Advanced Topics in Statistics

October 29, 2013

Author: James Brofos
james.a.brofos.15@dartmouth.edu
http://www.cs.dartmouth.edu/∼james/
Dartmouth College

Abstract

Presented here is a detailed discussion of select topics in statistics. We offer notes on linear algebra as it pertains to statistics, maximum likelihood theory, matrix formulations of least-squares, partial correlation, Fisher information, statistical power, expectation-maximization, and logistic regression. We hope that these notes will be useful to students of statistics and will serve as a resource for those wishing to refresh their memory. Notice that these notes are not intended as an introduction to statistics, and some prior understanding will be necessary for reading this document.

1 Linear Algebra

1.1 Fundamentals

Suppose we have a vector $x^{n\times 1} \in \mathbb{R}^n$ where $x = \{x_1, x_2, \ldots, x_n\}^T$. If $y \in \mathbb{R}^n$ also, then $x + y = \{x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n\}^T$ and $ax = \{ax_1, ax_2, \ldots, ax_n\}^T$ if a is a scalar (that is, $a \in \mathbb{R}$). Notice in each case that x and y are column vectors. We will maintain this as the convention throughout the text.

We may define also the inner product between two vectors x and y as follows: $(x, y) = \sum_{i=1}^n x_i y_i$. In the special case that $y = x$, we have that $(x, x) = \sum_{i=1}^n x_i^2$, which is the sum of squared elements. The magnitude of the vector x is computed as $\sqrt{(x, x)} = ||x||$.

For vectors $a_1, \ldots, a_p \in \mathbb{R}^n$, we say that these are independent if $\lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_p a_p = \vec{0}$ implies that $\lambda_i = 0 \forall\ i$. Notice that if $p > n$, then the vectors must be linearly dependent. The column vectors $a_1, \ldots, a_p$ may be "concatenated" to generate a matrix. For example $A \in \mathbb{R}^{n\times p}$ may be written as $A = [a_1, a_2, \ldots, a_p]$. For a vector $x \in \mathbb{R}^p$, we have that $A^{n\times p}x^{p\times 1} = (Ax)^{n\times 1} = x_1 a_1 + x_2 a_2 + \ldots + x_p a_p$. And for a matrix B, if the matrix multiplication AB is defined then $B \in \mathbb{R}^{p\times m}$ for some natural number m and the product $AB \in \mathbb{R}^{n\times m}$.

We offer as well some review of matrix transposes. In statistics, it is common to denote the vector transpose as one of either x^T or x'. Both are acceptable and mean the same thing. Particularly, $x^T = x' = \{x_1, x_2, \ldots, x_n\}$, and $(A^{n\times p})' = B^{p\times n}$ such that $a_{ij} = b_{ji}$. In this text, we will use the symbol $\mathbb{1}^{m\times p}$ to refer to a matrix of size $m \times p$ that is filled with unity.

The following is a property of the matrix transpose:

Proposition 1.1 *Let $A \in \mathbb{R}^{n\times m}$. Then the product AA' is symmetric.*

Proof This is straightforward. Notice that $(A')' = A$ and $(AB)' = B'A'$

$$(AA')' = (A')' A' = A'A \tag{1}$$

Thus, AA' is symmetric.

Definition The rank of A, written $\text{rank}\,[A]$, is defined to be the greatest number of columns of A that are linearly independent. The following are properties of the matrix rank:

1. $\text{rank}\,[A] \leq \min[m, p]$.
2. $\text{rank}\,[A] = \text{rank}\,[A']$.
3. $\text{rank}\,[A'A] = \text{rank}\,[AA'] = \text{rank}\,[A]$.
4. $\text{rank}\,[A + B] \leq \text{rank}\,[A] + \text{rank}\,[B]$.
5. $\text{rank}\,[AB] \leq \min[\text{rank}\,[A], \text{rank}\,[B]]$.

Example

$$\text{rank}\left[\mathbb{1}^{n\times 1}\mathbb{1}^{1\times n}\right] = 1 \tag{2}$$

$$\text{rank}\left[\mathbb{1}^{1\times n}\mathbb{1}^{n\times 1}\right] = n \implies \text{rank}\,[n] = 1 \tag{3}$$

Definition The trace of A, written as $\text{tr}\,[A]$, is the sum of the diagonal elements of A. That is $\text{tr}\,[A] = \sum_{i=1}^{n} a_{ii}$. The value of $\text{tr}\,[A'A]$ is the sum of all squared elements. A property of the matrix trace is that $\text{tr}\,[ABC] = \text{tr}\,[BCA] \implies \text{tr}\,[AB] = \text{tr}\,[BA]$, provided the matrix multiplications are defined.

Definition Suppose we have a square matrix $A \in \mathbb{R}^{n\times n}$. Then the determinant of A, written $\det A$, is the hypervolume of the parallelogram generated by A. In the case that $A \in \mathbb{R}^{2\times 2}$, the determinant is easy to calculate:

$$\det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21} \tag{4}$$

Furthermore, the matrix inverse A^{-1} of A is matrix such that $A^{-1}A = \mathbb{I}^{n\times n}$. A^{-1} exists if and only if $\det A \neq 0$. Notice that here we denote the identity matrix of size $n \times n$ and $\mathbb{I}^{n\times n}$.

Remark If $A : \mathbb{R}^3 \to \mathbb{R}^3$ and $\det A = 0$, then A maps to a plane.

Remark For matrices A and B, the inverse of the matrix product AB is $(AB)^{-1} = B^{-1}A^{-1}$.

Definition An orthogonal matrix is a matrix A that has the property $A'A = \mathbb{I}$. That is to say, $A' = A^{-1}$.

Example

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, R'R = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \mathbb{I}^{2\times 2} \tag{5}$$

Example

$$A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}, \det[A - \lambda\mathbb{I}] = 0 \tag{6}$$

$$\implies \det\begin{bmatrix} \cos\theta - \lambda & \sin\theta \\ \sin\theta & -\cos\theta - \lambda \end{bmatrix} = -\left[\cos\theta - \lambda\right]^2 - \sin^2\theta = \lambda^2 - 1 = 0 \tag{7}$$

$$\implies \lambda_1 = 1, \lambda_2 = -1 \tag{8}$$

1.2 Spectral Matrix Decomposition

If A is symmetric then it may be decomposed as $A = P\Lambda P^T$, where P is a matrix of the eigenvectors and $\Lambda = \text{diag}\,[\lambda_1, \ldots, \lambda_n]$ is a matrix diagonalization of the eigenvalues. P may also be normalized such that $P^TP = \mathbb{I} = PP^T$. An equivalent formation of the matrix A is $A = \sum_{i=1}^{n} \lambda_i p_i p_i^T$.

