma})$. Then if $\tilde{\gamma} \geq \tilde{\sigma}_e^2$ we set $(\tilde{\sigma}_e^2, \tilde{\gamma}) = (\hat{\sigma}_e^2, \hat{\gamma})$. But if not, then we remaximize ℓ_r and set $(\hat{\sigma}_e^2, \hat{\gamma})$ to be maximizer.

$$\hat{\sigma}_e^2 = \hat{\gamma}, \hat{\gamma} = \hat{\gamma}$$

Here is the result,

Theorem 5.3.1 The REMLE for (γ, σ_e^2) takes the following forms,

1. Unconstrained: the maximizer is of $\ell_R((\gamma, \sigma_e^2))$, over $\sigma_e^2 \geq 0, \gamma \geq 0$ are

$$\tilde{\sigma}_e^2 = ||v||^2 / (np - p)$$

$$\tilde{\gamma} = ||u||^2 / (p - 1)$$

2. Constrained: if $\tilde{\gamma} \geq \hat{\sigma}_e^2$ then

$$(\hat{\sigma}_e^2, \hat{\gamma}) = (\tilde{\sigma}_e^2, \tilde{\gamma})$$

Otherwise,

$$\hat{\sigma}_e^2 = \hat{\gamma} = (||u||^2 + ||v||^2 / (np - 1))$$

Also, this second case in part 2 of the above theorem gives us that,

$$\sigma_a^2 = 0, \hat{\sigma}_e^2 = (||u||^2 + ||v||^2 / (np - 1))$$

Further note that

$$||u||^2 \text{ is SSA}$$

$$||v||^2 \text{ is SSE}$$

In the first case of part 2,

$$\sigma_a^2 = \frac{||u||^2}{np - p} - \frac{||v||^2}{p - 1}$$

5.4 Unbalanced Case of One Way Random Effect ANOVA

$$\{Y_{ij} : j = 1, \dots, n_i; i = 1, \dots, p\}$$

Definition 5.4.1 We sat that the above model follows a random effect (unballance) one way ANOVA moel if,

$$Y_{ij} = \theta + \alpha_i + \varepsilon_{ij}$$

where

$$\alpha_i \sim N(0, \sigma_a^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_e^2)$$

$$\{\alpha_i\} \perp \{\varepsilon_{ij}\}$$

As before, let $E_{ij} = \{\delta_{uvw}^{ij} : v = 1, \dots, n_i, i = 1, \dots, p\}$.

Let $N = \sum_{i=1}^p n_i$

Let

$$\mathcal{T}_1 = \text{span}\{1_N\}$$

$$\mathcal{T}_2 = \text{span}\{E_1, \dots, E_p\}$$

$$\mathcal{S}_3 = \mathbb{R}^N$$

Then let

$$\begin{aligned}\mathcal{S}_1 &= \mathcal{T}_1 \\ \mathcal{S}_2 &= \mathcal{T}_2 \ominus \mathcal{T}_1 \\ \mathcal{S}_3 &= \mathcal{T}_3 \ominus \mathcal{T}_2\end{aligned}$$

So that $\mathcal{S}_1 \perp \mathcal{S}_2 \perp \mathcal{S}_3$ and $\mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3 = \mathbb{R}^N$

Lemma 6.3 We have that

$$P_{\mathcal{S}_2} = \begin{pmatrix} P_{1_{n_1}} & & \\ & \ddots & \\ & & P_{1_{n_p}} \end{pmatrix}$$

where $P_{n_m} = \frac{1_m 1_m^T}{m}$ for $m = 1, \dots, p$.

Now $Y = \{Y_{ij} : j = 1, \dots, n_i; i = 1, \dots, p\}$.

We may calculate variance of Y.

Theorem 5.4.1 Under the one-way random effect model, we have that

$$\text{Var}(Y) = \sigma_e^2 I_N + \bar{n} \sigma_a^2 J_N P_{\mathcal{S}_2}$$

where $\bar{n} = \frac{n_1 + \dots + n_p}{p}$

$$\text{and } J_N = \begin{pmatrix} \frac{n_1}{\bar{n}} I_{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{n_p}{\bar{n}} I_{n_p} \end{pmatrix}$$

Proof. ■

Lemma 6.4

$$J_N P_{\mathcal{S}_2} = P_{\mathcal{S}_2} J_N$$

Theorem 5.4.2 — Orthogonality and Sample Distribution Theorem. Under the assumptions of definition above, we have that

1. $P_{\mathcal{S}_1} \perp P_{\mathcal{S}_2} \perp P_{\mathcal{S}_3}$
2. $P_{\mathcal{S}_2} Y \sim N(0, P_{\mathcal{S}_2} (\sigma_e^2 I_N + \bar{n} \sigma_a^2 J_N) P_{\mathcal{S}_2})$
3. $P_{\mathcal{S}_3} Y \sim N(0, \sigma_e^2 P_{\mathcal{S}_3})$
4. Under H_0 ,

$$\|P_{\mathcal{S}_2} Y\|^2 \sim \sigma_e^2 \chi_{p-1}^2$$

- 5.

$$\|P_{\mathbb{R}^N \ominus \mathcal{S}} Y\|^2 \sim \sigma_e^2 \chi_{N-p}^2$$

Proof. 1. Proved similarly to balance case.

2. Need calculations of moments, which has been completed.

3. Calculate moments.
4. Under $H_0 : \sigma_a^2 = 0$,

$$P_{\mathcal{J}_2}Y \sim N(0, \sigma_e^2 P_{\mathcal{J}_2})$$

where $\text{rank}(P_{\mathcal{J}_2}) = p - 1$.

and by Chochran's Theorem we have $\|P_{\mathcal{J}_2}Y\|^2 \sim \sigma_e^2 \chi_{p-1}^2$.

5. Under $H_1 : \sigma_a^2 = 0$,

$$P_{\mathcal{J}_3}Y \sim N(0, \sigma_e^2 P_{\mathcal{J}_3})$$

where $\text{rank}(P_{\mathcal{J}_3}) = N - p$.

and by Chochran's Theorem we have $\|P_{\mathcal{J}_3}Y\|^2 \sim \sigma_e^2 \chi_{N-p}^2$.

■

Wednesday November 2

Finished proof.

Hypothesis testing

$$H_0 : \sigma_a^2 = 0$$

For our test statistic we may again use

$$F = \frac{\|P_{\mathcal{J}_2}Y\|^2 / (p - 1)}{\|P_{\mathcal{J}_3}Y\|^2 / (N - p)}$$

Under H_0 ,

$$F \sim F_{p-1, N-p}$$

Under H_1 ,

$$\|P_{\mathcal{J}_2}Y\|^2 \sim (\sigma_e^2 + \bar{n}\sigma_a^2)\chi_{p-1}^2$$

So,

$$F \sim \frac{(\sigma_e^2 + \bar{n}\sigma_a^2)}{(\sigma_e^2)} F_{p-1, N-p}$$

5.5 Balance, Nested Random Effect Model

Recall our carmaker example. We have different models of cars nested under various makers. Cannot possibly have the different companies make the same model of cars.

Also recall our fixxed effect case,

$$Y_{ijk} = \theta + \alpha_i + \delta_{ij} + \varepsilon_{ijk}$$

Now,

$$Y_{ijk} = \theta + a_i + d_{ij} + e_{ijk}$$

where,
 $k = 1, \dots, n$
 $j = 1, \dots, c$
 $i = 1, \dots, p$

$$a_i \sim N(0, \sigma_a^2)$$

$$d_{ij} \sim N(0, \sigma_a^2)$$

$$e_{ijk} \sim N(0, \sigma_e^2)$$

Our set of random variables, $\{a_i\}, \{d_{ij}\}, \{e_{ijk}\}$ are all independent.

Let $Y = Y_{ijk}$ and $N = pcn$.

$$E_{ijk} = \{\delta_{uvw}^{ij} : u = 1, \dots, r; v = 1, \dots, c; w = 1, \dots, n\}$$

$$E_{i\cdot} = \sum_{j=1}^c E_{ij}$$

$$\begin{aligned}\mathcal{T}_1 &= \text{span}(1_N) \\ \mathcal{T}_2 &= \text{span}(E_{i\cdot}) \\ \mathcal{T}_3 &= \text{span}(E_{ij}) \\ \mathcal{T}_4 &= \mathbb{R}^N\end{aligned}$$

With these we can create,

$$\begin{aligned}\mathcal{S}_1 &= \mathcal{T}_1 \\ \mathcal{S}_2 &= \mathcal{T}_2 \ominus \mathcal{T}_1 \\ \mathcal{S}_3 &= \mathcal{T}_3 \ominus \mathcal{T}_2 \\ \mathcal{S}_4 &= \mathcal{T}_4 \ominus \mathcal{T}_3\end{aligned}$$

So we have that each of the above spaces are independent.

