STA302H5

Al Nosedal

Fall 2018

"Simple can be harder than complex: You have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains."

Steve Jobs.

MULTIPLE LINEAR REGRESSION

Linear Algebra (background)

Theorem A

Generalized Theorem of Pythagoras
If **u** and **v** are orthogonal vectors in an inner product space, then

$$||\mathbf{u}+\mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2$$

Definition

Let W be a subspace of an inner product space V. A vector $\mathbf{u}$ in V is said to be **orthogonal to** W if it is orthogonal to every vector in W, and the set of all vectors in V that are orthogonal to W is called the **orthogonal complement of** W.

Note. The orthogonal complement of a subspace W is denoted by $W^{\perp}$

Theorem B

Properties of Orthogonal Complements

If W is a subspace of finite-dimensional inner product space V, then:

1. $W^{\perp}$ is a subspace of V
2. The only vector common to W and $W^{\perp}$ is $\mathbf{0}$.
3. The orthogonal complement of $W^{\perp}$ is W; that is $(W^{\perp})^{\perp} = W$

Theorem C

Geometric Link between Nullspace and Row Space
If A is an $m \times n$ matrix, then:

1. The nullspace of A and the row space of A are orthogonal complements in $\Re^n$ with respect to the Euclidean inner product.
2. The nullspace of A^T and the column space of A are orthogonal complements in $\Re^m$ with respect to the Euclidean inner product.

Theorem D

Projection Theorem
If W is a finite-dimensional subspace of an inner product space V, then every vector $\mathbf{u}$ in V can be expressed in exactly one way as

$$\mathbf{u} = \mathbf{w}_1 + \mathbf{w}_2$$

where $\mathbf{w}_1$ is in W and $\mathbf{w}_2$ is in $W^{\perp}$.

Theorem E

Best Approximation Theorem
If W is a finite-dimensional subspace of an inner product space V, and if $\mathbf{u}$ is a vector in V, then $\text{proj}_W\mathbf{u}$ is the **best approximation** to $\mathbf{u}$ from W in the sense that

$$||\mathbf{u} - \text{proj}_W\mathbf{u}|| < ||\mathbf{u} - \mathbf{w}||$$

for every vector $\mathbf{w}$ in W that is different from $\text{proj}_W\mathbf{u}$.

Least Squares Problem

Given a linear system $X\beta = \mathbf{y}$ of m equations in n unknowns, find a vector $\hat{\beta}$, if possible, that minimizes $||X\beta - \mathbf{y}||$ with respect to the Euclidean inner product on $\Re^m$. Such vector is called a **least squares solution** of $X\beta = \mathbf{y}$.

To solve the least squares problem, let W be the column space of X

$$W = C(X) = \{X\beta \ \text{ where } \beta \ \text{ in } \Re^m\}$$

Thus, for a vector $\hat{\beta}$ to be a least squares solution of $X\beta = \mathbf{y}$, this vector must satisfy

$$X\hat{\beta} = \text{proj}_{C(X)}\mathbf{y}.$$

We could try to find least squares solutions of $X\beta = \mathbf{y}$ by first calculating $\text{proj}_{C(X)}\mathbf{y}$ and then solving for $\hat{\beta}$. But there is a better way.
We know that for any $\mathbf{y}$ in $\Re^n$

$$\mathbf{y} \;\; = \;\; [\text{proj}_{C(X)}\mathbf{y}] + [\mathbf{y} - \text{proj}_{C(X)}\mathbf{y}]$$

where $\text{proj}_{C(X)}\mathbf{y}$ is in $W = C(X)$ and $\mathbf{y} - \text{proj}_{C(X)}\mathbf{y}$ is in $W^{\perp} = C(X)^{\perp}$.

We know that $\mathbf{y} - \text{proj}_{C(X)}\mathbf{y}$ is orthogonal to $C(X)$. Recalling that $C(X)$ represents the column space of X and, by Theorem C 2), $C(X)^{\perp} = N(X^T)$, we have that

$$\begin{aligned} X^T(\mathbf{y} - \text{proj}_{C(X)}\mathbf{y}) &= \mathbf{0} \\ X^T(\mathbf{y} - X\hat{\beta}) &= \mathbf{0} \\ X^T\mathbf{y} - X^TX\hat{\beta} &= \mathbf{0} \\ X^T\mathbf{y} = X^TX\hat{\beta} \ \ (X^TX \text{ invertible}) & \\ (X^TX)^{-1}X^T\mathbf{y} &= \hat{\beta} \end{aligned}$$