We say that a matrix A is positive definite if the all the eigenvalues of A are greater than zero. That is, if $\lambda_i > 0 \forall\ i$. Using this decomposition, we arrive at a natural way to perform matrix operations. For example, e^A.

Example

$$e^A = Pe^{\Lambda}P^T, e^{\Lambda} = \begin{bmatrix} e^{\lambda_1} & \cdots & \vec{0} \\ \vdots & \ddots & \vdots \\ \vec{0} & \cdots & e^{\lambda_n} \end{bmatrix} \tag{9}$$

Proposition 1.2 *To calculate A^{-1} using this decomposition, we need only invert Λ, which is trivial. In particular, $A^{-1} = P\Lambda^{-1}P^T$.*

Proof

$$\left(P\Lambda^{-1}P^T\right)\left(P\Lambda P^T\right) = \left(P\Lambda^{-1}\right)\mathbb{I}\left(\Lambda P^T\right) = P\mathbb{I}P^T = \mathbb{I} \tag{10}$$

Example $A^{\frac{1}{2}} = P\Lambda^{\frac{1}{2}}P^T$ and $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$. This is because

$$\left(P\Lambda^{\frac{1}{2}}\right)\left(P\Lambda^{\frac{1}{2}}\right) = P\Lambda P^T = A \tag{11}$$

Remark Calculating $A^n = P\Lambda^n P^T \to \mathcal{O}\left(\lambda_{\max}^n\right)$ as $n \to \infty$.

The determinant of the matrix A may be expressed as the product of the eigenvalues. Indeed, $\det A = \prod_{i=1}^n \lambda_i$. Thus, if any $\lambda_i = 0$ then A is not invertible. Furthermore, $\operatorname{tr}[A]$ is the sum of the eigenvalues $\operatorname{tr}[A] = \sum_{i=1}^n \lambda_i = \operatorname{tr}\left[P\Lambda P^T\right] = \operatorname{tr}\left[\Lambda P P^T\right] = \operatorname{tr}[\Lambda]$.

Definition The generalized matrix inverse is $A^+ = P\Lambda^+P^T$ where $\Lambda^+ = \begin{cases} \frac{1}{\lambda_i} & \text{if } \lambda_i \neq 0 \\ 0 & \text{if } \lambda_i = 0 \end{cases}$. This is also called the Moore-Penrose Generalized Inverse. The matrix A^+ has the property that $AA^+A = A$.

Remark The matrix A^+ has additional properties. In particular:

1. $A^+AA^+ = A^+$
2. $AA^+ = \left(AA^+\right)^T$
3. $A^+A = \left(A^+A\right)^T$

Remark For a matrix A that can be decomposed by the spectral method, $A^+ = \lim_{\lambda \to 0}\left(A^TA + \lambda\mathbb{I}\right)^{-1}A^T$.

Example Let $A^{n\times n} = \mathbb{1}^{n\times 1}\mathbb{1}^{1\times n}$. Then $A^+ = \frac{1}{n^2}\mathbb{1}^{n\times 1}\mathbb{1}^{1\times n}$. For example:

$$AA^+A = \left[\mathbb{1}^{n\times 1}\mathbb{1}^{1\times n}\right]\left[\frac{1}{n^2}\mathbb{1}^{n\times 1}\mathbb{1}^{1\times n}\right]\left[\mathbb{1}^{n\times 1}\mathbb{1}^{1\times n}\right] = \frac{1}{n^2}nn\left(\mathbb{1}^{n\times 1}\mathbb{1}^{1\times n}\right) = A \tag{12}$$

The reader is encouraged to work out the additional properties of the generalized inverse for this example at their leisure!

2 Random Variables and Vectors

Let $x_1, \ldots, x_n$ be random variables. Then we can concatenate these random variables into a random vector $\vec{x} = \{x_1, \ldots, x_n\}^T$. Also denote $\vec{y} = \{y_1, \ldots, y_n\}^T$. We may write that $\mu_x = \{\mathbb{E}[x_1], \ldots, \mathbb{E}[x_n]\}^T$ and $\operatorname{cov}[\vec{x}] = \mathbb{E}\left[(x-\mu_x)(x-\mu_x)^T\right]$. Similarly, $\operatorname{cov}[x,y] = \mathbb{E}\left[(x-\mu_x)(y-\mu_y)^T\right]$. We list properties of covariance:

1. $\operatorname{cov}[x,y] = \operatorname{cov}[x,y]^T$
2. $\operatorname{cov}[x+y,z] = \operatorname{cov}[x,z] + \operatorname{cov}[y,z]$
3. If $y = b^Tx$ then $\operatorname{cov}[y] = b^T\operatorname{cov}[x]\,b$
4. If $y = B^{m\times n}x$ then $\operatorname{cov}[y] = B\operatorname{cov}[x]\,B^T$

Remark We present a proof of the last property above as follows:

$$\operatorname{cov}[y] = \mathbb{E}\left[(y-\mu_y)(y-\mu_y)^T\right], \mu_y = \mathbb{E}[Bx] = B\mu_x \tag{13}$$

$$\implies \operatorname{cov}[y] = E\left[(Bx - B\mu_x)(Bx-B\mu_x)^T\right] \tag{14}$$

$$= B\mathbb{E}\left[(x-\mu_x)(x-\mu_x)^T\right]B^T = B\operatorname{cov}[x]\,B^T \tag{15}$$

Let $z \in \mathbb{R}^{n\times 2}$ be the matrix $z = \begin{bmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{bmatrix}^T$ and $z \sim \mathcal{N}(0, \mathbb{I})$. Then $\text{cov}\left[\Omega^{\frac{1}{2}} z\right] = \Omega^{\frac{1}{2}} \text{cov}[z] \Omega^{\frac{1}{2}} = \Omega$ where $\Omega^{\frac{1}{2}}$ us symmetric because $\Omega = P\Lambda P^T$. Further, $\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ and $z\Omega^{\frac{1}{2}} + \mathbb{1}^{n\times 1}\mu^T \sim \mathcal{N}(\mu, \Omega)$.

Definition Cholesky matrix decomposition states that a positive definite matrix A may be written as $A = T'T$, where T is an upper triangular matrix.