Let A_1, \dots, A_k be arbitrary matrices.

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

Lemma 6.5

$$P_{\mathcal{T}_3} = \oplus_{i=1}^p \oplus_{j=1}^n P_{1_N}$$

$$P_{\mathcal{T}_2} = \oplus_{i=1}^p P_{1_{nc}}$$

$$P_{\mathcal{T}_1} = P_{1_{pcn}}$$

Theorem 5.5.1 Under the nested random effect ANOVA, we have

$$\text{Var}(Y) = \sigma_e^2 I_N + ncP_{\mathcal{J}_2} + n\sigma_d^2 P_{\mathcal{J}_3}$$

Theorem 5.5.2 Under the nested random effect ANOVA model we have that

1. $P_{\mathcal{J}_1}Y \perp\!\!\!\perp P_{\mathcal{J}_2}Y \perp\!\!\!\perp P_{\mathcal{J}_3}Y \perp\!\!\!\perp P_{\mathcal{J}_4}Y$
2. $P_{\mathcal{J}_1}Y \sim N(\theta 1_N, (\sigma_e^2 + nc\sigma_a^2 + n\sigma_d^2)P_{\mathcal{J}_1})$
3. $P_{\mathcal{J}_2}Y \sim N(0, (\sigma_e^2 + nc\sigma_a^2 + n\sigma_d^2)P_{\mathcal{J}_2})$
4. $P_{\mathcal{J}_3}Y \sim N(0, (\sigma_e^2 + n\sigma_d^2)P_{\mathcal{J}_3})$
5. $P_{\mathcal{J}_4}Y \sim N(0, (\sigma_e^2)P_{\mathcal{J}_4})$
6. $\|P_{\mathcal{J}_1}Y\|^2 \sim (\sigma_e^2 + nc\sigma_a^2 + n\sigma_d^2)\chi_1^2(\frac{npc\theta^2}{(\sigma_e^2 + nc\sigma_a^2 + n\sigma_d^2)})$
7. $\|P_{\mathcal{J}_2}Y\|^2 \sim (\sigma_e^2 + nc\sigma_a^2 + n\sigma_d^2)\chi_{p-1}^2$
8. $\|P_{\mathcal{J}_3}Y\|^2 \sim (\sigma_e^2 + n\sigma_d^2)\chi_{pc-p}^2$
9. $\|P_{\mathcal{J}_4}Y\|^2 \sim (\sigma_e^2)\chi_{pcn-pc}^2$

Proof. 1. Only need to show $P_{\mathcal{J}_2}Y \perp\!\!\!\perp P_{\mathcal{J}_4}Y$ through covariance.
 2. Show moments.
 3. Moments
 4. Moments

■

Monday November 7

Hypothesis Testing

$$H_0 : \sigma_d^2 = 0, H_1 : \sigma_d^2 > 0$$

$$T(d) = \frac{\|P_{\mathcal{J}_3}Y\|^2 (pc - p)}{\|P_{\mathcal{J}_4}Y\|^2 (pcn - pc)}$$

Under the null hypothesis,

$$T(d) \sim F_{pc-p, pcn-pc}$$

Under the alternative hypothesis,

$$\|P_{\mathcal{J}_3}Y\|^2 (pc - p) \sim (\sigma_e^2 + n\sigma_d^2)\chi_{pc-p}^2$$

$$\|P_{\mathcal{J}_4}Y\|^2 (pcn - pc) \sim (\sigma_e^2)\chi_{pcn-pc}^2$$

$$\frac{[\|P_{\mathcal{J}_3}Y\|^2 / (\sigma_e^2 + n\sigma_d^2)] / (pc - p)}{[\|P_{\mathcal{J}_4}Y\|^2 / \sigma_e^2] / (pcn - pc)} \sim F_{pc-p, pcn-pc}$$

So if we rewrite above in the form of $T(d)$ we get,

$$T(d) \sim (1 + n\frac{\sigma_d^2}{\sigma_e^2})F_{pc-p, pcn-pc}$$

Now we may test σ_a^2

$$H_0 : \sigma_a^2 = 0, H_1 : \sigma_a^2 > 0$$

Which corresponds to the fixed effects version of

$$H_0 : \alpha_1 = \dots = \alpha_p$$

The test statistic for the fixed effect model was

$$\frac{||P_{\mathcal{J}_2}Y||^2 / \dim(\mathcal{J}_2)}{||P_{\mathcal{J}_4}Y||^2 / \dim(\mathcal{J}_4)}$$

Now instead we will use

$$T(a) = \frac{||P_{\mathcal{J}_2}Y||^2 / \dim(\mathcal{J}_2)}{||P_{\mathcal{J}_3}Y||^2 / \dim(\mathcal{J}_3)}$$

because their likelihoods are different. We derive UMPU-tests under different likelihood. It turns out that each are the respective UMPU tests for fixed and random effects respectively. This proof is beyond the scope of this course, will be covered in STAT 561.

Under the null hypothesis,

$$||P_{\mathcal{J}_2}Y||^2 \sim (\sigma_e^2 + n\sigma_a^2)\chi_{p-1}^2$$

$$||P_{\mathcal{J}_3}Y||^2 \sim (\sigma_e^2 + n\sigma_d^2)\chi_{pc-p}^2$$

$$T(a) \sim F_{p-1, pc-p}$$

Under the alternative hypothesis,

$$||P_{\mathcal{J}_2}Y||^2 \sim (\sigma_e^2 + nc\sigma_a^2n\sigma_d^2)\chi_{p-1}^2$$

$$||P_{\mathcal{J}_3}Y||^2 \sim (\sigma_e^2 + n\sigma_d^2)\chi_{pc-p}^2$$

$$T(a) \sim \left(\frac{\sigma_e^2 + nc\sigma_a^2n\sigma_d^2}{\sigma_e^2 + n\sigma_d^2} \right) F_{p-1, pc-p}$$

REML for $\sigma_e^2, \sigma_a^2, \sigma_d^2$

Based on the same principle, we will discard fixed effect from the likelihoods. $\ell_R(\sigma_e^2, \sigma_a^2, \sigma_d^2)$ is the distribution of

$$P_{\mathcal{J}_2}Y, P_{\mathcal{J}_3}Y, P_{\mathcal{J}_4}Y$$

This is

$$\ell_R \sim N(0, (\sigma_e^2 + nc\sigma_a^2 + n\sigma_d^2)P_{\mathcal{J}_2}) * N(0, (\sigma_e^2 + n\sigma_d^2)P_{\mathcal{J}_3}) * N(0, (\sigma_e^2)P_{\mathcal{J}_4})$$

Let the variance coefficients be denoted as: $\gamma_1, \gamma_2, \sigma_e^2$ where

$$\gamma_1 \geq \gamma_2 \geq \sigma_e^2$$

We want to maximize $\ell_R(\gamma_1, \gamma_2, \sigma_e^2)$ subject to the above inequality.

It is straightforward to show that the unconstrained maximizers are

$$\tilde{\gamma}_1 = \|P_{\mathcal{S}_3} Y\|^2 / (pc - p)$$

$$\tilde{\gamma}_2 = \|P_{\mathcal{S}_2} Y\|^2 / (p - 1)$$

$$\tilde{\sigma}_e^2 = \|P_{\mathcal{S}_4} Y\|^2 / (pcn - pc)$$

Let $A = \{(\gamma_1, \gamma_2, \sigma_e^2) : \gamma_1 \geq \gamma_2 \geq \sigma_e^2\}$.

Before in the one way case,

Use the same idea, but now the boundary is more complicated.

