Cumulative Knowledge (Frist 2 Years of Gradschool)

Summarized by Svetlana Eden

Month Day, 2015

Contents

List of Figures

List of Tables

1 Matrix algebra, Linear models

From *Linear Models in Statistics*, second edition, A.C.Rencher and G.B.Schaalje

1.0.1 Matrix Inversion, page 23-24

1) $(A^{-1})' = (A')^{-1}$

2) $(AB)^{-1} = B^{-1}A^{-1}$

3)

$$If \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad and \quad B = A_{22} - A_{21}A_{11}^{-1}A_{12}; \quad then \quad A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}B^{-1}A_{21}A_{11}^{-1} & A_{11}^{-1}A_{12}B^{-1} \\ -B^{-1}A_{21}A_{11}^{-1} & B^{-1} \end{bmatrix}$$

4)

$$If \quad A = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{a}_{12} \\ \boldsymbol{a}'_{12} & a_{22} \end{bmatrix} \quad and \quad b = a_{22} - \boldsymbol{a}'_{12}\boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}; \quad then \quad A^{-1} = \frac{1}{b}\begin{bmatrix} b\boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}\boldsymbol{a}'_{12}\boldsymbol{A}_{11}^{-1} & \boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12} \\ -\boldsymbol{a}'_{12}\boldsymbol{A}_{11}^{-1} & 1 \end{bmatrix}$$

5)

$$If \quad A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}^{-1} \quad then \quad A^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix}$$

6) $(B + cc')^{-1} = B^{-1} - \frac{B^{-1}cc'B^{-1}}{1+cB^{-1}c}$

7) $(A + PBQ)^{-1} = A^{-1} - A^{-1}PB(B + BQA^{-1}PB)^{-1}BQA^{-1}$

8)

$$If \quad A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \quad then \quad A^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

1.0.2 Positive Definite Matrices, page 27-28

Definition: Matrix A is *positive definite* if for any y: $y'Ay > 0$.
Definition: Matrix A is *positive semidefinite* if there exists such $y \neq 0$ that: $y'Ay = 0$.

1) If B is $n \times p$, where $n > p$ and $rank(B) = p(< p)$ then $B'B$ is positive definite (semidefinite)

2) If A is positive definite then A^{-1} is positive definite.

3) If A is positive definite and is partitioned in the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} were square, then A_{11} and A_{22} are positive definite

4) If A is positive definite (semidefinite) then diagonal elements $a_{ii} > 0$ ($a_{ii} \geq 0$)

5) If P is nonsingular and A is positive definite (semidefinite) then $P'AP$ is positive definite (semidefinite).

6) If B is an $n \times p$ and if $rank(B) = p$ ($rank(B) < p$) then $B'B$ is positive definite (semidefinite).

7) If A is positive definite $p \times p$ and B is $k \times p$ with $k \leq p$ then BAB' is positive definite.

8) A is positive definite if and only if there exists a non-singular matrix such that $A = P'P$ (Cholesky decomposition).

1.0.3 Generalized Inverse, page 32-

Definition: A *generalized inverse*(also called a *conditional inverse*) of an $n \times p$ matrix A is any matrix A^- that satisfies:

$$AA^-A = A \tag{1}$$

1) If B is $n \times p$, where $n > p$ and $rank(B) = p(< p)$ then $B'B$ is positive definite (semidefinite)

2) If A is positive definite then A^{-1} is positive definite.

3) If A is positive definite and is partitioned in the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} were square, then A_{11} and A_{22} are positive definite

4) Assume that A is $n \times p$ and $rank(A) = r$, and A_{11} is $r \times r$ and $rank(A_{11}) = r$:

$$If \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad then \ \ A^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

1.0.4 Orthogonal vectors and matrices, page 41-

Definition: Two vectors a and b of size $n \times 1$ are said to be *orthogonal* if $a'b = 0$. **Definition:** A set of $p \times 1$ vectors $c_1, c_2, ..., c_n$ that are normalized and pairwise orthogonal is said to be *orthogonal.* **Definition:** A set of $p \times 1$ vectors $c_1, c_2, ..., c_n$ that are normalized and pairwise orthogonal is said to be *orthogonal.*

1) $C'C = I$, $CC' = I$

2) Mulitplication by an orthogonal matrix is equivalent to axes rotation therefore, if $z = Cx$ then $z'z = (Cx)'(Cx) = x'C'Cx = x'Ix = x'x$.

If A is any $p \times p$ matrix and C is orthogonal:

1) $|C| = +1 \ \ or \ \ -1$

2) $|C'AC| = |A|$

3) $-1 \geq c_{ij} \geq 1$, where c_{ij} is any element of C.