Example

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix}^T \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix} = \begin{bmatrix} t_{11}^2 & t_{11}t_{12} \\ t_{11}t_{12} & t_{12}^2 t_{22}^2 \end{bmatrix} \tag{16}$$

$$\implies t_{11} = \sqrt{a_{11}}, t_{12} = \frac{a_{12}}{\sqrt{a_{11}}}, t_{22} = \sqrt{a_{22} - \frac{a_{12}^2}{a_{11}}} \tag{17}$$

Remark Let $z = T'u$ where $u \sim \mathcal{N}(0, \mathbb{I})$. Then $\text{cov}[z] = T'\text{cov}[u]\, T = T'T = \Omega$.

2.1 Quadratic Forms

Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ such that $A' = A$. Then define the quadratic form operator:

$$\mathcal{Q}(x) = x^T A x \tag{18}$$

We are guaranteed that $\mathcal{Q}(x)$ is positive definite (occasionally written $\mathcal{Q}(x) > 0$) for all $x \neq \vec{0}$. The maximum eigenvalue, denoted $\lambda_{\max} = \max_{||x||=1} x^T A x$ and similarly for the minimum eigenvalue $\lambda_{\min} = \min_{||x||=1} x^T A x$.

We can define an operator $\mathcal{L}(x;\lambda) = \mathcal{Q}(x) - \lambda\left((x,x)^2 - 1\right)$. Notice that this is precisely the challenge of finding the maximum eigenvalue set in the form of a optimization problem using Lagrange multipliers. And using a result from matrix calculus that $\frac{\partial\left(x^T A x\right)}{\partial x} = 2Ax$ we find that $\frac{\partial \mathcal{L}}{\partial x} = 2Ax - 2\lambda x$. Using the technique from calculus of setting the derivative equal to zero to perform optimization, it is immediate to obtain $Ax = \lambda x$, which is the classical eigenvalue problem, and it is apparent that the λ yielding the largest value is indeed the maximum eigenvalue.

Suppose we have estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ such that $\mathbb{E}\left[\hat{\theta}_1\right] = \mathbb{E}\left[\hat{\theta}_2\right] = \theta$.

Definition For matrices A and B, where $A, B \in \mathbb{R}^{n\times n}$, we say that $A \leq B$ if and only if the matrix $B - A$ is non-negative definite.

Suppose we have a matrix $\mathcal{X} \in \mathbb{R}^{n\times m}$ where the i^{th} row is an observation and the j^{th} column is a feature. Then we may seek a vector a where each value is a measurement of the relevance of each feature. Define $y = \mathcal{X}a$. We wish to maximize the variance of y, yet this problem is not well-defined since allowing $a \to \infty$ will produce infinite variance. Therefore, we can constrain $||a|| = 1$. Indeed:

$$\text{var}[y] = \sum_{i=1}^{n} (y_i - \bar{y})^2 = ||y - \mathbb{1}^{n\times 1}\bar{y}||^2 \tag{19}$$

$$\bar{y} = \frac{1}{n}\mathbb{1}^{1\times n}\mathcal{X}a \implies y - \mathbb{1}^{n\times 1}\bar{y} = \left(\mathbb{I} - \frac{1}{n}\mathbb{1}\mathbb{1}^T\right)\mathcal{X}a \tag{20}$$

Definition We say that a matrix M is idempotent if $M^2 = MM = M$. For the purposes of statistics, it is often the case that $M = M'$ so that M is symmetric as well.

Remark

$$\left(\mathbb{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)\left(\mathbb{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)^T=\left(\mathbb{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right) \tag{21}$$

So this matrix is idempotent. Since $\text{var}[y] = ||\left(\mathbb{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)\mathcal{X}a||^2 = ||za||^2 = a^Tz^Tza$ we still wish to maximize the variance subject to the constraint that $||a||^2 = 1$. Remember that $\lambda_{\max} = \max_{||x||=1} x^TAx \implies a$ is the maximum eigenvector. This process begins to approach the technique known as principle component analysis.

3 Linear Modeling

We have observations $\mathcal{X}$ for targets $y = \{y_1, \ldots, y_n\}^T \in \mathbb{R}^n$. Suppose we believe that there exists a model for the data of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1}x_{i(p-1)} + \epsilon_i \tag{22}$$

Such that there are p coefficients in the model. Assume that $\mathbb{E}[\epsilon_i] = 0$ and that $\text{var}[\epsilon_i] = \sigma^2$. The matrix $\mathcal{X}$ is a matrix of fixed numbers and ϵ_i and $epsilon_j$ are uncorrelated. We also assume that $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. In expectation then $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1}x_{i(p-1)}$. In matrix notation this becomes:

$$y^{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathcal{X}^{n\times p} = \begin{bmatrix} \mathbb{1}^{n\times 1} & x_{ij}\forall i \in \{1,\ldots,n\}; j \in \{2,\ldots,p\} \end{bmatrix}, \beta^{p\times 1} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \tag{23}$$

$$\implies y = \mathcal{X}\beta + \epsilon, \mathbb{E}[\epsilon] = \vec{0}, \text{var}[\epsilon] = \sigma^2\mathbb{I} \tag{24}$$

Then to discover an optimal choice of β we usually consider the objective function $||y - \mathcal{X}\beta||^2$ which is the residual sum of squares. It goes without saying that we seek to minimize this objective function. Deferring to Newton:

$$\frac{\partial||y-\mathcal{X}\beta||^2}{\partial\beta} = 2\mathcal{X}^T(y - \mathcal{X}\beta) = \vec{0} \tag{25}$$

$$\implies \mathcal{X}^Ty = \mathcal{X}^T\mathcal{X}\beta \implies \beta = \left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^Ty \tag{26}$$

To confirm that this estimator is "preferable" to a different (though still plausible) estimator, consider the following example.

Example Let $y_i = \beta x_i + \epsilon_i$ where one estimator $\tilde{\beta} = \frac{y_{\max}}{x_{\max}}$. It is easily confirmed that $\mathbb{E}\left[\tilde{\beta}\right] = \frac{\beta x_{\max}}{x_{\max}} = \beta$ and that the least-squares estimator $\hat{\beta} = \frac{\sum_{i=1}^n x_iy_i}{\sum_{i=1}^n x_i}$. Then the variance of $\hat{\beta}$ is $\text{var}\left[\hat{\beta}\right] = \frac{\sigma^2}{\sum x_i^2}$ and $\text{var}\left[\tilde{\beta}\right] = \frac{\sigma^2}{x_{\max}^2}$. So clearly the least-squares estimator has smaller variance.