So if $(\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\sigma}_e^2) \in A$ then REML is

$$(\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\sigma}_e^2) = (\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\sigma}_e^2)$$

If $(\tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\sigma}_e^2) \notin A$ then we have three regions (in order we obtain A_1, A_2, A_3),

$$\{(\gamma_1, \gamma_2, \sigma_e^2) : \gamma_2 = \sigma_e^2\} \cup \{(\gamma_1, \gamma_2, \sigma_e^2) : \gamma_1 = \gamma_2\} \cup \{(\gamma_1, \gamma_2, \sigma_e^2) : \gamma_1 = \gamma_2 = \sigma_e^2\}$$

So $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\sigma}_e^2) = \max(\sup_{A_1} \ell_R, \sup_{A_2} \ell_R, \sup_{A_3} \ell_R)$.

5.6 Nested Mixed Effect Model

Mixed: mixed fixed and random effects.

So in our earlier car example the numerous car makers would be fixed effects, but the nested models would be random effects.

R Recall that greek letters denote fixed effects, and latin letters denote random effects.

Definition 5.6.1 — Nested Mixed Effect Model.

$$Y_{ijk} = \theta + \alpha_i + d_{ij} + e_{ijk}$$

where $\theta \in \mathbb{R}, \sum \alpha_i = 0$
 $d_{ij} \sim^{iid} N(0, \sigma_d^2)$
 $e_{ijk} \sim^{iid} N(0, \sigma_e^2)$
 $\{d_{ij}\} \perp \{e_{ijk}\}$

Let $\mu = \{\theta + \alpha_i : \dots\}$.

$\mathcal{T}_1, \dots, \mathcal{T}_4, \mathcal{S}_1, \dots, \mathcal{S}_4$ are defined as before.

Theorem 5.6.1 Under the nested mixed effect model we have that

$$E(Y) = \mu = P_{\mathcal{T}_2} \mu = \{\theta + \alpha_i : \dots\}$$

$$\text{Var}(Y) = \sigma_e^2 I_{pcn} + n\sigma_d^2 P_{\mathcal{T}_3}$$

Proof. Simple, left as exercise. ■

Theorem 5.6.2 Under nested mixed effect ANOVA model,

1. $P_{\mathcal{J}_1}Y \perp\!\!\!\perp P_{\mathcal{J}_2}Y \perp\!\!\!\perp P_{\mathcal{J}_3}Y \perp\!\!\!\perp P_{\mathcal{J}_4}Y$
2. $P_{\mathcal{J}_1}Y \sim N(\theta 1_N, (\sigma_e^2 + n\sigma_d^2)P_{\mathcal{J}_1})$
3. $P_{\mathcal{J}_2}Y \sim N(P_{\mathcal{J}_2}\mu, (\sigma_e^2 + n\sigma_d^2)P_{\mathcal{J}_2})$

where $P_{\mathcal{J}_2}\mu = \{\alpha_i - \bar{\alpha} : \dots\}$.

4. $P_{\mathcal{J}_3}Y \sim N(0, (\sigma_e^2 + n\sigma_d^2)P_{\mathcal{J}_3})$
5. $P_{\mathcal{J}_4}Y \sim N(0, (\sigma_e^2)P_{\mathcal{J}_4})$
6. MAKE SURE HE REWRITES 6.

Wednesday November 9

Friday November 11

■ **Example 5.1**

$$Y \sim \text{Pois}(\lambda)$$

$$\begin{aligned} f_\lambda(y) &= \frac{\lambda^y}{y!} e^{-\lambda} \\ &= c(\lambda) e^{\lambda \log y} \frac{1}{y!} \end{aligned}$$

So in exponential family form we have that $\lambda = \vartheta, c(\lambda) = e^{-\lambda}, t_0(y) = \log y, \mu_0(y) = \frac{1}{y!} \kappa(y)$.

Which gives us that $\mu_0(A) = \sum_{y \in A} \frac{1}{y!}$.

■

Let $Y_1, \dots, Y_n \sim^{iid} E_p(\theta t_0, \mu_0)$

$$f_{y_1}, \dots, f_{y_n}(y_1, \dots, y_n) = e^{\theta^T t_0(y_i)} \quad (5.1)$$

$$= \left[\prod_{i=1}^n c(\theta) \right] e^{\theta^T \sum_{i=1}^n t(y_i)} \quad (5.2)$$

This is still in exponential family form.

$$E_p(\theta, t, \mu)$$

$$t(y) = \sum_{i=1}^n t_0(y_i)$$

$$\mu = \mu_0 x \dots x \mu_0$$

Sometimes ϑ can be replaced by one to one function of ϑ .

5.7 Mean Parameterization of Link Function

Take ϑ as in $c(\vartheta)e^{\theta^T t_0(y)}$ is call canonical parameterization.

In GLM there is another parameterization, mean parameteraization. Take $t_0(y) = y$ so that $t_0 = I$.

$$c(\theta)e^{\theta^T y}$$

which may be rewritten as

$$c(\theta) = \frac{1}{\int e^{\theta^T y} d\mu_0(y)}$$

Let

$$b(\theta) = \log \int e^{\theta^T y} d\mu_0(\theta)$$

"b(.)" is called the cumulant generating function.

Then we have, with respect to $\mu_0(y)$:

$$f_\theta(y) = e^{\theta^T y - b(\theta)}$$

We claim that

$$E_\vartheta(Y) = b'(\vartheta) = \frac{db(\vartheta)}{d\vartheta}$$

$$Var_\vartheta(Y) = b''(\vartheta) = \frac{d^2 b(\vartheta)}{d\vartheta^2}$$

$$\begin{aligned} \int e^{\theta^T y - b(\theta)} d\mu_0(y) &= 1 \\ \frac{\delta}{\delta \vartheta} \int e^{\theta^T y - b(\theta)} d\mu_0(y) &= 0 \\ \int \frac{\delta}{\delta \vartheta} e^{\theta^T y - b(\theta)} d\mu_0(y) &= 0 \\ \int e^{\theta^T y - b(\theta)} (y - b(\theta)) d\mu_0(y) &= 0 \\ \int e^{\theta^T y - b(\theta)} y d\mu_0(y) - b(\theta) \int e^{\theta^T y - b(\theta)} d\mu_0(y) &= 0 \\ E_\theta(y) - b(\theta) &= 0 \\ E_\theta(y) &= b(\theta) \end{aligned}$$

By a similar argument (omitted here but will proved as a more general version later) we may show that

$$V_\vartheta(y) = b''(\vartheta)$$

Aside: In general, the cumulant generating function is

$$C_y(t) = \log M_y(t)$$

where $M_X(t) = E(e^{t^T y})$.

It is true that

$$\frac{d^k C_y(t)}{dt^k} = k^{th} \text{ cumulant of } y$$

What's cumulant? Not covered here but it note the first two cumulants are expectation and variance.

Since $b''(\theta) = \text{Var}_\theta(Y)$ if $\text{Var}(Y) > 0$ for all θ then

$$b(\theta) \uparrow \theta$$

So we have that

$$Y \sim E_p(b^{-1}(\mu), I, \mu_0)$$

Where this parameterization is called mean parameterization.

5.8 Bringing in the Predictors

Recall that in the linear model,

$$Y = \alpha + \beta^T X + \varepsilon$$

This was how you brought in predictors. But here our Y is Poisson, Hypergeometric, etc. This is not the best way to bring in predictors, but the above can be rewritten as

$$E(Y|X) = \alpha + f(X)$$

So in GLM,

$$E_\theta(Y|X) = \mu(\beta^T X)$$

that is we have the conditional distribution of Y given X ,

$$Y|X \sim E_p(\theta^T(\beta^T X), t_0, \mu_0)$$

We may write $\eta = \beta^T X$ where η is the "linear prediction" or "linear index".

Note that μ is a function of η and likewise η is an inverse function of μ , which is our link function.

In GLM, you chose the link function as part of the modeling process.

$$\eta = \ell(\mu)$$

$$\mu = \ell^{-1} \eta$$

Here ℓ is the link function.

To recap, there are three ways to parameterize exponential families,

- Canonical ϑ

- Mean μ
 - Linear Index η
- This is how they are related

$$\mu = b'(\theta)$$

$$\theta = b'^{-1}(\mu)$$

$$\eta = \ell(\mu)$$

$$\mu = \ell^{-1}(\eta)$$

$$\eta = (\ell \circ b')(\theta)$$

$$\theta = (\ell \circ b')^{-1}(\eta) = (b'^{-1} \circ \ell^{-1})(\eta)$$

Although you chose the link you want, that fits the data, there is a canonical link, most commonly used link.