Theorem

If X is an $m \times n$ matrix with linearly independent column vectors, then for every $m \times 1$ matrix $\mathbf{y}$, the linear system $X\beta = \mathbf{y}$ has a unique least squares solution. This solution is given by

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

Moreover, if W is the column space of X, then the orthogonal projection of $\mathbf{y}$ on W is

$$\text{proj}_W \mathbf{y} = X\hat{\beta} = X(X^T X)^{-1} X^T \mathbf{y}$$

Example 1. Simple Linear Regression

Consider regression model

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

$E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ 1 & x_3 - \bar{x} \\ \vdots & \\ 1 & x_n - \bar{x} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$X^T X$

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^{n} x_i - n\bar{x} \\ \sum_{i=1}^{n} x_i - n\bar{x} & \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix}$$

$(X^TX)^{-1}$

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix}$$

$X^T Y$

$$X^T Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (x_i - \bar{x}) y_i \end{pmatrix}$$

$\hat{\beta} = (X^TX)^{-1}X^TY$

$$\hat{\beta} = (X^TX)^{-1}X^TY = \begin{pmatrix} \frac{\sum_{i=1}^{n} y_i}{n} \\ \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix}$$

Example 2. Simple Linear Regression (again...)

Consider regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

$E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$X^T X$

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$(X^TX)^{-1}$

$$(X^TX)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i^2\right)} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

$(X^TX)^{-1}$

$$(X^TX)^{-1} = \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

$X^T Y$

$$X^T Y = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

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

$$\hat{\beta} = \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \left[\sum_{i=1}^{n} x_i^2\right]\left[\sum_{i=1}^{n} y_i\right] - \left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} x_i y_i\right] \\ n\sum_{i=1}^{n} x_i y_i - \left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right] \end{pmatrix}$$

$\hat{\beta} = (X^TX)^{-1}X^TY$

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \left[\sum_{i=1}^{n} x_i^2\right] [\bar{y}] - [\bar{x}] \left[\sum_{i=1}^{n} x_i y_i\right] \\ \sum_{i=1}^{n} x_i y_i - n\,[\bar{x}]\,[\bar{y}] \end{pmatrix}$$

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

$$\hat{\beta} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \left[\sum_{i=1}^{n} x_i^2\right][\bar{y}] - n\bar{x}^2\bar{y} + n\bar{x}^2\bar{y} - [\bar{x}]\left[\sum_{i=1}^{n} x_i y_i\right] \\ \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \end{pmatrix}$$

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

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \bar{x} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix}$$

Classical Linear Regression Model

$$\underbrace{Y}_{(n\times 1)} = \underbrace{X}_{n\times(r+1)} \underbrace{\beta}_{(r+1)\times 1} + \underbrace{\epsilon}_{(r+1)\times 1},$$

$E(\epsilon) = \underbrace{0}_{(n\times 1)}$ and $Cov(\epsilon) = \underbrace{\sigma^2 I}_{(n\times n)}$,
where β and σ^2 are unknown parameters and the design matrix X has jth row $[1, x_{j1}, \cdots, x_{jr}]$.

Proposition

Let A and B be matrices of appropriate dimensions.

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

The last equality only holds when A and B both have inverses. The second to the last property implies that the inverse of a symmetric matrix is symmetric.

Result 1

Let X have full rank $r+1 \leq n$. The least squares estimate of β in the Classical Linear Regression Model is given by

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

Let $\hat{y} = X\hat{\beta} = Hy$ denote the fitted values of y, where $H = X(X^T X)^{-1} X^T$ is called "hat" matrix. Then the residuals

$$\hat{\epsilon} = y - \hat{y} = [I - X(X^T X)^{-1} X^T]y = (I - H)y$$

satisfy $X^T \hat{\epsilon} = 0$ and $\hat{y}^T \hat{\epsilon} = 0$. Also, the

$$\begin{aligned} \text{residual sum of squares} &= \hat{\epsilon}^T \hat{\epsilon} \\ &= y^t [I - X(X^T X)^{-1} X^T] y \\ &= y^T y - y^T X \hat{\beta} \end{aligned}$$

Mean Vectors

A random vector is a vector whose elements are random variables. The expected value of a random vector is the vector consisting of the expected values of each of its elements.