Example Let $\mathcal{X} \in \mathbb{R}^{n\times 2}$ then $\mathcal{X}^T\mathcal{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$. Then the inverse of $\mathcal{X}^T\mathcal{X}$ is $\left(\mathcal{X}^T\mathcal{X}\right)^{-1} = \frac{1}{\Delta}\begin{bmatrix} n & -\sum x_i \\ -\sum x_i & \sum x_i^2 \end{bmatrix}$, where Δ is the determinant of $\mathcal{X}^T\mathcal{X}$.

Proposition 3.1 *The expectation of $\hat{\beta}$ is $\mathbb{E}\left[\hat{\beta}\right] = \beta$ and the covariance matrix of $\hat{\beta}$ is $cov\left[\hat{\beta}\right] = \sigma^2\left(\mathcal{X}^T\mathcal{X}\right)^{-1}$.*

Proof

$$\hat{\beta} = \left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T y = \left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T\left(\mathcal{X}\beta+\epsilon\right) = \beta + \left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T\epsilon \implies \mathbb{E}\left[\hat{\beta}\right] = \beta \tag{27}$$

$$\text{cov}\left[\hat{\beta}\right] = \mathbb{E}\left[\left(\hat{\beta}-\beta\right)^T\left(\hat{\beta}-\beta\right)\right] = \mathbb{E}\left[\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T\epsilon\epsilon^T\mathcal{X}\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\right] \tag{28}$$

$$= \sigma^2\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\left(\mathcal{X}^T\mathcal{X}\right)\left(\mathcal{X}^T\mathcal{X}\right)^{-1} = \sigma^2\left(\mathcal{X}^T\mathcal{X}\right)^{-1} \tag{29}$$

Theorem 3.2 ***(The Gauss-Markov Theorem)*** *Ordinary least-squares is the best linear unbiased estimator.*

Proof Suppose there is an alternative estimator $\tilde{\beta}$ which is linear in the sense that for some choice of L, $\tilde{\beta} = L^{m\times n}y^{n\times 1}$. Let it also be the case that $\mathbb{E}\left[\tilde{\beta}\right] = \beta$. Then $\mathbb{E}\left[\tilde{\beta}\right] = \mathbb{E}\left[L\mathcal{X}\beta + L\epsilon\right] = L\mathcal{X}\beta = \beta \implies L\mathcal{X} = \mathbb{I}$. We must have that $\text{cov}\left[\tilde{\beta}\right] = \sigma^2 LL^T$. Then $\forall\ L$ where $L\mathcal{X} = \mathbb{I}$ we have $LL^T - \left(\mathcal{X}^T\mathcal{X}\right)^{-1} \geq 0$. The minimum bound is only achieved when $L = \left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T$.

3.1 Confidence Intervals for Linear Modeling

Suppose that the model is $y = \mathcal{X}\beta + \epsilon$. Consider a specific $x_*^{p\times 1}$. Then $\hat{y}_* = \hat{\beta}^T x_*$. The variance $\text{var}\left[\hat{y}_*\right] = x_*^T\text{cov}\left[\hat{\beta}\right]x_* = \hat{sigma}^2 x_*^T\left(\mathcal{X}^T\mathcal{X}\right)^{-1}x_*$. Let $s_* = \sqrt{\text{var}\left[\hat{y}_*\right]}$. Then the $1-\alpha$ confidence interval is given as $\hat{\beta}^T x_* \pm t_{1-\frac{\alpha}{2}}\left(s_*\right)$.

4 Random Variable Theory

Let $x^{n\times 1} = \{x_1, \ldots, x_n\}^T$ where each of the $x_i \sim \mathcal{N}(0,1)$.

Theorem 4.1 *If $a \in \mathbb{R}^n$ such that $||a|| = 1$ then $x^T a \sim \mathcal{N}(0,1)$. In addition, $var\left[x^T a\right] = a^T cov[x]\, a = a^T a = 1$.*

Theorem 4.2 ***(χ^2-Distribution)*** *For a matrix $A \in \mathbb{R}^{n\times n}$ let it be symmetric and idempotent. Then $x^T Ax \sim \chi^2\left(tr[A]\right)$*

Example Let $y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Then we have:

$$\frac{1}{\sigma^2}\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 = \sum_{i=1}^n\left(\frac{y_i}{\sigma}-\frac{\bar{y}}{\sigma}\right)^2 = \sum_{i=1}^n\left(z_i-\bar{z}\right)^2, z_i \sim \mathcal{N}(0,1) \tag{30}$$

$$\implies \frac{1}{\sigma^2}\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 = z^T\left(\mathbb{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)z, z \sim \mathcal{N}\left(\vec{0}, \mathbb{I}\right) \tag{31}$$

$$\implies \text{tr}\left[\left(\mathbb{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)\right] = n - \text{tr}\left[\frac{1}{n}\mathbb{1}\mathbb{1}^T\right] = n - \frac{1}{n}n = n-1 \tag{32}$$

Thus, $\frac{1}{\sigma^2}\sum_{i=1}^n\left(y_i-\bar{y}\right)^2 \sim \chi^2(n-1)$.

Theorem 4.3 ***(t-Distribution)*** *Let $a \in \mathbb{R}^n$ be normalized such that $||a|| = 1$. As above, also permit A to be a symmetric and idempotent matrix with the added condition that $Aa = \vec{0}$. Then:*

$$\frac{a^T x}{\sqrt{\frac{a^T Aa}{tr[A]}}} \sim t\left(tr[A]\right) \tag{33}$$

Theorem 4.4 ***(F-Distribution)*** *Suppose now that there are two matrices A and B where both are symmetric and idempotent and are mutually orthogonal. That is, $AB = \vec{0}^{n\times n}$. Then:*

$$\frac{x^T Ax/tr[A]}{x^T Bx/tr[B]} \sim F\left(tr[A], tr[B]\right) \tag{34}$$

Example

4.1 Partial Correlation

Let $\mathcal{X} = \{x_1, \ldots, x_n\} \in \mathbb{R}^n\}$. It is easy to see that $\mathbb{E}[\mathcal{X}] = \mu = \{\mu_1, \ldots, \mu_n\}^T$ and $\operatorname{cov}[\mathcal{X}] = \Omega^{n\times n} = \mathbb{E}\left[(\mathcal{X}-\mu)^T(\mathcal{X}-\mu)\right]$. Assuming then that the x_i are normally distributed then, we may write that $\mathcal{X} \sim \mathcal{N}(\mu, \Omega)$. Indeed, it is further the case that $\forall\ a \in \mathbb{R}^n$, $a^T\mathcal{X} \sim \mathcal{N}\left(a^T\mu, a^T\Omega a\right)$.