Definition 5.8.1 The canonical link is the link that makes $\ell \circ b'$ identity so that $\theta = \eta$. So that the canonical parameter is the linear index.

So we have that

$$\ell \circ b' = I$$

$$\ell = b'^{-1}$$

where b'^{-1} is the canonical link.

So, using the link function we may write

$$Y|X E_p((b'^{-1} \circ \ell^{-1})(\beta^T X), t_0, \mu_0)$$

as (under the canonical link)

$$Y|X \sim E_p(\beta^T X, t_0, \mu_0)$$

5.9 Dispersion Parameters

As is normal we have μ, σ^2 . So when we bring σ^2 to GLM, this is called **Dispersion**.

Introduce segmented exponential family

$$f(y, \theta, \phi) = C(y, \phi) e^{\theta y - b(\theta)/a(\phi)}$$

where ϕ is the dispersion parameter, $a()$ is a (given) function.

Theorem 5.9.1 If $f(y, \theta, \phi)$ is of the form above, then

$$E_\theta(Y) = b'(\theta)$$

$$\text{Var}_\theta(Y) = a(\phi) b''(\theta)$$

Proof.

$$\begin{aligned}
 \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} d\mu_0(y) &= 1 \\
 \frac{\partial}{\partial \theta} \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} d\mu_0(y) &= 0 \\
 \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} y - b'(\theta) d\mu_0(y) &= 0 \\
 \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} y - b'(\theta) d\mu_0(y) &= 0 \\
 E(Y) - b'(\theta) &= 0 \\
 E(Y) &= b'(\theta)
 \end{aligned}$$

$$\begin{aligned}
 \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} (y - b(\theta)) d\mu_0(y) &= 0 \\
 \frac{\partial}{\partial \theta} \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} (y - b(\theta)) d\mu_0(y) &= 0 \\
 \int \frac{\partial}{\partial \theta} C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} (y - b(\theta)) d\mu_0(y) &= 0 \\
 \frac{1}{a(\phi)} \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} \frac{y - b'(\theta)}{a(\phi)} d\mu_0(y) - \int C(y, \phi) e^{\theta y - b(\theta)/a(\phi)} b''(\theta) d\mu_0(y) &= 0 \\
 \frac{1}{a(\phi)} \text{Var}_{\theta}(Y) - b''(\theta) &= 0 \\
 \text{Var}_{\theta}(Y) &= a(\phi) b''(\theta)
 \end{aligned}$$

■

Monday November 14

Finished proof.

So in μ -parameterization we have our variance function,

$$\text{Var}_{\theta}(Y) = a(\phi) b''(b^{-1}(\mu)) = a(\phi) V(\mu)$$

6. Estimation in GLM

6.1 Overview

$(X_1, Y_1), \dots, (X_n, Y_n)$ are independent and identically distributed from (X, Y) . This assumption is not crucial, you can treat X as fixed as well.

If we treat X_i as fixed, then Y_i are non identical, but still independent. So, for asymptotics we use Lindeberg-Feller Theorem.

If we are assuming that the X are also random then we can use the Lindeberg-Levy (iid CLT).

Another point, as in ANOVA, the X may not be explicit, that is the X could be subspaces described by dummy variables.

Anyway, let

$$(X_1, Y_1), \dots, (X_n, Y_n) \sim^{iid} (X, Y)$$

where (X, Y) has density

$$Y_i | X_i \sim C(\phi y_i) e^{\theta_i - b(\theta_i) / a(\phi)}$$

where our predictor comes in through $\theta_i = b'^{-1}(\mu(\beta^T X_i))$.

We often take

$$a(\phi) = \frac{\phi}{\Theta_i}$$

so that we may weight the values according to context. The joint density of (X_i, Y_i) is

$$f_{X_i}(x_i) f_{Y_i|X_i}(y_i|x_i)$$

But the marginal of X does not matter because it doesn't have any parameter terms after taking the log and then the derivative.

$$\log f_{X_i} + \log(\phi y_i) + \omega_i(\theta y_i - b(\theta))(\phi)$$

Take the partial derivatives with respect to θ_i and β .

Skipping some steps we get

$$\frac{\partial}{\partial \beta} [\log f_{X_i} + \log(\phi y_i) + \omega_i(\theta y_i - b(\theta))(\phi)] = w_i[(y_i - b'(\theta_i)/\phi)] \frac{1}{V(\mu_i)} \mu'(\beta^T X_i) X_i$$

So our score (derivative of log density) for β for a single observation is

$$w_i[(y_i - b'(\theta_i)/\phi)] \frac{\mu'(\beta^T X_i) X_i}{V(\mu_i)}$$

and the score equation of the whole sample is

$$\sum_{i=1}^n w_i X_i \frac{\mu'(\beta^T X_i) X_i}{V(\beta^T X_i)} [(y_i - b'(\theta_i)/\phi)]$$

We may solve this equation using Newton - Raphson algorithm.

R Recall the Newton-Raphson algorithm. Suppose $f : \mathbb{R}^p \rightarrow \mathbb{R}^p$ and we want to solve for $f(x) = 0$. We can take the Taylor Expansion around x_0 and get

$$f(x) \approx f(x_0) + \frac{\partial f}{\partial x^T}(x_0)(x - x_0) - \hat{f}(x)$$

Instead of solving nonlinear equation $f(x) = 0$ we may solve linear equation $\hat{f}(x) = 0$.

We can iterate to find next iteration of x using previous iterations,

$$x = x_0 - [\frac{\partial f}{\partial x^T}(x_0)]^{-1} f(x_0)$$

Iterate until convergence. This method works best if the objective function $\int f(x) dx$ is convex or concave (which is the case for all in exponential family - useful for GLM). Usually can converge to the 7th digit within 10 iteration using only 4 or 5 lines of code in R.

In our case,

$$S(\beta, X, Y) = \sum_{i=1}^n w_i a(\beta^T X_i) X_i (Y_i - \mu(\beta^T X_i))$$

$$\frac{\partial S}{\partial \beta^T} = \sum w_i a'(\eta_i) X_i X_i^T (Y_i - \mu(\theta^T X)) = \sum w_i a(\eta) X_i X_i^T (-\mu'(\beta^T X_i))$$

So the Newton-Raphson iteration will be

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \frac{\partial S(\hat{\beta}^{(k)}, X, Y)}{\partial \beta^T}^{-1} S(\hat{\beta}^{(k)}, X, Y)$$

A modification of this algorithm is called Fisher's Scoring method which ignores the first term in the partial derivative with respect to β^T and only focusing on the second to aide the algorithm.

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I^{-1}(\beta) S(\hat{\beta}^{(k)}, X, Y)$$

Wednesday November 16

Suppose we have U_1, \dots, U_n independent sample with the expectation of each being 0 and

$$\sum_{i=1}^n U_i = O_p\left(\frac{1}{\sqrt{n}}\right)$$

why might

$$\frac{\sum U_i}{\sqrt{n}} \rightarrow N(0, \Sigma)$$

but, if $E(U_i) \neq 0$ then

$$\frac{\sum U_i - n\mu}{\sqrt{n}} \rightarrow N(0, \Sigma)$$

here,

$$\sum_{i=1}^n U_i = n\mu + O_p(\sqrt{n}) = O_p(n)$$

and

$$\max(U_1, \dots, U_n) = O_p(\log n)$$

But we have that, using results from before

$$\begin{aligned} E(U_i) &= E\left(W_I \frac{\partial a(\eta_i)}{\partial \eta_i} X_i X_i^T (Y_i - \mu(\eta_i))\right) \\ &= E\left(W_I \frac{\partial a(\eta_i)}{\partial \eta_i} X_i X_i^T E(Y_i - \mu(\beta^T X_i | X_i))\right) \\ &= 0 \end{aligned}$$

So we can ignore the $O_p(\sqrt{n})$ term in the partial derivative of S with respect to β^T . We may now use

$$\frac{\partial S}{\partial \beta^T} \approx - \sum W_i a(\beta^T X_i) \mu(\beta^T X_i) X_i X_i^T$$

which is proportional to Fisher information and we may write it as $J_n(\beta)$.

So, can use iteration as follows

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + J_n(\hat{\beta}^{(k)}) S(\hat{\beta}^{(k)}, X, Y)$$

This turns out to be easier to compute via the Newton- Raphson method and is often more stable, but takes slightly longer to converge.