1.0.5 Trace, page 44-

1) If A and B are $n \times n$ matrices, then $tr(A \pm B) = tr(A) \pm tr(B)$

2) If A is $n \times p$ and B is $p \times n$, then $tr(AB) = tr(BA)$

3) If A is $n \times p$, then $tr(A'A) = \sum_{i=1}^{p} a_i'a_i$, where a_i is the i^{th} *column* of A AND $tr(AA') = \sum_{i=1}^{n} a_i'a_i$, where a_i is the i^{th} *row* of A OR $tr(A'A) = tr(AA') = \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2$

4) If A is any $n \times n$ matrix and P is any $n \times n$ nonsingular matrix, then $tr(P^{-1}AP) = tr(A)$

5) If A is any $n \times n$ matrix and C is any $n \times n$ orthogonal matrix, then $tr(C'AC) = tr(A)$

6) If A is any $n \times p$ of rank r and A^- is a generalized inverse of A, then $tr(A^-A) = tr(AA^-) = r$

1.0.6 Idempotent matrices, page 54,55

1.0.7 Vector and matrix calculus, page 56-

Theorem 2.14.a, p.56: Let $u = \boldsymbol{a}'\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{a}$ where $\boldsymbol{a}' = (a_1, a_2, ..., a_p)$ is a vector of constants. Then

$$\frac{\partial u}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}'\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

Theorem 2.14.b, p.56: Let $u = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A}$ is a symmetric matrix of constants. Then

$$\frac{\partial u}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$$

Theorem 2.14.c, p.57: Let $u = tr(\boldsymbol{X}\boldsymbol{A})$ where $\boldsymbol{X}$ is a $p \times p$ positive definite matrix and $\boldsymbol{A}$ is a $p \times p$ matrix of constants. Then

$$\frac{\partial u}{\partial \boldsymbol{X}} = \frac{\partial [tr(\boldsymbol{X}\boldsymbol{A})]}{\partial \boldsymbol{X}} = \boldsymbol{A} + \boldsymbol{A}' - diag\boldsymbol{A}$$

Theorem 2.14.d, p.58: Let $u = ln|\boldsymbol{X}|$ where $\boldsymbol{X}$ is a $p \times p$ Then

$$\frac{\partial ln|\boldsymbol{X}|}{\partial \boldsymbol{X}} = 2\boldsymbol{X}^{-1} - diag(\boldsymbol{X}^{-1})$$

Theorem 2.14.e, p.58: Let $\boldsymbol{A}$ be nonsingular of order n with derivative $\frac{\partial \boldsymbol{A}}{\partial x}$ Then

$$\frac{\partial \boldsymbol{A}^{-1}}{\partial x} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$$

Theorem 2.14.f, p.58: Let $\boldsymbol{A}$ be $n \times n$ positive definite matrix. Then

$$\frac{\partial ln|\boldsymbol{A}|}{\partial x} = tr\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\right)$$

1.0.8 Means and variances, page 80-

Theorem 3.6.a,b, Corollary 1, p.80-1: Let $\boldsymbol{a}$ be a $p \times 1$ vector of constatns and $\boldsymbol{y}$ be a $p \times 1$ random vector with mean vector $\boldsymbol{\mu}$. Also, if $\boldsymbol{X}$ is a random matrix, and $\boldsymbol{a}$, $\boldsymbol{b}$ are vectors of constants, and $\boldsymbol{A}$, $\boldsymbol{B}$ are matrices of constants, then

(i) $\mu_z = E(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{a}'E(\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{\mu}$

(ii) $E(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}E(\boldsymbol{y})$

(iii) $E(\boldsymbol{a}'\boldsymbol{X}\boldsymbol{b}) = \boldsymbol{a}'E(\boldsymbol{X})\boldsymbol{b}$

(iv) $E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$

(v) $E(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}) = \boldsymbol{A}E(\boldsymbol{y}) + \boldsymbol{b}$

Theorem 3.6.c, Corollary 1, p.80-1, 76: Let $\boldsymbol{a}$ be a $p \times 1$ vector of constatns and $\boldsymbol{y}$ be a $p \times 1$ random vector with covariance matrix $\boldsymbol{\Sigma}$. Also, let $\boldsymbol{a}$, $\boldsymbol{b}$ be $p \times 1$ (or other dimension (play it by ear)), and $\boldsymbol{A}$ be $k \times p$ matrix of constants, $\boldsymbol{B}$ be an $m \times p$ matrix of constants, then

(i) $\sigma^2 = var(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$

(ii) $cov(\boldsymbol{a}'\boldsymbol{y}, \boldsymbol{b}'\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{b}$

(iii) $cov(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$

(iv) $cov(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$

(v) $cov(\boldsymbol{A}\boldsymbol{y}, \boldsymbol{B}\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{x}}\boldsymbol{B}'$

(vi) $\boldsymbol{\Sigma} = E[(\boldsymbol{y} - \boldsymbol{\mu})][(\boldsymbol{y} - \boldsymbol{\mu})'] = E(\boldsymbol{y}\boldsymbol{y}' - \boldsymbol{\mu}\boldsymbol{\mu}')$

(vii) p. 107 $E(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$

1.0.9 Theorems related to normal distribution

Theorem 4.4a, p.92: Let $p \times 1$ random vector $\boldsymbol{y}$ be $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $\boldsymbol{a}$ be any $p \times 1$ vector of constants, and let $\boldsymbol{A}$ be any $k \times p$ matrix of constatns with rank $k \leq p$. Then

(i) $z = \boldsymbol{a}'\boldsymbol{y}$ is $N(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$

(ii) $z = \boldsymbol{A}'\boldsymbol{y}$ is $N(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$

Theorem 7.6b, p.159: Suppose that $\boldsymbol{y}$ is $N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, where $\boldsymbol{X}$ is $n \times (k+1)$ of rank $k+1 < n$ and $\beta = (\beta_0, \beta_1, ..., \beta_k)'$. Then the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ given in theorem 7.6a have the following distributional properties:

(i) $\hat{\boldsymbol{\beta}}$ is $N_{k+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$

(ii) $\frac{n\hat{\sigma}^2}{\sigma^2}$ is χ^2_{n-k-1}, or equivalently, $\frac{(n-k-1)s^2}{\sigma^2}$ is χ^2_{n-k-1}

(iii) $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ (or s^2) are independent.