If $\det\Omega = 0$, the there exists an a for which $a^T\Omega a = 0$, but we are guaranteed that Ω is non-negative definite so we have that $a^T\Omega a \geq 0$. Let $D = \operatorname{diag}\left[\{\sigma_1^2, \ldots, \sigma_n^2\}\right]$. Then define $R = D^{\frac{-1}{2}}\Omega D^{\frac{-1}{2}}$ and $\Omega = D^{\frac{1}{2}}RD^{\frac{1}{2}}$.

Definition The multivariate Gaussian p.d.f. is written as follows:

$$f_x(x;\mu,\Omega) = (2\pi)^{\frac{-n}{2}}[\det\Omega]^{\frac{-1}{2}}\exp\left[\frac{-1}{2}(x-\mu)^T\Omega^{-1}(x-\mu)\right] \tag{35}$$

The reader is encouraged to show that this p.d.f. decomposes into the familiar formula for the case $n = 1$.

Allow that $\mathcal{X} \sim \mathcal{N}(\mu, \Omega)$ for a matrix A this becomes $A\mathcal{X} \sim \mathcal{N}\left(A\mu, A\Omega A^T\right)$. Suppose instead that we have $\mathcal{X} \sim \mathcal{N}(\mu_x, \Omega_x)$ and $\mathcal{Y} \sim \mathcal{N}(\mu_y, \Omega_y)$. If $\mathcal{X}$ and $\mathcal{Y}$ are uncorrelated, then $\begin{bmatrix}\mathcal{X}\\ \mathcal{Y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}, \begin{bmatrix}\Omega_x & \vec{0}\\ \vec{0} & \Omega_y\end{bmatrix}\right)$ and $\mathcal{X}+\mathcal{Y} \sim \mathcal{N}(\mu_x+\mu_y, \Omega_x+\Omega_y)$.

Definition The marginal distribution is defined as the following integral:

$$\int_{-\infty}^{\infty} f(x,y)\,dy = f_x(x) \tag{36}$$

From this, we have the result that $\mathcal{Y}|\mathcal{X} = x \sim \mathcal{N}\left(\mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x-\mu_x), \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}\right)$. Given the situation that:

$$\begin{bmatrix}y\\ \mathcal{X}^{m\times 1}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_y\\ \mu_x^{m\times 1}\end{bmatrix}, \begin{bmatrix}\sigma_y^2 & \Omega_{yx}^T\\ \Omega_{yx} & \Omega_x\end{bmatrix}\right) \tag{37}$$

$$\implies y|\mathcal{X} = x \sim \mathcal{N}\left(\mu_y + \Omega_{xy}^T\Omega_x^{-1}(x-\mu_x), \sigma_y^2 - \Omega_{yx}^T\Omega_x^{-1}\Omega_{yx}\right) \tag{38}$$

Or, even more generally in the case that y is a vector:

$$y|\mathcal{X} = x \sim \mathcal{N}\left(\mu_y + \left[\Omega_{yx}^T\Omega_x^{-1}\Omega_{yx}\right][x-\mu_x], \Omega_y - \Omega_{yx}^T\Omega_x^{-1}\Omega_{yx}\right) \tag{39}$$

5 Maximum Likelihood

Let $y_i \overset{i.i.d.}{\sim} f(y;\theta)$. Then the likelihood function (and the log-likelihood function) is:

$$\mathcal{L}(\theta) = \prod_{i=1}^{k} f(y_i;\theta) \implies l(\theta) = \sum_{i=1}^{k}\log f(y_i;\theta) \tag{40}$$

In many cases, we seek to maximize the likelihood of the data by finding the choice for θ that maximizes the log-likelihood function. Because log is a strictly increasing function, notice that this optimal choice for θ will also optimize the vanilla likelihood.

Definition

$$\hat{\theta}_{\text{maximum likelihood}} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(\theta) \equiv \underset{\theta}{\operatorname{argmax}}\ l(\theta) \tag{41}$$

Theorem 5.1 *In the classical problem of linear modeling, we have* $\hat{\beta}_{\textit{least-squares}} = \hat{\beta}_{\textit{maximum likelihood}}$ *when the model assumes* $y \sim \mathcal{N}\left(\mathcal{X}\beta, \sigma^2\mathbb{I}\right)$.

Proof The proof of this is perhaps more straightforward than one might imagine. Consider that $f\left(y;\beta,\sigma^2\right) = (2\pi)^{\frac{-n}{2}}\left(\sigma^2\right)^{\frac{-n}{2}}\exp\left[\frac{-1}{2\sigma^2}||y-\mathcal{X}\beta||^2\right]$. Then the log-likelihood function is expressed as:

$$l\left(\beta,\sigma^2\right) = \frac{-1}{2}\left(n\log 2\pi + n\log\sigma^2 + \frac{1}{\sigma^2}||y-\mathcal{X}\beta||^2\right) \tag{42}$$

Notice that the only term in the log-likelihood involving β is $\frac{1}{\sigma^2}||y-\mathcal{X}\beta||^2$, and the fractional term in the front is essentially a constant scale, which means it may be disregarded. It is immediate to see that maximizing log-likelihood with respect to β requires the minimization of $||y-\mathcal{X}\beta||$. In particular, $\hat{\beta}_{\text{maximum likelihood}} = \underset{\beta}{\text{argmax}}\ \ ||y-\mathcal{X}\beta||$. This is the same minimization problem that was solved in the case of least-squares. Therefore, $\hat{\beta}_{\text{maximum likelihood}} = \hat{\beta}_{\text{least-squares}}$.

Remark It is also easy to obtain the maximum likelihood estimate for the variance σ^2. Differentiating:

$$\frac{\partial l}{\partial\sigma^2} = \frac{n}{\sigma^2} - \frac{||y-\mathcal{X}\beta||}{\sigma^2} = 0 \implies \hat{\sigma}^2_{\text{maximum likelihood}} = \frac{||y-\mathcal{X}\beta||^2}{n} \tag{43}$$

Example Suppose we have $f\left(y;\mu,\lambda\right) = \frac{1}{2\lambda}\exp\left[-\frac{-\mu|}{\lambda}\right]$. Then the log-likelihood function is $l\left(\mu,\lambda\right) = -n\log 2 - n\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n}|y_i-\mu|$. To obtain the maximum likelihood estimate of μ we seek to minimize $\sum_{i=1}^{n}|y_i-\mu|$. With some thought, it can be seen that this quantity may be minimized by setting $\hat{\mu}_{\text{maximum likelihood}} = \text{median}$.