6.2 Estimation of ϕ

Commonly used two methods to estimate ϕ .

Maximum Likelihood Estimator

Assume we have estimated β as $\hat{\beta}$. Then the likelihood is

$$\prod_{i=1}^n C(\phi, Y_i) e^{(\hat{\theta}_i Y_i - b(\hat{\theta}_i)) / a(\phi)}$$

where $\hat{\theta}_i = b'^{-1}(\mu(\hat{\eta}_i))$ and $\hat{\eta}_i = \hat{\beta}^T X_i$.

We want to maximize the log likelihood with respect to ϕ (and then set equal to zero).

$$\frac{\partial \log L}{\partial \phi} = \sum \frac{C'(\phi, Y_i)}{C(\phi, Y_i)} - \sum (\hat{\theta}_i Y_i - b(\hat{\theta}_i)) (-1) a^{-2}(\phi) = 0$$

With Newton Raphson we get that the derivative of the above result gives us

$$\sum_{i=1}^n \left\{ \left[\frac{C''(\phi, Y_i)}{C(\phi, Y_i)} - \frac{C'(\phi, Y_i)}{C(\phi, Y_i)} \right] - (\hat{\theta}_i Y_i - b(\hat{\theta}_i)) \left(\frac{a''(\phi)}{a(\phi)} - 2 \frac{a'(\phi)^2}{a(\phi)^3} \right) \right\}$$

Use Newton-Raphson iteration to find ϕ , usually $a(\phi) = \phi$.

Method of Moments

This method is used in an asymptotic sense. Some detailed asymptotics are omitted (will be covered in STAT 561). We know that

$$\text{Var}(Y_i | X_i) = a(\phi) V(\mu(\beta^T X_i))$$

Let $\mu_i = \mu(\beta^T X_i)$ and $\hat{\mu} = \mu(\hat{\beta}^T X_i)$. Then it can be shown that

$$\sum_{i=1}^n \left[\frac{(Y_i - \mu_i)^2}{V(\mu_i)} - \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu})} \right] \rightarrow a(\phi) \chi_p^2$$

by vonMises expansion therefore,

$$E(\text{above equation}) \rightarrow a(\phi)P$$

In the meantime

$$\begin{aligned} E \left(\sum_{i=1}^n \frac{(Y_i - \mu_i)^2}{V(\mu_i)} \right) &= \sum_{i=1}^n E \left(\frac{(Y_i - \mu_i)^2}{V(\mu_i)} \right) \\ &= \sum_{i=1}^n E \left(\frac{E[(Y_i - \mu_i)^2 | X_i]}{V(\mu_i)} \right) \\ &= \sum_{i=1}^n a(\phi) E \left(\frac{V(\mu_i)}{V(\mu_i)} \right) \\ &= na(\phi) \end{aligned}$$

Now want to take $E(\dots)$ of both sides and equate them.

$$na(\phi) - E\left(\sum_{i=1}^n \frac{(Y_i - \hat{\mu})^2}{V(\hat{\mu})}\right) = pa(\phi)$$

$$E\left(\sum_{i=1}^n \frac{(Y_i - \hat{\mu})^2}{V(\hat{\mu})}\right) = (n-p)a(\phi)$$

So by Method of Moments we have

$$\sum_{i=1}^n \frac{(Y_i - \hat{\mu})^2}{V(\hat{\mu})} = (n-p)a(\phi)$$

and thus

$$a(\hat{\phi}) = \frac{1}{(n-p)} \sum_{i=1}^n \frac{(Y_i - \hat{\mu})^2}{V(\hat{\mu})}$$

which has the same form as the MSE for linear regression.

So if $a(\phi) = \phi$ then

$$\hat{\phi} = \frac{1}{(n-p)} \sum_{i=1}^n \frac{(Y_i - \hat{\mu})^2}{V(\hat{\mu})}$$

6.3 Statistical Inference for Generalized Linear Model

Let Y_1, \dots, Y_n be independent with density

$$C(\phi, y_i) e^{\ell_i(\beta, Y_i)/\phi}$$

where $a(\phi) = \phi$

$$\ell_i(\beta, Y_i) = W_i(Y_i \theta_i - b(\theta_i))$$

$$\theta_i = b'^{-1}(\mu(\beta^T X_i))$$

let $\hat{\beta}$ be the MLE, that is the maximizer of

$$\sum_{i=1}^n [\log C(\phi, y_i) + \ell_i(\beta, Y_i)/\phi]$$

or equivalently the maximizer of

$$\sum_{i=1}^n \ell_i(\beta, Y_i)$$

By standard theory of statistical inference we have that

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, I(\beta))$$

where $I(\beta)$ is the fisher information, which is the limit of

$$-\frac{1}{\phi} \frac{1}{n} \sum_{i=1}^n E \left(\frac{\partial^2 \ell_i(\beta, Y_i)}{\partial \beta \partial \beta^T} \right)$$

In practice we approximate $I(\beta)$ by

$$-\frac{1}{\phi} \frac{1}{n} \sum_{i=1}^n \frac{\mu^{2'}(\hat{\beta}^T X_i)}{V(\hat{\beta}^T X_i)} X_i X_i^T$$

Furthermore, if $\beta_0 \in \mathcal{S} \subseteq \mathbb{R}^p$ under the null hypothesis and suppose $\tilde{\beta}$ is the constrained MLE, then what is the distribution of

$$\sqrt{n}(\tilde{\beta} - \beta_0) \rightarrow N(0, \phi)$$

Friday November 18

This is the MLE under the full model. As in linear models what to test where

$$H_0 : \beta_0 \in \mathcal{S} \subseteq \mathbb{R}^T$$

Let $\tilde{\beta}$ be the constrained MLE under the null hypothesis. It can be shown (omitted - covered in STAT 561) that


$$\sqrt{n}(\tilde{\beta} - \beta_0) = P_{\mathcal{S}}^T(I^{-1}(\beta_0))\sqrt{n}(\hat{\beta} - \beta_0) + O_p(n^{-\frac{1}{2}})$$

$$P_{\mathcal{S}}^T(I^{-1}(\beta_0)) = V(V^T I^{-1}(\beta_0)V)^{-1}V^T I^{-1}(\beta_0)$$

where $\text{span}(V) = \mathcal{S}$

It follows that

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \tilde{\beta}) &= \sqrt{n}(\hat{\beta} - \beta_0) - \sqrt{n}(\tilde{\beta} - \beta_0) \\ &= \sqrt{n}(\hat{\beta} - P_{\mathcal{S}}^T(I^{-1}(\beta_0))\sqrt{n}(\hat{\beta} - \beta_0) + O_p(n^{-\frac{1}{2}})) \\ &= [I_p - P_{\mathcal{S}}^T(I^{-1}(\beta_0))] \sqrt{n}(\hat{\beta} - \beta_0) + O_p(n^{-\frac{1}{2}}) \\ &= Q_{\mathcal{S}}^T(I^{-1}(\beta_0)) \sqrt{n}(\hat{\beta} - \beta_0) + O_p(n^{-\frac{1}{2}}) \end{aligned}$$

 Here I is the Fisher information and I_p is the $p \times p$ identity matrix.

So, all the geometric structure reemerges at the asymptotic level after ignoring the small term, $O_p(n^{-\frac{1}{2}})$.