Let **y** represent a random vector of p variables measured on a sampling unit (subject or object). The mean of **y** over all possible values in the population is called the population mean vector or expected value of **y**. It is defined as a vector of expected values of each variable,

$$E(\mathbf{y}) = E\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

Covariance Matrices

If $\mathbf{y}$ is a random vector taking on any possible value in a multivariate population, the population covariance matrix is defined as

$$\mathbf{\Sigma} = cov(\mathbf{y}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$

The diagonal elements $\sigma_{jj} = \sigma_j^2$ are the population variances of the y's, and the off-diagonal elements σ_{jk} are the population covariances of all possible pairs of y's.

The population covariance matrix can also be found as

$\mathbf{\Sigma} = E[(\mathbf{y} - E(\mathbf{y}))(\mathbf{y} - E(\mathbf{y}))^{'}]$

Correlation Matrices

The population correlation matrix is defined as

$$\mathbf{P}_{\rho} = (\rho_{jk}) = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix}$$

where $\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$.

Proposition

Let A be a fixed $r \times n$ matrix, let c be a fixed $r \times 1$ vector, and let Y be an $n \times 1$ random vector, then

1. $E(AY + c) = AE(Y) + c$
2. $Cov(AY + c) = ACov(Y)A^T$

Proposition

Let X and Y be random matrices of the same dimension, and let A and B be conformable matrices of constants.

1. $E(X+Y) = E(X) + E(Y)$
2. $E(AXB) = AE(X)B$

Definition

Let $A = \{a_{ij}\}$ be a $k \times k$ square matrix. The **trace** of the matrix A, written $tr(A)$, is the sum of the diagonal elements; that is,

$$tr(A) = \sum_{i=1}^{k} a_{ij}$$

Proposition

Let A and B be $k \times k$ matrices and c be a scalar.

- $tr(cA) = c\ tr(A)$
- $tr(A \pm B) = tr(A) \pm tr(B)$
- $tr(AB) = tr(BA)$
- $tr(B^{-1}AB) = tr(A)$

Proposition

Let A be a $k \times k$ symmetric matrix and x be a $k \times 1$ vector. Then

$$x^T Ax = tr(x^T Ax) = tr(Axx^T)$$

Result 2

Under the general linear regression model, the least squares estimator $\hat{\beta} = (X^TX)^{-1}X^TY$ has $E(\hat{\beta}) = \beta$ and $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$. The residuals $\hat{\epsilon}$ have the properties $E(\hat{\epsilon}) = 0$ and $Cov(\hat{\epsilon}) = \sigma^2[I - H]$. Also, $E(\hat{\epsilon}^T\hat{\epsilon}) = (n - r - 1)\sigma^2$, so defining

$$s^2 = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n - (r+1)} = \frac{Y^T[I - H]Y}{n - r - 1}$$

we have $E(s^2) = \sigma^2$.
Moreover, $\hat{\beta}$ and $\hat{\epsilon}$ are uncorrelated.

$E(\hat{\beta})$

$$\begin{aligned} E(\hat{\beta}) &= E\left((X^TX)^{-1}X^TY\right) \\ &= (X^TX)^{-1}X^TE(Y) \\ &= (X^TX)^{-1}X^TX\beta \\ &= \beta \end{aligned}$$

$Cov(\hat{\beta})$

$$
\begin{aligned}
Cov(\hat{\beta}) &= Cov\left((X^TX)^{-1}X^TY\right) \\
&= \left[(X^TX)^{-1}X^T\right] Cov(Y) \left[(X^TX)^{-1}X^T\right]^T \\
&= \left[(X^TX)^{-1}X^T\right] Cov(Y)X \left[(X^TX)^{-1}\right]^T \\
&= (X^TX)^{-1}X^TCov(Y)X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}
$$

Gauss-Markov theorem

Let $Y = X\beta + \epsilon$, where $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I$, and X has full rank $r+1$. For any c, the estimator

$$c^T\hat{\beta} = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \cdots + c_r\hat{\beta}_r$$

of $c^T\beta$ has the smallest possible variance among all linear estimators of the form

$$a^T Y = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

that are unbiased for $c^T\beta$.
(Note. $\hat{\beta}$ represents the least squares estimator)

Multivariate Normal Distribution.