5.1 Properties of (Asymptotic) Maximum Likelihood

Consistency: The maximum likelihood estimator is consistent. With large numbers of observations, the maximum likelihood estimator converges in probability to the true value. Mathematically, we have $\hat{\theta}_{\text{maximum likelihood}} \overset{\text{prob.}}{\to} \theta$

Identification: The maximum likelihood must be identifiable in order to be consistent, but we list this property explicitly in any case. The maximum likelihood estimator is identifiable in the sense that $f\left(y;\hat{\theta}_1\right) = f\left(y;\hat{\theta}_2\right) \iff \hat{\theta}_1 = \hat{\theta}_2$.

Normality: The maximum likelihood estimator has an asymptotically normal distribution. In particular, $\sqrt{n}\left(\hat{\theta}_{\text{maximum likelihood}} - \theta\right) \to \mathcal{N}\left(0, \mathcal{I}^{-1}\left(\theta\right)\right)$, where $\mathcal{I}\left(\theta\right)$ is the Fisher Information matrix (more on that later).

Efficiency: If there exists a minimum-variance unbiased estimator of θ, it will be produced by the maximum likelihood method. The mathematics behind this is expressed as follows: $\forall\ \tilde{\theta}_n \to \theta$ and $\sqrt{n}\left(\hat{\theta}_{\text{maximum likelihood}} - \theta\right) \to \mathcal{N}\left(0, \mathcal{A}\left(\theta\right)\right) \implies \mathcal{A}\left(\theta\right) - \mathcal{I}\left(\theta\right) \geq 0$.

6 Fisher Information

Definition The Fisher Information matrix is defined to be (for $\theta \in \mathbb{R}^m$):

$$\mathcal{I}^{m\times m}\left(\theta\right) = \mathbb{E}\left[\left(\frac{\partial l\left(\theta\right)}{\partial\theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 l\left(\theta\right)}{\partial\theta^2}\right] = \text{cov}\left[\frac{\partial l\left(\theta\right)}{\partial\theta}\right] \tag{44}$$

Lemma 6.1

$$\mathbb{E}\left[\frac{\partial l\left(\theta\right)}{\partial\theta}\right] = 0 \tag{45}$$

Proof Recall that $l(\theta) = \log f(y;\theta)$

$$\mathbb{E}\left[\frac{\partial l(\theta)}{\partial \theta}\right] = \mathbb{E}\left[\frac{\frac{\partial f(y;\theta)}{\partial \theta}}{f(y;\theta)}\right] = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(y;\theta)\, dy = \frac{\partial}{\partial \theta} 1 = 0 \tag{46}$$

Lemma 6.2

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} l(\theta) \overset{\sim}{\to} \mathcal{N}(0, \mathcal{I}(\theta)) \tag{47}$$

Example Let $y \sim \mathcal{N}\left(\mathcal{X}\beta, \sigma^2 \mathbb{I}\right)$. Then the multivariate p.d.f. is $f\left(y;\beta,\sigma^2\right) = (2\pi)^{\frac{-n}{2}} \left(\sigma^2\right)^{\frac{-n}{2}}$ $\exp\left[\frac{1}{2\sigma^2}||y - \mathcal{X}\beta||^2\right]$. Then the log-likelihood was previously derived to be:

$$l\left(\beta,\sigma^2\right) = \frac{-n}{2}(2\pi) - \frac{n}{2}\sigma^2 - \frac{1}{2\sigma^2}||y - \mathcal{X}\beta||^2 \implies \frac{\partial l\left(\beta,\sigma^2\right)}{\partial \beta} = \frac{1}{\sigma^2}\mathcal{X}^T\left[y - \mathcal{X}\beta\right] \tag{48}$$

$$\mathcal{I}(\beta) = \mathbb{E}\left[\frac{1}{\sigma^4}\mathcal{X}^T \epsilon \epsilon^T \mathcal{X}\right] = \frac{1}{\sigma^2}\mathcal{X}^T\mathcal{X} \implies \frac{\partial^2 l\left(\beta,\sigma^2\right)}{\partial \beta^2} = \frac{1}{\sigma^2}\mathcal{X}^T\mathcal{X} \tag{49}$$

Remark (The Cramer-Rao Bound) Let $\hat{\theta}$ be an unbiased estimator of θ. Then it must be the case that $\text{cov}\left[\hat{\theta}\right] \geq \mathcal{I}^{-1}(\theta)$. Least-squares is best according to the Cramer-Rao bound when the assumption of normality is true. In particular, $\text{cov}\left[\hat{\beta}\right] = \sigma^2\left(\mathcal{X}^T\mathcal{X}\right)^{-1} = \mathcal{I}^{-1}(\beta)$ and has achieved the lower bound and cannot be surpassed.

6.1 Iterative Scoring Algorithms

Suppose we have a mapping function $f : \mathbb{R}^m \to \mathbb{R}^m$ and $y^{m\times 1} = f\left(x^{m\times 1}\right)$. Then $f(x) \approx f(x_0) + J(x_0)(x - x_0)$. The value $J(x_0)$ is the matrix:

$$J(x_0) = \left[\begin{array}{ccc} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_m} \end{array}\right]^{m\times m}\Bigg|_{x=x_0} \tag{50}$$

$$\implies x_k = x_{k-1} - J^{-1}(x_{k-1})\, f(x_{k-1}) \tag{51}$$

In the limit as $k \to \infty$, $x_k \to x_*$ which solves $f(x_*) = \vec{0}$. For maximum likelihood, two approaches to iterative optimization have emerged. One is the Newton-Raphson algorithm, the other is the Fisher scoring algorithm. The two are fairly identical, with a minor difference between them, and either will suffice for convergence to the maximum likelihood estimator. The algorithms are:

$$\theta_k = \theta_{k-1} - \left(\frac{\partial l(\theta)}{\partial \theta}\Big|_{\theta=\theta_{k-1}}\right) \times \begin{cases} \left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\Big|_{\theta=\theta_{k-1}}\right)^{-1} & \text{Newton-Raphson} \\ \left(\mathbb{E}\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right]\right)^{-1} & \text{Fisher Scoring} \end{cases} \tag{52}$$

7 Statistical Tests

7.1 Wald Test

$$\hat{\theta}_{\text{maximum likelihood}} \to \mathcal{N}\left(\theta, \mathcal{I}^{-1}(\theta)\right) \tag{53}$$

$$\implies \hat{\theta}_{\text{maximum likelihood},i} \to \mathcal{N}\left(\theta_i, \mathcal{I}^{-1}_{(i,i)}(\theta)\right) \tag{54}$$