Theorem 8.4a, p.199: If $\boldsymbol{y}$ is distributed $N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{C}$ is $q \times (k+1)$, then:

(i) $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ is $N_q(\boldsymbol{C}\boldsymbol{\beta}, \sigma^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}')$.

(ii) $\frac{SSH}{\sigma^2} = \frac{(\boldsymbol{C}\hat{\boldsymbol{\beta}})'\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}}{\sigma^2}$ is $\chi^2_{q,\lambda}$, where $\lambda = \frac{(\boldsymbol{C}\boldsymbol{\beta})'\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}\boldsymbol{\beta}}{2\sigma^2}$.

(iii) $\frac{SSE}{\sigma^2} = \frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X})\boldsymbol{y}}{\sigma^2}$ is χ^2_{n-k-1}.

(iv) SSH and SSE are independent.

Theorem 8.4b, p.199: If $\boldsymbol{y}$ be $N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ and define the statistic:

$$F = \frac{SSH/q}{SSE/(n-k-1)} = \frac{(\boldsymbol{C}\hat{\boldsymbol{\beta}})'\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}/q}{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X})\boldsymbol{y}/(n-k-1)}$$

where $\boldsymbol{C}$ is $q \times (k+1)$ of rank $q \leq k+1$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ the distribution of F is as follows:

(i) If $H_0: \boldsymbol{C}\boldsymbol{\beta} = 0$ is false, then F is distributed as $F_{(q,\ n-k-1,\ \lambda)}$, where $\lambda = \frac{(\boldsymbol{C}\boldsymbol{\beta})'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}\boldsymbol{C}\boldsymbol{\beta}}{2\sigma^2}$.

(ii) If $H_0: \boldsymbol{C}\boldsymbol{\beta} = 0$ is true, then F is distributed as $F_{(q,\ n-k-1)}$.

2 Probability, Book: STATISTICAL INFERENCE, Casella & Berger, Second Edition

2.1 Distributions

2.1.1 Normal

2.1.2 Geometric

2.1.3 Hypergeometric

2.1.4 Binomial

2.1.5 Poisson

2.1.6 Gamma, p. 63

Support $0 < x < \infty$, $\alpha, \beta > 0$. Density and MGF:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad MGF_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha$$

Transformations:

1) If $X \sim \Gamma(\alpha, \beta)$ and $\alpha = 1$ then If $X \sim exp(\beta)$
2) If $X_i \overset{independent}{\sim} \Gamma(\alpha_i, \beta)$ then $\sum_{i=1}^n X_i \sim (p.)\Gamma(\sum_{i=1}^n \alpha, \beta)$, page 183, proof using MGF.
3) If $X_i \overset{iid}{\sim} \Gamma(\alpha, \beta)$ and $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ then $\bar{X} \sim \Gamma(\alpha \cdot n, \beta/n)$.
4) If $X \sim \Gamma(\alpha, \beta)$ and $\beta = 2$ then $X \sim \chi^2_{2\alpha}$
5) If $X \sim \Gamma(\alpha, \beta)$ and $\alpha = 3/2$ and $Y = \sqrt{X/\beta}$ then $Y \sim Maxwell$
6) If $X \sim \Gamma(\alpha, \beta)$ and $Y = 1/X$ then If $Y \sim Inverted\ Gamma$
7) If $X_i \sim \Gamma(\alpha_i, \beta)\ for\ i = 1, 2$ and $Y = \frac{X_1}{X_1 + X_2}$ then $Y \sim Beta(\alpha_1, \alpha_2)$
8) If $X \sim \Gamma(\alpha, \beta)$ and then $cX \sim \Gamma(\alpha, c\beta)$
9) If $U_i \overset{iid}{\sim} Uniform[0, 1]$ and then $-\sum_{i=1}^n ln(U_i) \sim \Gamma(n, 1)$
10) If $X \sim \Gamma(\alpha \in \boldsymbol{Z}, \beta)$ and $Y \sim Pois\left(\frac{x}{\beta}\right)$ then $P\{X > x\} = P\{Y < \alpha\}$
11) If $X_n \sim \Gamma(\alpha_n, \beta)$ and $\alpha_n \to \infty$ then $X_n\ is\ distributed\ more\ like\ Normal(\alpha_n\beta,\ \alpha_n\beta^2)$
12) see more in Wikipedia

2.1.7 Beta

2.1.8 Moment Generating function

2.2 Conditional Moments Identities

2.3 Inequalities

2.4 Wald, Likelihood, Score tests

2.5 Delta method

2.5.1 First order method

If a sequence of random variables satisfies: $\sqrt{n}(x_n - \theta) \to\ N(0, \sigma^2)$ in distribution and $g(\theta)$ is a function with existing derivative at point θ and such that $g(\theta) \neq 0$. Then:

$$\sqrt{n}(g(x_n) - g(\theta)) \overset{d}{\to}\ N(0, \sigma^2 \cdot (g'(\theta))^2)$$