We can find the joint asymptotic distribution of

$$(\sqrt{n}(\hat{\beta} - \tilde{\beta}), \sqrt{n}(\tilde{\beta} - \beta_0))$$

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\tilde{\beta} - \beta_0) \\ \sqrt{n}(\hat{\beta} - \tilde{\beta}) \end{pmatrix} &= \begin{pmatrix} \sqrt{n}P_{\mathcal{S}}^T(I^{-1}(\beta_0)) \\ \sqrt{n}Q_{\mathcal{S}}^T(I^{-1}(\beta_0)) \end{pmatrix} + O_p(n^{-\frac{1}{2}}) \\ &= \begin{pmatrix} P_{\mathcal{S}}^T(I^{-1}) \\ Q_{\mathcal{S}}^T(I^{-1}) \end{pmatrix} \sqrt{n}(\hat{\beta} - \beta_0) + O_p(n^{-\frac{1}{2}}) \\ &\rightarrow \begin{pmatrix} P_{\mathcal{S}}^T(I^{-1}) \\ Q_{\mathcal{S}}^T(I^{-1}) \end{pmatrix} N(0, \phi I^{-1}) \\ &= N(0, \phi \begin{pmatrix} P_{\mathcal{S}}^T(I^{-1}) \\ Q_{\mathcal{S}}^T(I^{-1}) \end{pmatrix} I^{-1} (P_{\mathcal{S}}^T(I^{-1}), Q_{\mathcal{S}}^T(I^{-1}))) \end{aligned}$$

Note that

$$\Omega = \begin{pmatrix} P_{\mathcal{S}}^T(I^{-1})(I^{-1})P_{\mathcal{S}} & P_{\mathcal{S}}^T(I^{-1})(I^{-1})Q_{\mathcal{S}} \\ Q_{\mathcal{S}}^T(I^{-1})(I^{-1})P_{\mathcal{S}} & Q_{\mathcal{S}}^T(I^{-1})(I^{-1})Q_{\mathcal{S}} \end{pmatrix} = \begin{pmatrix} P_{\mathcal{S}}^T(I^{-1})(I^{-1})P_{\mathcal{S}} & 0 \\ 0 & Q_{\mathcal{S}}^T(I^{-1})(I^{-1})Q_{\mathcal{S}} \end{pmatrix}$$

So the asymptotic variance matrix for

$$\begin{pmatrix} \sqrt{n}(\tilde{\beta} - \beta_0) \\ \sqrt{n}(\hat{\beta} - \tilde{\beta}) \end{pmatrix}$$

is block -diagonal. So they are asymptotically independent (asymptotic normal with 0 covariance). Note the parallels between this and the linear model.

Let

$$\Delta(\tilde{\beta}, \beta_0) = \sqrt{n}(\tilde{\beta} - \beta_0)^T [P_{\mathcal{S}}^T(I^{-1})I^{-1}P_{\mathcal{S}}(I^{-1})] \sqrt{n}(\tilde{\beta} - \beta_0) + O_p(n^{-\frac{1}{2}})$$

Then by the continuous mapping theorem we have that

$$\Delta(\tilde{\beta}, \beta_0) \rightarrow \phi \chi_q^2$$

where $q = \dim(\mathcal{S})$.

Let

$$\Delta(\hat{\beta}, \tilde{\beta}) = \sqrt{n}(\hat{\beta} - \tilde{\beta})^T [Q_{\mathcal{S}}^T(I^{-1})I^{-1}Q_{\mathcal{S}}(I^{-1})] \sqrt{n}(\hat{\beta} - \tilde{\beta})$$

And again by CMT we have

$$\Delta(\hat{\beta}, \tilde{\beta}) \rightarrow \phi \chi_{p-q}^2$$

Furthermore the two are asymptotically independent.

Again note the parallels to the linear model. Finally, it can be shown (omitted - STAT 561) that

$$W_1 = 2 \sum_{i=1}^n [\ell_i(\tilde{\beta}, X_i) - \ell_i(\beta_0, X_i)] = \Delta(\tilde{\beta}, \beta_0) + O_p(n^{-\frac{1}{2}})$$

$$W_2 = 2 \sum_{i=1}^n [\ell_i(\hat{\beta}, X_i) - \ell_i(\tilde{\beta}, X_i)] = \Delta(\hat{\beta}, \tilde{\beta}) + O_p(n^{-\frac{1}{2}})$$

Which gives us the same results as above (convergence in distribution and asymptotic independence). It follows that

$$W_1 + W_2 \rightarrow \phi \chi_p^2$$

Note that X_1, \dots, X_n represent a generic sequence, so it's actually (Y_1, \dots, Y_n) . This leads naturally to something called deviance which is the generalization of sum of squares.



7. Omitted



8. Omitted

9. Statistically Inference for GLM

9.1 Asymptotic Distribution

9.2 Estimation of Asymptotic Varariance of $(\hat{\beta})$ and its Confidence Interval

By Section 9.1

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow N(0, \phi I^{-1}(\beta_0))$$

Which gives us that

$$AVar(\sqrt{n}(\hat{\beta} - \beta_0)) = \phi I^{-1}(\beta_0)$$

which can be estimated using results from chapters 6 and 7

$$\hat{\phi} I^{-1}(\hat{\beta})$$

Recall that

$$I_n(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n w_i X_i X_i^T \frac{\mu^2(X_i^T \hat{\beta})}{V(X_i^T \hat{\beta})}$$

So if you want to construct the C.I. for $C^T \beta_0$ it is

$$C^T \beta_0 \pm \xi_{\alpha/2} \sqrt{C^T AVar(\hat{\beta})}$$

Monday November 28

9.3 Deviance Function

This corresponds to SSE in the linear model. It is the 2 * loglikelihood ratio between the saturated model (i.e. p = n) and your Generalized Linear Model. This reduces to SSE.

Again, assume Y_1, \dots, Y_n independent where

$$Y_i = C_i(Y_i, \phi) e^{w_i(\theta_i Y_i - b(\theta_i)) / \phi}$$

The log likelihood ratio is

$$\sum_i \log C_i(Y_i, \phi) + w_i(\theta_i Y_i - b(\theta_i)) / \phi$$

So, to estimate θ_i you only need to consider

$$\sum_i w_i(\vartheta_i Y_i - b(\vartheta_i))$$

MLE under Saturated Model

Note that $\vartheta^* = \begin{pmatrix} \theta_1^* \\ \vdots \\ \theta_n^* \end{pmatrix}$.

$$\begin{aligned} \frac{\partial}{\partial \vartheta_j} \sum_i w_i(\theta_i Y_i - b(\vartheta_i)) &= \frac{\partial}{\partial \vartheta_j} [w_i(\theta_i Y_i - b(\vartheta_i))] \\ &= w_j(Y_j - b(\vartheta_j)) = 0 \end{aligned} \quad \Rightarrow \theta_j^* = b'^{-1}(y_j)$$

MLE under Full Model

Here $\eta \in \mathcal{S} \subseteq \mathbb{R}^p$. Also, recall that $\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$ is our linear predictor. Because $\eta \in \mathcal{S}$ we have that $\dim(\mathcal{S}) = \phi$.

$$\eta = X\beta$$

$$X \in \mathbb{R}^{n \times p}$$

Here X is a basis matrix of \mathcal{S} , that is to say that $\mathcal{S} = \text{span}(X)$. Recall that

$$\vartheta_i = b'^{-1}(\ell^{-1}(\eta_i)) = b'^{-1}(\mu(\eta_i))$$

So we can maximize $\sum_i w_i(\vartheta_i Y_i - b(\vartheta_i))$ (mentioned before as important to consider) over

$$\left\{ \begin{pmatrix} b'^{-1}(\ell^{-1}(\beta^T X_1)) \\ \vdots \\ b'^{-1}(\ell^{-1}(\beta^T X_n)) \end{pmatrix} : \beta \in \mathbb{R}^p \right\}$$

We explained how to do this via Newton Raphson or Fisher scoring to get $\hat{\beta}$. From here you can find that

$$\hat{\vartheta}_i = b'^{-1}(\ell^{-1}(\hat{\beta}^T X_i))$$

is the MLE of ϑ under the full model.

So, the deviance is defined to be

$$D(\theta^*, \hat{\vartheta}) = 2[\ell(\theta^*) - \ell(\hat{\theta})]$$

Where ℓ is defined before as the log likelihood ratio.

The scaled deviance is

$$\frac{1}{\phi} D(\theta^*, \hat{\vartheta})$$

It can be shown under the null hypothesis, $H_0 : \vartheta \in \mathcal{S}$, that

$$D(\theta^*, \vartheta_0) - D(\theta^*, \hat{\vartheta}) = 2(\ell(\hat{\theta}) - \ell(\vartheta_0)) \rightarrow \phi \chi_p^2$$

Furthermore, consider a submodel where $H_0 : \eta \in \mathcal{S}' \subseteq \mathcal{S}$ where the dimension of \mathcal{S} is $q < p$. (Again, this is a parallel of LM where we consider $\mu = \theta = \eta$ and it made no difference to state that $\mu \in \mathcal{S}'$, $\eta \in \mathcal{S}'$, $\theta \in \mathcal{S}'$. But here, we only say that $q \in \mathcal{S}'$).