Univariate Normal (Definition)

A random variable X_1 has the Normal distribution with mean μ_1 and variance σ_1^2, denoted $X_1 \sim N(\mu_1, \sigma_1^2)$ whose density is given by

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} exp\left\{ -\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 \right\}$$

where $-\infty < x_1 < \infty$.

The joint density of two **independent** Normal variates is thus $f(x_1, x_2) = f(x_1)f(x_2)$.

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} exp\left\{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_{2\times 1}$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}_{2\times 1}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}_{2\times 2}$$

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix}_{2\times 2}$$

$|\mathbf{\Sigma}| = det(\mathbf{\Sigma}) = \sigma_1^2\sigma_2^2.$

Note that

$$(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2.$$

The joint density can be written as

$$f(x_1, x_2) = \frac{1}{(2\pi)^{2/2}|\mathbf{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$$

Multivariate Normal Density

It can be shown that a p-dimensional Normal density for the random vector $\mathbf{x}^{'} = (x_1, x_2, ..., x_p)$ has the form

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

We say that $\mathbf{x}$ is distributed as $N_p(\mu, \mathbf{\Sigma})$, or simply $\mathbf{x}$ is $N_p(\mu, \mathbf{\Sigma})$.

Example. Bivariate Normal Density

Let us evaluate the $p = 2$ variate Normal density in terms of the individual parameters $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$, $\sigma_1^2 = Var(X_1)$, $\sigma_2^2 = Var(X_2)$, $cov(X_1, X_2) = \sigma_{12}$, and $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = cor(X_1, X_2)$.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}_{2\times 2}$$

$$|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2(1-\rho^2)$$

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}_{2\times 2}$$

Note that

$$(\mathbf{x} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)$$

$$= \frac{1}{1-\rho^2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right].$$

Hence,

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho)}} \times$$

$$e^{\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}}$$

This last expression is useful for discussing certain properties of the Normal Distribution. For example, if the random variables X_1 and X_2 are uncorrelated, so that $\rho = 0$, the joint density can be written as the product of two univariate Normal densities, that is $f(x_1, x_2) = f(x_1)f(x_2)$.

Definition

A $p - vector$ $\mathbf{X}$ has a p-variate Normal distribution iff $\mathbf{a}^T\mathbf{X}$ has a univariate Normal distribution for all constant p-vectors $\mathbf{a}$.

Proposition

i) Any linear transformation of a Multinormal p-vector is Multinormal.
ii) Any vector of elements from a Multinormal p-vector is Multinormal. In particular, the components are univariate Normal.

Proof

Definition

The moment generating function of a multivariate random variable $\mathbf{X}$ is given by

$$M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t}^T}\mathbf{X}]$$

provided this expectation exists in a rectangle that includes the origin. More precisely, there exists $h_i > 0$, $i = 1, \cdots, p$, so that the expectation exists for all $\mathbf{t}$ such that $-h_i < t_i < h_i$, $i = 1, \cdots, p$.

The following two results, which will not be proven, provide the rules for handling multivariate mgf's.

Result

If moment generating functions for two random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ exist, then the cdf's for $\mathbf{X}_1$ and $\mathbf{X}_2$ are identical iff the mgf's are identical in an open rectangle that includes the origin.

Result

Assume the random vectors $\mathbf{X}_1$, $\mathbf{X}_2$, $\cdots$, $\mathbf{X}_p$ each have mgf's $M_{\mathbf{X}_j}(\mathbf{t}_j)$, $j = 1, \cdots, p$, and that $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T, \cdots, \mathbf{X}_p^T)^T$ has mgf $M_{\mathbf{X}}(\mathbf{t})$, where $\mathbf{t}$ is partitioned similarly. Then $\mathbf{X}_1$, $\mathbf{X}_2$, $\cdots$, $\mathbf{X}_p$ are mutually independent iff

$$M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}_1}(\mathbf{t}_1) \times M_{\mathbf{X}_2}(\mathbf{t}_2) \times \cdots \times M_{\mathbf{X}_p}(\mathbf{t}_p)$$

for all $\mathbf{t}$ in an open rectangle that includes the origin.

Theorem

If X is p-variate Normal with mean μ and covariance matrix Σ, its MGF is

$$M(\mathbf{t}) = exp\left[t^T\mu + \frac{1}{2}t^T\Sigma t\right]$$

Proof

Corollary

The components of $\mathbf{X}$ are independent iff Σ is diagonal.