Example Let y_i be binary in the sense that $y_i \in \{0,1\}$. Suppose further that there exists a probability p such that $\mathbb{P}[y_i = 1] = p$. Then suppose further that for a choice of $0 \leq p_0 \leq 1$ we wish to test the hypothesis that $p = p_0$. We write the likelihood function as follows:

$$\mathcal{L}(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \implies l(p) = \sum_{i=1}^{n} y_i \log p + (1-y_i)\log[1-p] \tag{55}$$

$$\implies \frac{\partial l(p)}{\partial p} = \sum_{i=1}^{n}\left[\frac{y_i}{p} - \frac{(1-y_i)}{1-p}\right] = \frac{\sum_{i=1}^{n} y_i}{p} - \frac{\sum_{i=1}^{n}(1-y_i)}{p} = \frac{m}{p} - \frac{(n-m)}{1-p} = 0 \tag{56}$$

$$\implies \hat{p}_{\text{maximum likelihood}} = \frac{m}{n} \tag{57}$$

Where in this case m represents the number of times that y_i was equal to 1 across all i. We continue then to develop the Fisher Information matrix:

$$\mathcal{I}(p) = -\mathbb{E}\left[\frac{\partial^2 l}{\partial p^2}\right] = \text{var}\left[\frac{\partial l}{\partial p}\right] \tag{58}$$

$$\implies \frac{\partial^2 l}{\partial p^2} = \frac{-m}{p^2} - \frac{(n-m)}{(1-p^2)}, \mathbb{E}[m] = np \tag{59}$$

$$\mathbb{E}\left[\frac{\partial^2 l}{\partial p^2}\right] = \frac{-np}{p^2} - \frac{n-np}{1-p^2} = \frac{-n}{p(1-p)} = \mathcal{I}(p) \implies \hat{p} \to \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \tag{60}$$

Or, as an alternative calculation:

$$\text{var}\left[\frac{m}{p} - \frac{n-m}{1-p}\right] = \text{var}\left[\frac{m-pn}{p(1-p)}\right] = \frac{p(1-p)n}{p^2(1-p)^2} = \frac{n}{p(1-p)} \tag{61}$$

Which quickly yields an identical result. Then we may make up numbers for the Wald test to demonstrate the functional form of the statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{35}{365} - \frac{30}{365}}{\sqrt{\frac{\frac{30}{365}\left(1-\frac{30}{265}\right)}{365}}} = .95 \tag{62}$$

Example Suppose the model is $y = \beta_1 x_1 + \ldots + \beta_m x_m + \epsilon$. Then we may wish to test the hypothesis that $\beta_1 = 0$. Recall that $t = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{(\mathcal{X}^T\mathcal{X})^{-1}_{(1,1)}}}$. It is the case that $\hat{\sigma}^2 = \frac{1}{n-m}||y - \mathcal{X}\beta||^2$. As $n \to \infty$ this estimator goes to the maximum likelihood value. Indeed $\hat{\sigma}^2_{\text{maximum likelihood}} = \frac{1}{n}||y - \mathcal{X}\beta||^2$. From this, we have the implication that $\frac{\hat{\beta}_1}{\hat{\sigma}_{\text{maximum likelihood}}\sqrt{(\mathcal{X}^T\mathcal{X})^{-1}_{(1,1)}}} \sim \mathcal{N}(0,1)$.

7.2 Likelihood Ratio Test

Consider the result from Fisher:

$$-2\left(\max_{\theta^*} l\left(\hat{\theta}, \theta^*\right) - \max_{\theta, \theta^*} l(\theta, \theta^*)\right) \sim \chi^2(1) \tag{63}$$

Under the null hypothesis that the parameter $\theta = \hat{\theta}$. When this null hypothesis is true, we expect the value to be small and positive.

8 Statistical Power

We begin with two definitions that should be well-known and recognizable:

Type I Error: The null hypothesis is true, but rejected in any case. This can be expressed as the α level in a statistical test.

Type II Error: The instance when the null hypothesis is not true, but it is accepted in any case.

Example Define $\hat{y} = (y_1, y_2, \ldots, y_n)$ and $\mathcal{S}(\hat{y})$ and standardization function of the input $\hat{y}$. Thus, we obtain:

$$\mathbb{P}\left[|\mathcal{S}(\hat{y}) - \theta_0| > S_0 \middle| \theta_0\right] = \alpha \tag{64}$$

$$\mathbb{P}\left[|\mathcal{S}(\hat{y}) - \theta_0| \leq S_0 \middle| \theta\right] = \beta \tag{65}$$

$$\text{Power}\,(\theta) = 1 - \beta = \mathbb{P}\left[|\mathcal{S}(\hat{y}) - \theta_0| > S_0 \middle| \theta\right] \tag{66}$$

Example Consider that our null hypothesis will be $\mu = \mu_0$ with an alternative $\mu \neq \mu_0$. Then define a standardization function $\mathcal{S}(y) = \frac{y - y_0}{\sigma/\sqrt{n}}$.

$$\text{Power}\,(\mu) = \mathbb{P}\left[|\mathcal{S}(y)| > Z_{1-\alpha} \middle| \mu\right] = \mathbb{P}\left[\mathcal{S}(y) < -Z_{1-\alpha/2}\right] + \mathbb{P}\left[\mathcal{S}(y) > Z_{1-\alpha/2}\right] \tag{67}$$

$$= \Phi\left(-\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - Z_{1-\alpha/2}\right) + 1 - \Phi\left(-\frac{\mu - \mu_0}{\sigma/\sqrt{n}} + Z_{1-\alpha/2}\right) \tag{68}$$

Remark Notice that $\text{Power}(\mu_0) = \alpha$ and that $\underset{\mu \to \infty}{\text{Power}}(\mu) = \underset{\mu \to -\infty}{\text{Power}}(\mu) = 1$

Example Consider $x_i \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), y_i \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$. We may have a null hypothesis that $\mu_1 = \mu_2$ and an alternative of the form $\mu_1 - \mu_2 = \delta \neq 0$. The quantity:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1) \tag{69}$$

Therefore, we have that the power of detection is:

$$\text{Power} = \Phi\left(-Z_{1-\alpha/2} - \frac{\delta}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) + 1 - \Phi\left(Z_{1-\alpha/2} - \frac{\delta}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \tag{70}$$

8.1 Fisher Z-Transformation

Consider ordered pairs of data $(x_i, y_i) \sim \mathcal{N}(\mu, \Omega)$ with $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. Then define:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{71}$$

Then the Fisher Z-Transformation is given as:

$$\sqrt{n-3}\left(\frac{1}{2}\log\left[\frac{1+r}{1-r}\right] - \frac{1}{2}\log\left[\frac{1+\rho}{1-\rho}\right]\right) \sim \mathcal{N}(0, 1) \tag{72}$$

$$\implies \frac{1}{2}\log\left[\frac{1+r}{1-r}\right] \sim \mathcal{N}\left(\frac{1}{2}\log\left[\frac{1+\rho}{1-\rho}\right], \frac{1}{n-3}\right) \tag{73}$$

In other words, the test statistic is $\sqrt{n-3}\frac{1}{2}\left[\frac{1+r}{1-r}\right]$ and we reject the null hypothesis if and only if $|\sqrt{n-3}\frac{1}{2}\log\left[\frac{1+r}{1-r}\right]| > Z_{1-\alpha/2}$.