Proof: From Taylor series expansion we have:

$$
\begin{aligned}
g(x_n) &= g(\theta) + \frac{g'(\theta)(x_n-\theta)}{1!} + R(x_n,\theta) \\
g(x_n) - g(\theta) &= g'(\theta)(x_n-\theta) + R(x_n,\theta) \\
\sqrt{n}(g(x_n) - g(\theta)) &= g'(\theta)\sqrt{n}(x_n-\theta) + \sqrt{n}R(x_n,\theta)
\end{aligned}
$$

Because we somehow know that $\sqrt{n}R(x_n,\theta) \to 0$, it follows that $\sqrt{n}(g(x_n) - g(\theta)) \xrightarrow{d} N(0, \sigma^2 \cdot (g'(\theta))^2)$.

2.5.2 Second order method (Bryan's class notes)

If a sequence of random variables satisfies: $\sqrt{n}(x_n - \theta) \to N(0,\sigma^2)$ in distribution and $g(\theta)$ is a function with existing derivative at point θ and such that $g(\theta) = 0$. Then:

$$
\frac{2n(g(x_n) - g(\theta))}{\sigma^2 g''(\theta)} \xrightarrow{d} \chi^2_1
$$

Proof: From Taylor series expansion we have:

$$
\begin{aligned}
g(x_n) &= g(\theta) + \frac{g'(\theta)(x_n-\theta)}{1!} + \frac{g''(\theta)(x_n-\theta)^2}{2!} + R^*(x_n,\theta) \\
g(x_n) &= g(\theta) + \frac{g''(\theta)(x_n-\theta)^2}{2!} + R^*(x_n,\theta) \\
n(g(x_n) - g(\theta)) &= n\frac{\sigma^2}{\sigma^2}\frac{g''(\theta)(x_n-\theta)^2}{2!} + n \cdot R^*(x_n,\theta)
\end{aligned}
$$

Because again, we somehow know that $n \cdot R^*(x_n,\theta) \to 0$, and $\frac{n(x_n-\theta)^2}{\sigma^2} \xrightarrow{d} \chi^2_1$ the result follows.

2.5.3 Multivariate method (Bryan's class notes)

Suppose we have a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$ and its estimate $\hat{\boldsymbol{\theta}}$.
And also suppose $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\sigma}$ is a variance covariance matrix for $\boldsymbol{\theta}$. Then, we have:

$$
\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{d} N(\mathbf{0}, g'(\boldsymbol{\theta})^T \boldsymbol{\Sigma}\ g'(\boldsymbol{\theta}))
$$

Where $g'(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1}g(\boldsymbol{\theta}), \frac{\partial}{\partial\theta_2}g(\boldsymbol{\theta}), ..., \frac{\partial}{\partial\theta_p}g(\boldsymbol{\theta})\right)$

3 Advanced probability. Book: *Probability. Alan F.Karr*

3.1 Sets and measures

Definition 1.8, p.20: Let $A_1, A_2, A_3, ...$ and A be subsets of Ω.

a) The *lim sup* of (A_n) is the set of ω such that $\omega \in A_n$ for *infinitely many* values of n:

$$\{A_n,\ i.o.\} \stackrel{def}{=} \limsup_n A_n = \cap_{k=1}^{\infty} \cup_{n=k}^{\infty} A_n$$

b) The *lim inf* of (A_n) is the set of ω such that $\omega \in A_n$ for *all byt infinitely many* values of n:

$$\{A_n,\ ult.\} \stackrel{def}{=} \liminf_n A_n = \cup_{k=1}^{\infty} \cap_{n=k}^{\infty} A_n$$

c) The sequence (A_n) *converges to* A, which we write as $A = \lim_{n\to\infty} A_n$ or simply $A_n \to A$ if :

$$\limsup_n A_n = \liminf_n A_n = A$$

Theorem 1.27 *(Borel-Cantelli lemma)*, p.27: If $\sum_{n=1}^{\infty} P(A_n) < \infty$ then $P\{A_n, i.o.\} = 0$
Theorem 3.23 *(Borel-Cantelli lemma)*, p.81: If $A_1, A_2, A_3, ...$ are independent events such that $\sum_{n=1}^{\infty} P(A_n) = \infty$ then $P\{A_n, i.o.\} = 1$

3.2 Types of convergence and implications among forms of convergence, p. 135-

Definition: Sequence X_n converges to X ***almost surely*** (probabilistic version of pointwise conversion), denoted by $X_n \xrightarrow{a.s.} X$, if

$$P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$$

Definition: Sequence X_n converges to X ***in probability***, denoted by $X_n \xrightarrow{P} X$, if

$$\lim_{x\to\infty} P\{|X_n - X|\} = 0$$

Definition: Suppose $X, X_1, X_2, ..., X_i, ...$ belongs to L^2, then X_n converges to X ***in quadratic mean***, denoted by $X_n \xrightarrow{q.m.} X$ or $X_n \xrightarrow{L^2} X$, if