Proof

Corollary

Let $X \sim N_p(\mu, \Sigma)$, and $Y_1 = a_1 + B_1X$, $Y_2 = a_2 + B_2X$, then Y_1 and Y_2 are independent iff $B_1 \Sigma B_2^T = 0$.

Recall from Linear Algebra that λ is an **eigenvalue** of a matrix A with **eigenvector** $x(\neq 0)$ if

$$Ax = \lambda x$$

(x is normalized if $x^T x = \sum_{i=1} x_i^2 = 1$).

Recall also that if A is a real symmetric matrix, then A can be diagonalized by an orthogonal transformation B, to D, say:

$$B^T AB = D$$

(it can be shown that if λ is an eigenvalue of A, then λ is an eigenvalue of D too).

Then a quadratic form in Normal variables with matrix A is also a quadratic form in Normal variables with matrix D, as

$$x^T A x = x^T B D B^T x = y^T D y, \quad \text{(define } y = B^T x\text{)}$$

Proposition

If P is idempotent, its eigenvalues λ are 0 or 1.

Definition

The **rank** of a matrix A is the dimension of the row space of A.

Proposition

If P is symmetric and idempotent, its trace is its rank.

We will be interested in symmetric projection (so idempotent) matrices P. Because their eigenvalues are 0 or 1, we can diagonalize them by orthogonal transformations to a diagonal matrix of 0s and 1s. So if P has rank r, a quadratic form $x^T Px$ can be reduced to a sum of r squares of standard normal variates. By relabelling variables,

$$x^T Px = y_1^2 + y_2^2 + \cdots + y_r^2$$

Result

Let $Y = X\beta + \epsilon$, where X has full rank $r+1$ and ϵ is distributed as $N_n(0, \sigma^2 I)$. Then,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \text{ is distributed as } N_{r+1}(\beta, \sigma^2 (X^T X)^{-1})$$

and is distributed independently of the residuals $\hat{\epsilon} = Y - X\hat{\beta}$. Further,

$$\hat{\epsilon}^T \hat{\epsilon} \quad \text{is distributed as } \sigma^2 \chi^2_{n-r-1}$$

Let Y_0 denote the value of the response when the predictor variables have values $x_0^T = [1, x_{01}, \cdots, x_{0r}]$. According to the model $Y = X\beta + \epsilon$, where X has full rank $r+1$ and ϵ is distributed as $N_n(0, \sigma^2 I)$, the expected value of Y_0 is

$$E(Y_0|x_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_r x_{0r} = x_0^T \beta$$

Its least squares estimate is $x_0^T \hat{\beta}$.

Result

For the linear regression model $Y = X\beta + \epsilon$, where X has full rank $r + 1$ and ϵ is distributed as $N_n(0, \sigma^2 I)$, $x_0^T \hat{\beta}$ is the unbiased linear estimator of $E(Y_0|x_0)$ with minimum variance, $Var(x_0^T \hat{\beta}) = x_0^T (X^T X)^{-1} x_0 \sigma^2$. A $100(1-\alpha)\%$ confidence interval for $E(Y_0|x_0) = x_0^T \beta$ is provided by

$$x_0^T \hat{\beta} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{x_0^T (X^T X)^{-1} x_0 s^2}$$

where $t_{n-r-1}\left(\frac{\alpha}{2}\right)$ is the upper $100(\alpha/2)$th percentile of a t-distribution with $n-r-1$ d.f.

Prediction of a new observation, such as Y_0 at $x_0^T = [1, x_{01}, \cdots, x_{0r}]$. is more uncertain than estimating the expected value of Y_0. According to our regression model

$$Y_0 = x_0^T \beta + \epsilon_0$$

where ϵ_0 is distributed as $N(0, \sigma^2)$ and is independent of ϵ and, hence, of $\hat{\beta}$ and s^2.

Result

Given the linear regression model $Y = X\beta + \epsilon$, where X has full rank $r+1$ and ϵ is distributed as $N_n(0, \sigma^2 I)$, a new observation Y_0 has the unbiased predictor

$$x_0^T\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_r x_{0r}$$

The variance of the forecast error of $Y_0 - x_0^T\hat{\beta}$ is

$$Var(Y_0 - x_0^T\hat{\beta}) = \sigma^2(1 + x_0^T(X^TX)^{-1}x_0)$$

A $100(1-\alpha)\%$ confidence interval for $E(Y_0|x_0) = x_0^T\beta$ is provided by

$$x_0^T\hat{\beta} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{(1 + x_0^T(X^TX)^{-1}x_0)s^2}$$

where $t_{n-r-1}\left(\frac{\alpha}{2}\right)$ is the upper $100(\alpha/2)$th percentile of a t-distribution with $n-r-1$ d.f.