Typically, there exists a submatrix, \tilde{X} of X such that

$$\eta = \tilde{X}\tilde{\beta}$$

where $\tilde{\beta}$ is a subvector of β . In general, \tilde{X} can be any $n \times q$ matrix such that

$$\text{span}(\tilde{X}) = \text{span}(X)$$

and $\tilde{\beta}$ can be any q linear combination of β .

So let $\tilde{\vartheta}$ be the MLE under $H_0 : \eta \in \mathcal{S}'$ (this can be done using Newton Raphson with X replace by \tilde{X}).

Then you have that

$$D(\theta^*, \tilde{\vartheta}) = 2[\ell(\theta^*) - \ell(\tilde{\theta})]$$

As we discussed in Section 9.1 (?), in the deviance,

1. $D(\theta^*, \vartheta_0) - D(\theta^*, \tilde{\vartheta}) \rightarrow \phi \chi_q^2$
2. $D(\theta^*, \tilde{\vartheta}) - D(\theta^*, \hat{\vartheta}) \rightarrow \phi \chi_{p-q}^2$
3. $D(\theta^*, \vartheta_0) - D(\theta^*, \tilde{\vartheta})$ and $D(\theta^*, \tilde{\vartheta}) - D(\theta^*, \hat{\vartheta})$ are asymptotically independent.

So you can use deviance as a substitute for sum of squares in statistical inference.

9.4 Residuals

The same diagnostic issues arise as in GLM. We want to look at the residuals to see if any important is in trash, if there is anything in your model not good. We will consider three types of residuals.

9.4.1 Pearson's Residual

$$r_p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

where $\hat{\mu}_i = \mu(\hat{\beta}X_i) = \ell^{-1}(\hat{\beta}^T X_i)$.

However, this residual is typically skewed. For example, Y could be Poisson, Gamma, Chi Squared, etc.

9.4.2 Anscombe Residual

Used to correct for skewness. This is derived from asymptotic expansion not used in this course.

Let $A(\mu) = \int V^{-1/3}(\mu) d\mu$. Or rather that

$$\frac{dA(\mu)}{d\mu} = V^{-1/3}(\mu)$$

Then we have that

$$r_A = \frac{A(y_i) - A(\hat{\mu}_i)}{A(\hat{\mu}_i) \sqrt{V(\hat{\mu}_i)}}$$

■ Example 9.1

$$Y_i \sim \text{Pois}(\mu_i)$$

$$V(\mu_i) = \mu_i$$

$$V(\mu) = \mu$$

■

Wednesday November 30

9.5 Deviance Residual

The idea from before is that

$$SSE = \sum_i (Y_i - \hat{\beta}^T X_i)^2$$

For GLM,

$$D(\vartheta^*, \hat{\theta}) = \sum_i d_i$$

where $d_i \geq 0$.

From here we may calculate the deviance residual,

$$r_D = \text{sgn}(d_i) \sqrt{d_i}$$

More specifically,

$$d_i = 2w_i(\vartheta^* y_i - b(\vartheta_i^*)) - w - i(\hat{\vartheta}_i y_i - b(\hat{\beta}_i)) = 2w_i[(\vartheta^* - \hat{\vartheta}_i)y_i - (b(\vartheta_i^*) - b(\hat{\beta}_i))]$$

Recall that

$$\vartheta_i^* = b'^{-1}(y_i)$$

$$\hat{\vartheta}_i = b'^{-1}(\hat{\mu}_i)$$

$$\hat{\mu}_i = \ell^{-1}(\hat{\beta}^T X_i) = \mu(\hat{\beta}^T X_i)$$

This is a generalization of

$$Y_i = \hat{\beta}^T X_i = \text{sgn}(y_i - \hat{\mu}_i) \sqrt{(y_i - \hat{\mu}_i)^2}$$

This also is less skewed than Pearson (as seen in ?).



10. Omitted - adding to make next chapter 11

11. Nature of Predictors

11.1 Link to ANOVA

The difference between ANOVA and regression is that ANOVA prediction is categorical. Here, consider categorical prediction for GLM. So, we are doing ANOVA for GLM - ANODEV (Bing Li's abbreviation)

1. 1 - way Main Effect Model

$$\eta_{ij}C + \alpha_i; j = 1, \dots, n_i; i = 1, \dots, p$$

$$\sum_i \alpha_i = 0$$

Note that $\alpha_p = -\alpha_1 - \dots - \alpha_{p-1}$

Consider the example where $n_1 = 3, n_2 = 3, n_3 = 2$

$$\eta_{11} = c + \alpha_1$$

$$\eta_{12} = c + \alpha_1$$

$$\eta_{13} = c + \alpha_1$$

$$\eta_{21} = c + \alpha_2$$

$$\eta_{22} = c + \alpha_2$$

$$\eta_{23} = c + \alpha_2$$

$$\eta_{31} = c + \alpha_3$$

$$\eta_{32} = c + \alpha_3$$

$$\eta = \begin{pmatrix} \eta_{11} \\ \vdots \\ \eta_{32} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ -1 & -1 \\ -1 & -1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} c \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = 1_n C + A \alpha_{12}$$

Obviously we may also use $\alpha_2 = -\alpha_1 - \alpha_3$

We introduce algebraic structure to describe categorical predictors. Let A and B represent the design matrices for factor A and B.

1. *Addition* We use $A + B$ to represent $(A:B)$ and remove any dependent columns. (Note that $A + cA = A$ and $A + B = B + A$)
2. *Multiplication* $A \cdot B$ is defined as

$$\{A\}i \odot B_j : j = 1, \dots, q; i = 1, \dots, p\}$$

Where \odot is Hadamard matrix product which is point wise product.

3. *Subtraction* IF $f(A, B)$ is an algebraic factor, of the form $g(A, B) + A$, then $f(A, B) - A$ simply means $g(A, B)$.

The following applies largely to factor algebra.

Proposition 11.1 (Proved in HW 6)

1. Associative Law:

$$A \cdot (B \cdot C) = (A \cdot B) \cdot C$$

2. Associative Law:

$$A + (B + C) = (A + B) + C$$

3. Distributive Law:

$$A \cdot (B + C) = (A \cdot B) + (A \cdot C)$$

Using these algebraic notation, we can summarize interaction models and nested models nicely as follows

2-way ANODEV with Iteration:

$$\eta_{ijk} = c + \alpha_i + \beta_j + \gamma_{ij}$$

$$\begin{aligned} \sum_i \alpha_i &= 0 \\ \sum_j \beta_j &= 0 \\ \sum_i \gamma_{ij} &= 0 \\ \sum_j \gamma_{ij} &= 0 \end{aligned}$$

We may write this in matrix form as

$$\eta = 1_n C + A \alpha_{1:p-1} + B \beta_{1:q-1} + (A \cdot B) \gamma_{1:(p-1)(q-1)}$$

Then plug in all the formulates before to do estimation, inference, diagnostics, etc.

Friday December 2

Crossed design is abbreviated as $A * B$.

Nested design is abbreviated as $A + AB$. This is written as A/B and is called B neded in A.

Lot of the designed can be created from these basic designs.

1. 3 Way Crossed Design

$$A * B * C = A + B * C + A \cdot (B * C) = A + B + C + A \cdot B + A \cdot C + B \cdot C + A \cdot B \cdot C$$

2. Nested Nested Design

$$A/(B/C)$$

$$A + A \cdot (B/C) = A + A \cdot (B + B \cdot C)$$

$$= A + A \cdot B + A \cdot B \cdot C$$

You can write down the model in terms of η easily.

1. 3 Way Crossed Design

$$\eta_{ijk} = c + \alpha_i + \beta_j + \gamma_k + \kappa_{ij} + \nu_{jk} + \varepsilon_{ijk}$$

2. Nested Nested Design

$$\eta_{ijk} = c + \alpha_i + (\alpha\beta)_{ij} + (\alpha\beta\gamma)_{ijk}$$

This also applies to ANOVA.

11.2 Analysis of Deviance (ANODev)

You can ceate a table of deviance just like ANOVA table, except that orthogonality is lost. So the table is not additive. In fact, without orthogonal design, even ANOVA table is nonadditive.