$$\lim_{x\to\infty} E[(X_n - X)^2] = 0$$

Definition: Suppose $X, X_1, X_2, ..., X_i, ... \in L^1$, then X_n converges to X ***in*** L^1, denoted by $X_n \xrightarrow{q.m.} X$ or $X_n \xrightarrow{L^1} X$, if

$$\lim_{x\to\infty} E[|X_n - X|] = 0$$

Definition: The sequence (X_n) converges to X ***in distribution***, denoted by $X_n \xrightarrow{q.m.} X$ or $X_n \xrightarrow{d} X$, if

$$\lim_{x\to\infty} F_{X_n} t = F_X(t)$$

for all t at which F_X is continuous.
Definition 5.15, p.142: A sequence (X_n) is ***uniformly integrable*** if $X_n \in L^1$ for each n and if

$$\lim_{a\to\infty} \sup_n E[|X_n|; \{|X_n| > a\}] = 0$$

★, **p.143** Any sequence (X_n) dominated by an element from L^1 is uniformly integrable.
Definition, p.143: A sequence (X_n) is ***uniformly absolutely continuous*** if for each $\varepsilon > 0$ there is $\delta > 0$ such that

$$\sup_n E[|X_n|; A] < \varepsilon, \quad \textit{for } \forall A \textit{ such that } P(A) < \delta$$

Proposition, p.138: If $\sum_{n=1}^{\infty} P\{|X_n - X| > \varepsilon\} < \infty$ for every $\varepsilon > 0$, then $X_n \xrightarrow{a.s.} X$.

3.2.1 Implications, theorems (p. 138,140,-)

$$
\begin{array}{lllll}
L^2 \implies & L^1 \implies & \textit{In Probability} \implies & & \textit{In Distribution} \\
& L^1 \xleftarrow{Uniform Integr.} & \textit{In Probability} & & \\
& \textit{Almost Surely} \implies & \textit{In Probability} \implies & & \textit{In Distribution} \\
& & \textit{In Probability} & \xleftarrow{Constant\ Limit} & \textit{In Distribution}
\end{array}
$$

3.2.2 Algebraic operations and continuous mapping theorems (p. 145,-)

Theorems 5.19, 5.20, 5.21, 5.22, 5.23, p.145-8 Let X, X_n, Y, and Y_n be random variables, and $\odot \in \{a.s., P, q.m., L^1\}$, then it follows:

a) If $X_n \xrightarrow{\odot} X$ and $Y_n \xrightarrow{\odot} Y$, then $X_n + Y_n \xrightarrow{\odot} X + Y$

b) If $X_n \xrightarrow{\odot} X$ and $Y_n \xrightarrow{\odot} Y$, then $X_n Y_n \xrightarrow{\odot} XY$

c) **Slutsky's theorem** If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c \in \mathbb{R}$, then $X_n + Y_n \xrightarrow{d} X + c$

d) **Slutsky's theorem, bis** If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c \in \mathbb{R}$, then $X_n Y_n \xrightarrow{d} cX$

e) **Continuous mappings** Let $g : \mathbb{R} \to \mathbb{R}$ be a continuouts function. And let $\circledast \in \{a.s., P, d\}$. Then we have that if $X_n \xrightarrow{\circledast} X$, then $g(X_n) \xrightarrow{\circledast} g(X)$

3.3 Characteristic functions (p. 163-)

This chapter uses notations and facts for *imaginary numbers*:

$$
\begin{aligned}
&z = x + y \cdot i = \Re(i) + \Im(i) \cdot i \\
&\bar{z}\ is\ conjugate\ of\ z\ if\ z = x + y \cdot i \quad and \quad \bar{z} = x - y \cdot i \\
&|z| = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}}\ (modulus\ of\ z) \\
&e^{it} = cos(t) + i \cdot sin(t) \quad (Euler's\ formula)
\end{aligned}
$$

Definition: The ***characteristic function*** of a random variable X is the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itX} dF_X(x)$$

★ $|\varphi_X(0)| = 1$.
★ Characteristic function always exists because for all t: $|\varphi_X(t)| \leq 1$.
★ Characteristic function is *uniformly continuous*, where ***Uniform Continuity*** means that for any A and any ε, there is such $\delta > 0$ that $\sup\limits_{(x,y)\in A^2,\ |x-y|<} |\varphi(x) - \varphi(y)| < \varepsilon$.

Proposition 6.2 (p.164): The characteristic function is uniformly continuous and $\varphi_X(-t) = \overline{\varphi_X(t)},\ \forall t$.
Theorem 6.4 (p.165): If X and Y are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.
Theorem 6.5 (p.166): Whenever $a < b \in \mathbb{R}$ are continuity points of F_x,

$$P\{a < X < b\} = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{itb}}{it} \varphi_X(t) dt$$

Theorem 6.6 (p.167): If $\varphi_X(t) = \varphi_Y(t),\ \forall t$ then X and Y have the same distribution.
Theorem 6.7 (p.167) (Fourier inversion theorem): If

$$\int_{-\infty}^{\infty} |\varphi_X(t)| dt < \infty$$

then X is absolutely continuous with density

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itX} \varphi_X(t) dt$$

Theorem 6.11 (p.169): If $E[|X|^k] < \infty$ then the derivative $\varphi_X^{(k)}(t)$ exists and

$$E[X^k] = i^{-k} \varphi_X^{(k)}(0)$$

Theorem 6.12 (p.170): Let k be and even integer, and suppose that $\varphi_X^{(k)}(0)$ exists then, $E[|X^k|] < \infty$.
Theorem 6.15 (p.171): Suppose we have $X_n \xrightarrow{d} X$ **if and only if** $\varphi_{X_n}(t) \to \varphi_X(t)$, $\forall t \in \mathbb{R}$
Some Theorem not from the book: If X and Y areindependent real random variables with characteristic functions φ and ϕ respectively, then $Ee^{itXY} = E\varphi(tY) = E\phi(tX)$