Examples

Example

Let Y_1 and Y_2 be independent, Normal random variables ($E(Y_1) = \mu_1$, $E(Y_2) = \mu_2$, $V(Y_1) = \sigma_1^2$, and $V(Y_2) = \sigma_2^2$. Let a_1 and a_2 denote known constants. Find the probability distribution of the linear combination $U = a_1 Y_1 + a_2 Y_2$.

Solution

The MGF for a random variable X that has a Normal distribution with parameters μ and σ^2 is $M_X(t) = exp\{\mu t + \sigma^2 t^2/2\}$.

$$M_U(t) = E[e^{Ut}] = E[e^{(a_1 Y_1 + a_2 Y_2)t}] = E[e^{a_1 Y_1 t}]E[e^{a_2 Y_2 t}]$$

$$= E[e^{Y_1(a_1 t)}]E[e^{Y_2(a_2 t)}] = M_{Y_1}(a_1 t)M_{Y_2}(a_2 t)$$

$$= exp\{\mu_1 a_1 t + \sigma_1^2 a_1^2 t^2/2\} exp\{\mu_2 a_2 t + \sigma_2^2 a_2^2 t^2/2\}$$

$$M_U(t) = exp\{(a_1\mu_1 + a_2\mu_2)t + (a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2/2\}$$

Therefore U has a Normal distribution with mean $a_1\mu_1 + a_2\mu_2$ and variance $a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.

Result

If $\mathbf{x}$ is distributed as $N_p(\mu, \mathbf{\Sigma})$, then any linear combination of variables $\mathbf{a}^{'}\mathbf{x} = a_1x_1 + a_2x_2 + ... + a_px_p$ is distributed as $N(\mathbf{a}^{'}\mathbf{x}, \mathbf{a}^{'}\mathbf{\Sigma}\mathbf{a})$.

Example

Consider the linear combination $\mathbf{a}^{'}\mathbf{x}$ of a Multivariate Normal random vector determined by the choice $\mathbf{a}^{'} = (1, 0, 0, ..., 0)$ with

$$E(\mathbf{x}) = (\mu_1, \mu_2, ..., \mu_p)^{'}$$

and

$$V(\mathbf{x}) = \mathbf{\Sigma}_{p \times p}$$

Find the distribution of $\mathbf{a}^{'}\mathbf{x}$.

Solution

$$E(\mathbf{a}^{'}\mathbf{x}) = \mathbf{a}^{'}E(\mathbf{x}) = \mu_1$$

$$V(\mathbf{a}^{'}\mathbf{x}) = \mathbf{a}^{'}V(\mathbf{x})\mathbf{a} = \mathbf{a}^{'}\mathbf{\Sigma}\mathbf{a} = \sigma_1^2$$

More generally, the marginal distribution of any component x_i of $\mathbf{x}$ is $N(\mu_i, \sigma_i^2)$.

Result

If $\mathbf{x}_{p\times 1}$ is distributed as $N_p(\mu_{p\times 1}, \mathbf{\Sigma})$, the q linear combinations

$$\mathbf{A}_{q\times p}\mathbf{x}_{p\times 1} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p \\ \vdots a_{q1}x_1 + a_{q2}x_2 + ... + a_{qp}x_p \end{pmatrix}_{q\times 1}$$

are distributed as $N_q(\mathbf{A}\mu, \mathbf{A\Sigma A}')$. Also, $\mathbf{x}_{p\times 1} + \mathbf{d}_{p\times 1}$ where $\mathbf{d}$ is a vector of constants, is distributed as $N_p(\mu + \mathbf{d}, \mathbf{\Sigma})$.