9 Classification

Consider a random variable x such that $\mathbb{P}\left[x > c|\mu_1, \sigma_1^2\right] = 1 - \Phi\left(\frac{c-\mu_1}{\sigma_1}\right)$. This probability is referred to as the sensitivity of the classification. The specificity is defined as the quantity $\mathbb{P}\left[x < c|\mu_2, \sigma_2^2\right] =$

$\Phi\left(\frac{c-\mu_2}{\sigma_2}\right)$. Fisher's Linear Discriminant rule for classification provides the following rule for class membership:

$$(\mu_1 - \mu_2)\,\Omega^{-1}\left(x - \frac{\mu_1+\mu_2}{2}\right) > 0 \tag{74}$$

If $\delta^2 = (\mu_1-\mu_2)^T\,\Omega^{-1}\,(\mu_1-\mu)$, then the probability of misclassification is $\Phi\left(\frac{-\delta}{2}\right)$.

10 Expectation Maximization

Suppose there are observations that are obtained $x_1,\ldots,x_{n_1}$ and results that occurred but were unobtainable $z_1,\ldots,x_{n_2}$. Then the total number of instances is $n_1+n_2=n$. We seek to answer the question, "How can we predict missing values when we know only the joint and marginal probability densities?" In particular:

$$k\,(z|x,\theta) = \frac{h\,(x,z|\theta)}{g\,(x|\theta)} \tag{75}$$

We may define the observed likelihood as $\mathcal{L}\,(\theta|x) = g\,(x|\theta)$ and the complete likelihood $\mathcal{L}^c\,(\theta|x,z) = h\,(x,z|\theta)$. Thus:

$$\log\mathcal{L}\,(\theta|x) = \int \log\left[\mathcal{L}\,(\theta|x)\right] k\,(z|x,\theta_0)\,dz \tag{76}$$

$$= \int \left(\log\left[h(x,z|\theta)\right] - \log\left[k(z|x,\theta)\right]\right) k(z|x,\theta_0)dz \tag{77}$$

$$= \int \left(\log\left[h(x,z|\theta)\right] k(z|x,\theta_0)dz - \int \log\left[k(z|x,\theta)k(z|x,\theta_0)dz\right.\right. \tag{78}$$

$$= \mathbb{E}\left[\mathcal{L}^c\,(\theta|x,z)\,|x,\theta_0\right] - \mathbb{E}\left[\log\left[k(z|x,\theta)\right]|x,\theta_0\right] \tag{79}$$

Then the calculation $Q(\theta,\theta_0) = \mathbb{E}\left[\mathcal{L}^c\,(\theta|x,z)\,|x,\theta_0\right]$ is referred to as the expectation step. Further, the optimization $\max_\theta\ Q(\theta,\theta_0)$ is referred to as the maximization step. Therefore, this iterative process is referred to as the algorithm of expectation maximization.

11 Sufficient Statistics

We say that an estimator $\hat{\theta}$ is the minimum variance unbiased estimator if and only if $\mathbb{E}\left[\hat{\theta}\right] = \theta$ and, for any alternative unbiased estimator $\tilde{\theta}$ such that $\mathbb{E}\left[\tilde{\theta}\right] = \theta$, we have that $\text{var}\left[\hat{\theta}\right] \leq \text{var}\left[\tilde{\theta}\right]$. For $y = (y_1,\ldots,y_m)$ and $y_i \sim f\,(y|\theta)$, we say that $u(y) = u$ is sufficient for θ if and only if $\frac{f(y|\theta)}{g(u|\theta)} = H(y)$, where g is the probability density function of u.

Example Consider again $y_i \sim \mathcal{N}\,(\mu,\mathbb{I})$. We shall show that $\bar{y}$ is sufficient for μ. Then,

$$f(y|\mu) = (2\pi)^{-n/2}\exp\left[\frac{-1}{2}\sum(y_i-\mu)^2\right] \tag{80}$$

$$g(\bar{y}|\mu) = \frac{1}{\sqrt{2\pi/n}}\exp\left[\frac{-n}{2}(\bar{y}-\mu)^2\right] \tag{81}$$

Where the last equation follows because $\bar{y} \sim \mathcal{N}\left(\mu,\frac{1}{n}\right)$. Thus, we obtain:

$$\frac{f}{g} = C\exp\left[\frac{-1}{2}\sum(y_i-\mu)^2 - n(\bar{y}-\mu)^2\right] = C\exp\left[\frac{-1}{2}\sum(y_i-\bar{y})^2\right] \tag{82}$$

Theorem 11.1 *$u = u(y)$ is sufficient for $\theta \iff f(y|\theta) = p\,(u|\theta)g(y)$.*

Example *Consider* $f(y|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{-n/2}} \exp\left[\frac{-1}{2\sigma^2}\sum(y_i-\mu)^2\right]$. *Thus,* $\mathcal{L}\left(\mu,\theta^2\right) = f(y|\mu,\theta) = p(u|\mu,\sigma^2)g(y) \implies l\left(\mu,\sigma^2\right) = \log p + \log g \implies \hat{\theta}_{maximum\ likelihood} = \hat{\theta}_{maximum\ likelihood}$. *Therefore, maximum likelihood is a sufficient statistic.*

Theorem 11.2 ***(The Rao-Blackwell Theorem)*** *Suppose* $\mathcal{S} = \mathcal{S}(y)$ *is sufficient and that* $\mathcal{T} = \mathcal{T}(y)$ *is an unbiased estimator of* θ. *Then* $\mathcal{H}(s) = \mathbb{E}[\mathcal{T}|\mathcal{S}=s]$ *and if* $\mathcal{H}(s)$ *does not depend on* θ *then* $\mathcal{H}(s)$ *is unbiased and* $var[\mathcal{H}(s)] \leq var[\mathcal{T}]$.