3.4 Prediction and conditional expectation

See page 217: Let's assume that random variable Z is defined on a probability space $(\Omega, \mathcal{F}, P)$ and $Z \in L^2$ ($E[Z^2] < \infty$). We want to predict unobserved X using observed $Y_1, Y_2, ..., Y_n$. The book says that we predict X using some allowable functions of Y_i. So we need to build a *linear predictor*: $Z = f(Y_1, Y_2, ..., Y_n) = a_1^* \cdot Y_1 + a_2^* \cdot Y_1 + a_3^* \cdot Y_1 + ... + a_1^* \cdot Y_n$.
Definition(p. 217): the **mean squared error of Z (MSE)** is:

$$MSE(Z) = E[(Z - X)^2]$$

Assuming that the optimal predictor exists, **Definition**(p. 217): the **minimum mean squared error predictor (MMSE)** of X (let's denote it as $\hat{X}$), then satisfies:

$$E[(\hat{X} - X)^2] \leq E[(Z - X)^2], \quad for\ \forall\ Z\ \in\ V$$

Definition(my own definition (could not find it in the book)): **orthonormal base of space** V^n is a set of vectors Z_i ($i \in \{1, 2, ..., n\}$) such that there exists such numbers $a_i, a_2, ..., a_n$:

$$v = \sum_{i=1}^{n} a_i Z_i, \quad \forall\ v \in V^n$$
$$and \quad \langle Z_i Z_j \rangle = 0, \quad \forall\ i \neq j \quad and \quad ||Z_i|| = 1, \quad \forall\ i$$

Proposition 8.11 The MMSE linear predictor of $X \in L^2$ within V is $\hat{X} = \sum_{i=1}^{n} = a_i^* Y_i$, where $a_1^*, a_2^*, ..., a_n^*$ are such that

$$\sum_{j=1}^{n} \langle Y_i Y_j \rangle a_j^* = \langle Y_i X \rangle, \quad i = 1, 2, ...n \qquad for\ any\ basis\ Y_1, Y_2, ..., Y_n$$
$$a_i^* = \langle Y_i X \rangle,\ so\ \hat{X} = \sum_{j=1}^{n} \langle Y_i X \rangle Y_i, \quad i = 1, 2, ...n \quad for\ orthonormal\ basis\ Y_1, Y_2, ..., Y_n$$

Example If $V = \{aY + b : a, b \in R\}$, where $Y \in L^2$ and $\sigma^2 = Var(Y) > 0$ then it can be shown that $Z_1 = (Y - E[Y])/\sigma$ and $Z_2 \equiv 1$ is the orthonormal basis, and MMSE, $\hat{X} = \frac{cov(X,Y)}{\sigma^2}(Y - E[Y]) + E[X]$.

4 Survival analysis (under construction)

4.1 Tests

4.1.1 Log-rank test

4.2 Hazard, survival functions and identities

5 Maximization, Lagrange Multiplier, Karush-Kuhn-Tucker conditions (KKT)

We skip the KKT theorem here and give an example of minimization with inequality constraints. The example is taken from `http://myweb.clemson.edu/ pbelott/bulk/teaching/lehigh/ie426-f09/lecture20.pdf`. We have the following minimization problem:

$$\begin{aligned} \min\ & X^2 + 2y^2 \\ & x + y \geq 3 \\ & y - x^2 \geq 1 \end{aligned}$$

The Lagrangian is:

$$\begin{aligned} \mathcal{L} = x^2 + 2y^2 - \lambda_1(x + y - 3) + &\quad \lambda_2(y - x^2 - 1) \\ \lambda_1 \geq 0, &\qquad\qquad \lambda_2 \geq 0 \end{aligned}$$

Note that when we *minimize* we take the main expression with + sign and we **subtract** the equality constraint the way it is shown above. For inequality, we write it in a standard form (with variables on the left side, with $\geq$ or $>$ signs and a constant on the right), and also subtract it. When we *maximize* or when the inequality constraits are written in a different way, the $+$ and $-$ signs might work differently.

The KKT conditions are:

$$\begin{aligned} &A)\ \ x + y \geq 3 \\ &B)\ \ y - x^2 \geq 1 \\ &C)\ \ \partial\mathcal{L}/\partial x = 2x - \lambda_1 + 2\lambda_2 = 0 \\ &D)\ \ \partial\mathcal{L}/\partial y = 4y - \lambda_1 - \lambda_2 = 0 \\ &E)\ \ \lambda_1(x + y - 3) = 0 \\ &F)\ \ \lambda_2(y - x^2 - 1) = 0 \\ &G)\ \ \lambda_1,\ \lambda_2 \geq 0 \end{aligned}$$

We have to consider four cases:

1) $\lambda_1 = \lambda_2 = 0$. Then C and D become $x = 0$ and $y = 0$, which violates A.