Example

For $\mathbf{x}$ distributed as $N_3(\mu, \mathbf{\Sigma})$, find the distribution of

$$\begin{pmatrix} x_1 - x_2 \\ x_2 - x_3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Solution

Mean.

$$\mathbf{A}\mu = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \end{pmatrix}$$

Covariance matrix.

$$\mathbf{A}\mathbf{\Sigma}\mathbf{A}^{'} = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \sigma_{12} - \sigma_{13} + \sigma_{23} - \sigma_2^2 \\ \sigma_{12} - \sigma_{13} + \sigma_{23} - \sigma_2^2 & \sigma_2^2 + \sigma_3^2 - 2\sigma_{23} \end{pmatrix}$$

Example

If **x** is distributed as $N_5(\mu, \boldsymbol{\Sigma})$, find the distribution of

$$\begin{pmatrix} x_2 \\ x_4 \end{pmatrix}$$

Solution

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} & \sigma_{45} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_{55} \end{pmatrix}$$

Solution

Let us rearrange $\mathbf{x}$

$$\mathbf{x}^* = \begin{pmatrix} x_2 \\ x_4 \\ x_1 \\ x_3 \\ x_5 \end{pmatrix}$$

Solution

$$\mu^* = \begin{pmatrix} \mu_2 \\ \mu_4 \\ \mu_1 \\ \mu_3 \\ \mu_5 \end{pmatrix}$$

$$\mathbf{\Sigma}^* = \begin{pmatrix} \sigma_{22} & \sigma_{24} & \sigma_{21} & \sigma_{23} & \sigma_{25} \\ \sigma_{24} & \sigma_{44} & \sigma_{41} & \sigma_{43} & \sigma_{45} \\ \sigma_{21} & \sigma_{41} & \sigma_{11} & \sigma_{13} & \sigma_{15} \\ \sigma_{23} & \sigma_{43} & \sigma_{13} & \sigma_{33} & \sigma_{35} \\ \sigma_{25} & \sigma_{45} & \sigma_{15} & \sigma_{35} & \sigma_{55} \end{pmatrix}$$

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

Solution

Mean.

$$E(\mathbf{Ax}^*) = \mathbf{A}E(\mathbf{x}^*) = \begin{pmatrix} \mu_2 \\ \mu_4 \end{pmatrix}$$

Covariance matrix.

$$V(\mathbf{Ax}^*) = \mathbf{A}V(\mathbf{x}^*)\mathbf{A}^{'} = \mathbf{A}\mathbf{\Sigma}^*\mathbf{A}^{'} = \begin{pmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{pmatrix}$$

Result

All subsets of $\mathbf{x}$ are Normally distributed. If we respectively partition $\mathbf{x}$, its mean vector μ, and its covariance matrix $\boldsymbol{\Sigma}$ as

$$\mathbf{x}_{p\times 1} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$$

$$\mu_{p\times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

where $\mathbf{x}_1$ and μ_1 are $q \times 1$ vectors and $\mathbf{x}_2$ and μ_2 are $(p-q) \times 1$ vectors.

Result (cont.)

$$\mathbf{\Sigma}_{p\times p} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

where $\mathbf{\Sigma}_{11}$ is $q \times q$, $\mathbf{\Sigma}_{12}$ is $q \times (p-q)$, $\mathbf{\Sigma}_{21}$ is $(p-q) \times q$, and $\mathbf{\Sigma}_{22}$ is $(p-q) \times (p-q)$
then $\mathbf{x}_1$ is distributed as $N_q(\mu_1, \mathbf{\Sigma}_{11})$.

Result

If $\mathbf{x}_1$ $(q_1 \times 1)$ and $\mathbf{x}_2$ $(q_2 \times 1)$ are independent, then $cov(\mathbf{x}_1, \mathbf{x}_2) = 0$, a $q_1 \times q_2$ matrix of zeros.

Example

Let $\mathbf{x}_{3\times 1}$ be $N_3(\mu, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Are x_1 and x_2 independent? What about (x_1, x_2) and x_3?

Solution

Since x_1 and x_2 have covariance $\sigma_{12} = 1$, they are **not** independent. Partitioning $\mathbf{x}$ and $\boldsymbol{\Sigma}$ we see that $\mathbf{x}_1 = (x_1, x_2)'$ and x_3 have covariance matrix

$$\boldsymbol{\Sigma}_{12} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Therefore, (x_1, x_2) and x_3 are independent.

Result

If

$$\mathbf{x}_{(q_1+q_2)\times 1} = \left(\begin{array}{c} \mathbf{x}_1 \\ \mathbf{x}_2 \end{array} \right)$$

is Normally distributed with mean

$$\mathbf{x}_{(q_1+q_2)\times 1} = \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right)$$

and covariance matrix

$$\left(\begin{array}{cc} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{array} \right)$$

then $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent if and only if $\mathbf{\Sigma}_{12} = 0$.