We will illustrate ANODev tabel by crossed design with intercept.

$$1 + A + B + A \cdot B$$

Supose A has p-levels and B has q-levels. Then the total Degrees of Freedom will be

$$1 + (p - 1) + (q - 1) + (p - 1)(q - 1) = pq$$

In terms of deviance, in general for any submodel the deviance column is

$$-D(\theta^*, \hat{\theta}(full)) - D(\theta^*, \tilde{\theta}(complement of sub))$$

Which is analogous to

$$-SSE(full) + SSE(compliment of sub)$$

Factor	Df	DEV	Deviance
1	1	$2[\ell(\hat{\theta}) - \ell(\tilde{\theta}_{-1})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-1})$
A	p-1	$2[\ell(\hat{\theta}) - \ell(\tilde{\theta}_{-A})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-A})$
B	q-1	$2[\ell(\hat{\theta}) - \ell(\tilde{\theta}_{-B})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-B})$
A · B	(p-1)(q-1)	$2[\ell(\hat{\theta}) - \ell(\tilde{\theta}_{-A \cdot B})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-A \cdot B})$
A * B	pq - 1	$2[\ell(\hat{\theta}) - \ell(\tilde{\theta}_{-A * B})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-A * B})$

The role played by orthogonality in ANOVA when it exists,

$$SS(sub_1) = SSE(sub_1^C) - SSE(F)$$

$$SS(sub_2) = SSE(sub_2^C) - SSE(F)$$

Under orthogonality,

$$SS(sub_1 \cup sub_2) = SS(sub_1) + SS(sub_2)$$

Without orthogonality, this equation no longer holds.

$SS(sub)$, as defined by projection norm no longer adequately represents the variation of the submodel. However,

$$SSE(sub^C) - SSE(F)$$

still makes sense since in ANODEV, we only have the full.

11.3 Numerical Prediction

In this section, bring in numerical predictor on top of categorical predictor. In linear model this is called ANCOVA, Analysis of Covariance. For example, we may have,

$$Y_{ij} = \mu + \alpha_i + \beta^T x + \varepsilon_{ij}$$

where α_i is the categorical predictor and $\beta^T x$ is the numerical predictor.

A linear model with each cell having different intercepts, but common slope.

ANCODEV, or Analysis of Codeviance

$$\eta_i = C + \alpha_i + X_i^T \beta$$

In algebraic representation, $1_n : A : X$.

This can also apply to more complicated categorical predictors.

2 Way Main Effect Model

$$C + \alpha_i + \beta_i + \gamma^T X$$

$$1 + A + B + X$$

2 Way Crossed Design Model

$$1 + A + B + A \cdot B + X$$

Factor	Df	DEV	Deviance
1	1	$2[\ell(\hat{\theta}_{1+A*B+X}) - \ell(\tilde{\theta}_{A*B-X})]$	$-D(\theta^*, \hat{\theta}_{1+A*B+X}) + D(\theta^*, \tilde{\theta}_{A*B+X})$
A	p-1	$2[\ell(\hat{\theta}_{1+A*B+X}) - \ell(\tilde{\theta}_{1+A*B+X-A})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-A})$
B	q-1	$2[\ell(\hat{\theta}) - \ell(\tilde{\theta}_{-B})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-B})$
A · B	(p-1)(q-1)	$2[\ell(\hat{\theta}) - \ell(\tilde{\theta}_{-A \cdot B})]$	$-D(\theta^*, \hat{\theta}) + D(\theta^*, \tilde{\theta}_{-A \cdot B})$
X	r		
A*B + X	(pq-1) + r		

Note the general formulas for deviance are

$$2[\ell(\hat{\theta}_{full}) - \ell(\hat{\theta}_{sub^c})] = D(\theta^*, \tilde{\theta}_{sub^c}) - D(\theta^*, \hat{\theta}_{full})$$

11.4 Testing Hypothesis

Illustration by cross design ANCODEV

$$H_0 : A = 0$$

$$H_0 : \alpha_1 = \cdots = \alpha_p$$

$$2[\ell(\hat{\theta}_{1+A*B+X}) - \ell(\tilde{\theta}_{1+A*B+X-A})] \rightarrow^{\mathcal{D}} \chi_{pq+r-pq+r-(p-1)}^2 = \chi_{p-1}^2$$

That is, we want to use χ_{p-1}^2 as the reference distribution. We may wonder why not the F distribution, as it is so often used in prior models, but using a strict analogy to linear model we can see that

$$\frac{D(\theta^*, \tilde{\theta}_{sub^c}) - D(\theta^*, \hat{\theta}_{full})}{D(\theta^*, \hat{\theta}_{full})} \rightarrow Fdf$$

But the bottom assumption does not hold. When n is large, χ^2 large is constant. So even if you try to use this approach to getting an F distribution, it is still approximately a χ^2 distribution when n is large.

12. Two Special Cases of GLM

12.1 Logistic Regression

$$Y_i \sim b(m_i, p_i)$$

Likelihood

$$\binom{m_i}{y_i} e^{y_i \log(\frac{p_i}{1-p_i}) + m_i \log(1-p_i)}$$

Canonical Parameter: $\theta_i = \log(\frac{p_i}{1-p_i}) \rightarrow p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$

So we have our Logit and Expit links, respectively, above.

$$\begin{aligned} \log(1-p_i) &= \log(1 - \frac{e^{\theta_i}}{1+e^{\theta_i}}) \\ &= \log(\frac{1}{1+e^{\theta_i}}) \\ &= -\log(1+e^{\theta_i}) \end{aligned}$$

In terms of canonical parameterization,

$$\binom{m_i}{y_i} e^{\theta_i y_i - m_i \log(1+e^{\theta_i})}$$

So $b(\theta_i) = m_i \log(1+e^{\theta_i})$.

$$b'(\theta_i) = \mu_i = m_i \frac{e^{\theta_i}}{1+e^{\theta_i}}$$

Inverting this we get that

$$\theta_i = \log \frac{\mu_i/m_i}{1 - \mu_i/m_i}$$

So your natural link function is to make $\theta_i = \eta_i$,

$$\text{link}_{\text{nat}}(\mu_i) = \log \frac{\mu_i/m_i}{1 - \mu_i/m_i} = \log \frac{p_i}{1 - p_i}$$

For natural link, regression is logistic regression.

$$\log \frac{p_i}{1 - p_i} = X_i^T \beta$$

Where X_i may be categorical (ANODEV), numerical (regression), or mixed (ANCODEV).

Estimation

Log Likelihood

$$\sum_i (x_i^T \beta) y_i - m_i \log(1 + e^{\beta^T x_i})$$

Score

$$S(\beta, y) = \sum_i x_i (y_i - m_i (\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}))$$

Solve by Newton Raphson or Fisher Scoring for the Fisher information.

$$n^{-1} \sum_i m_i X_i X_i^T (\frac{e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})^2})$$

Two other types of link functions for binary response:

1. Probit Link (probability)

$$\text{link}_{\text{probit}} : \mu_i \mapsto \Phi^{-1}(\mu_i/m_i)$$

where Φ is the c.d.f. of $N(0, 1)$.

$$\text{link}_{\text{probit}}^{-1} : \mu(\eta_i) = m_i \Phi(\eta)$$

The idea of link function is for us to be able to go from $\eta_i \leftarrow \mathbb{R}, \mu_i \leftarrow \text{special range}$.

GLM using probit as link is call Probit Regression.

2. Complementary Log-Log Link

$$\text{link}_{\text{cloglog}}(\mu_i) = \log(-\log(1 - \frac{\mu_i}{m_i}))$$

$$\text{link}_{\text{cloglog}}^{-1}(\eta_i) = \mu(\eta_i) = m_i(1 - e^{-e^{\eta_i}})$$

For these 2 links and all (?) link, estimation goes as follows

$$S(\beta, y_i) = x_i \frac{\mu(\eta_i)}{V(\eta_i)} (y_i - \mu(\eta_i))$$

where $\eta_i = X_i^T \beta$.

By some algebra,

$$V(\mu_i) = b \circ b(\mu_i) = m_i \left(\frac{\mu_i}{m_i} \right) \left(1 - \frac{\mu_i}{m_i} \right)$$

So, the score function for observation of (X_i, Y_i) is

$$X_i \frac{\mu'(\eta_i)}{m_i \frac{\mu(\eta_i)}{m_i} \left(1 - \frac{\mu(\eta_i)}{m_i} \right)} (y_i - \mu(\eta_i))$$

Want to solve $\sum_i (\text{above}) = 0$ using Newton-Raphson or Fisher Scoring.
For the two link functions we discussed,