2) $\lambda_1 > 0,\ \lambda_2 = 0$. From E it follows that $x + y - 3 = 0$. From C and D we have: $2x - \lambda_1 = 0$ and $4y - \lambda_1 = 0$, which leads to a unique solution $(x, y) = (2, 1)$, which violates $y - x^2 \geq 1$.

3) $\lambda_1 = 0,\ \lambda_2 > 0$. From F we have $y - x^2 - 1 = 0$. From C we have: $2x + 2\lambda_2 x = 2x(1 + \lambda_2) = 0$. Because $\lambda_2 > 0$ we must have $x = 0$. From D we have $4y - \lambda_2 = 0$. Because $x = 0$ we have $y = 0 + 1 = 1$, which violates $x + y \geq 3$.

4) $\lambda_1 > 0,\ \lambda_2 > 0$. From E and F we have: $y - x^2 - 1 = 0$ and $x + y = 3$, therefore $(3 - x) - x^2 - 1 = 0$ or $x^2 + x - 2 = 0$. This equation has two solutions $x \in \{-2, 1\}$, therefore $(x, y) \in \{(-2, 5), (1, 2)\}$. For $(x, y) = (-2, 5)$, equation C becomes $-4 - \lambda_1 - 4\lambda_2 = 0$, which is impossible for $\lambda_1, \lambda_2 \geq 0$. For $(x, y) = (1, 2)$ we solve C and D and get $\lambda_1 = 6$, $\lambda_2 = 2$. This is the only point that satisfies KKT so this is the solution.

6 What to practice/ to code

1) Wald test (with C matrix), chunk test

2) Plot coefficient effects.

3) Spaghetti plot

4)

5)

7 Longitudinal Data Analysis

Estimation of β can be done through **response feature analysis** (two-step approach), or **weighted least squares**.
Estimation of Σ can be done through **robust sandwich estimator**, **maximum likelihood (ML)**, **restricted maximum likelihood (REML)**

8 OLS vs WLS

When $\mathbf{\Sigma} \neq \sigma^2 I$ there are two sources of error (slide 42, mod.1, ch.2):

- $\hat{\sigma}2$ is **biased**
- $\sigma^2(X^TX)^{-1}$ is **not correct** estimate for $cov(\hat{\beta})$

In order to fix this problem we can do robust sandwich estimator.

	OLS	WLS
Minimize		$\sum_{i=1}^n (Y_i - X_i\beta)^T W_i (Y_i - X_i\beta)$
Score	$X^T(Y - X\beta) = 0$	$X^T W(Y - X\beta) = 0$
$\hat{\beta}$	$(X^TX)^{-1}X^TY$	$\left(\sum_{i=1}^n X_i^T W_i X_i\right)^{-1} \left(\sum_{i=1}^n X_i^T W_i Y_i\right)$
Covariance	$\sigma^2(X^TX)^{-1}$	$\left(\sum_{i=1}^n X_i^T W_i X_i\right)^{-1}$
Sandwich	$(X^TX)^{-1}(X^T\hat{\sigma}^2 X)(X^TX)^{-1}$	$\left(\sum_{i=1}^n X_i^T W_i X_i\right)^{-1} \left(\sum_{i=1}^n X_i^T W_i \hat{\mathbf{\Sigma}}_0 W_i X_i\right) \left(\sum_{i=1}^n X_i^T W_i X_i\right)^{-1}$
cov estimator	$\hat{\sigma}^2 = \frac{1}{mn-p} \sum_{i=1}^n (Y_i - X_i\hat{\beta})^T (Y_i - X_i\hat{\beta})$	$\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i\tilde{\beta})^T (Y_i - X_i\tilde{\beta})$

9 Models for Longitudinal Data

9.1 Interpretation of coefficients

Population level 1: β_1 is expected change in weight difference b/w subjects associated difference of $1kg$ of weight at enrolled time holding other variable constant and assuming that no other factors explain this difference.
Population level 2: β_1 is expected difference in weight b/w two population of subjects that differ in weight by $1kg$ at enrolled time holding other variable constant and assuming that no other factors explain this difference.
Population level OR: The odds of receiving individualized information for treatment group were 12 percent less compared to the control group (OR=.88, 95%CI: [.78, 1.00]), assuming that these groups were identical given the rest of the covariates. In other words, we have evidence that the treatment reduced the odds of receiving the individualized information.

9.2 Alternating Logistic Regression

9.2.1 Model Definition

When we have binary longitudinal data, defining covariance structure is a problem because correlation for binary variables is limited by some expression of their means.
Alternating Logistic regression offeres a different approach. We need to solve $\sum_i D_i^T V_i^{-1}(Y_i - \mu_i) = 0$, where V_i is a working covariance structure. Instead of defining it, we assume:

$$\begin{aligned}
&log(E(Y_{ij}|Y_{ij'})) = \Delta_{ijj'} + \alpha Y_{ij'} \\
&\qquad\qquad or \\
&log(E(Y_{ij}|Y_{ij'})) = \Delta_{ijj'} + (\alpha_0 + \alpha_1[|j - j'|] - 1)Y_{ij'} \\
&\qquad\qquad or \\
&log(E(Y_{ij}|Y_{ij'})) = \Delta_{ijj'} + \alpha_0 + [\alpha_1 I[|j - j'| == 2) + \alpha_2 I[|j - j'| == 3) + \alpha_3 I[|j - j'| == 4) + ...]Y_{ij'}
\end{aligned}$$

9.2.2 Method of Solving

9.2.3 Hypothesis Testing

9.2.4 R-code

If data looks like this

id	time
11	0
11	1
11	2

Then the z.matrix is built like this:

j	j'	(j-j')
0	0	0
0	1	-1
0	2	-2
1	0	1
1	1	0
1	2	-1
2	0	2
2	1	1
2	2	0

So we can run the following code:

```
z.mat <- NULL
for (i in unique(data1$id)){
#for (i in c("1", "2")){
    tmp <- data1[data1$id==i,]
    tmp1 <- expand.grid(tmp$month,tmp$month)
    tmp2 <- tmp1[,1]-tmp1[,2]
    tmp2 <- tmp2[tmp2>0]
    len <- length(tmp2)
zForID = cbind(rep(tmp$id[1], len), tmp2, rep(tmp$gender[1], len), rep(tmp$age[1], len),
tmp2==2, tmp2==3, tmp2==4, tmp2==5, tmp2==6, tmp2>=7)
    z.mat <- rbind(z.mat, zForID)
}
z.mat <- data.frame(z.mat)
names(z.mat) <- c("id", "timediff", "gender", "age", "timeDiff=2", "timeDiff=3", "timeDiff=4",
"timeDiff=5", "timeDiff=6", "timeDiff>=7")
z.mat$timediff1 <- z.mat$timediff-1
z.mat$one <- 1

mod1b = ordgee(ordered(thought) ~ month*age + month*gender, data=data1,
z=z.mat[,c("one","gender", "timeDiff=2", "timeDiff=3", "timeDiff=4", "timeDiff=5", "timeDiff=6",
"timeDiff>=7")], corstr="userdefined")
```

9.2.5 Parameter interpretation

9.3 Generalized Linear Mixed Effects Models (GLMM)

9.3.1 Model Definition

9.3.2 Method of Solving

9.3.3 Hypothesis Testing

9.3.4 Parameter interpretation

Outcome	Coefficient	Random Intercept	Random Intercept/Slope
Continuous	Intercept	Marginal	Marginal
	Slope	Marginal	Marginal
Count	Intercept	Conditional	Conditional
	Slope	Marginal	Conditional
Binary	Intercept	Conditional	Conditional
	Slope	Conditional	Conditional

9.3.5 Model classification

	Parametric Likelihood	Semiparametric
Marginal	MM-LV, MM-T, MM-TLV	GEE-I, GEE-AR, ALR
Conditional (subject specific)	GLMM-RI, GLMM-RL	
Conditional (transitional)	GLTM	

9.3.6 Model Summary

Model	Description	Interp	Lik/Sem ML-max.lik RE-restr.ML CL-cond.lik QL-quasi lik	package	call	Mod.,Sl.	Section in this doc

GEE 1.0	$\sum D_i V_i(\alpha)(y_i - \mu_i) = 0$
	poisson, binary
		Popul	QL=semi		`geepack,`
					`geeglm`
					`geese`
						3
						3
						3	

GEE 2.0	GEE with Indep.	Popul	semi			3	
GEE-I (1.0 but it's not)	GEE with Indep. GEE with Indep.	Popul Popul	semi semi				
GEE-AR (1.0)	GEE with AR(1)	Popul	semi				
GLMM-RI/RL continuous count binary	Mixed Effects, Rand. Int. /+Slope	 Subj Subj Subj	ML,RE,CL		`glmer`	4	
GLMM-RL continuous count binary	Mixed Effects, Rand. Int. /+Slope	 Subj Subj Subj	ML,RE,CL		`glmer`	4	
ALR≡GEE 1.5	Alternating Logist. Regr. of response dependence ALR is just an example	 Popul.	semi semi semi		`ordgee` `ordgee` `ordgee`	3,sl.64 3,sl.64 3,sl.64	
GLTM	Gen.Lin.Transional	Popul.$\vert Y_{j-1}$	CL		`glm`	5,sl.23	
MM-LV	Marginalized, Latent Var	Popul	lik		`mm`	5	
MM-T	Marginalized, Transitional	Popul	lik		`mm`	5	
MM-TLV	Marginalized, Trans. Marginalized, Latent Var Marginalized, both	Popul Popul Popul	lik lik lik		`mm` `mm` `mm`	5 5 5	

Is GEE 2 quisi lik? Choosing b/w model wiht a time lag, which model to use
Populaiton

9.3.7 Missing

For gee, glm, we can compute weights =1/(probability of being observed) and this is how we can deal with missing data.

9.3.8 R-code

9.4 Random facts

GLM + robust standard errors = GEE + independence covariate structures (GEE-I)
If you fit GEE 2.0 then the second moment has to be something you are interested in. You estimate them together, but if you get second moment wrong, then the first moment might be wrong.
In GLMM(glmer() function) for not continuous outcome, if we get the correlaiton structure wrong then the first model would be biased.

10 Computational Statistic

10.1 Metropolis-Hastings Algorithm

The following is from the Wikipedia:

1) Choose and arbitrary point x_0