Linear Models with Matrices

James G. Booth

Cornell University

Fall 2015

1 Matrix Algebra

1.1 Basic Matrix Arithmetic

Let $\mathbf{a}$ and $\mathbf{b}$ denote column vectors of dimension n; that is,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Transpose: $\mathbf{a}' = (a_1, a_2, \ldots, a_n)$

Addition: $\mathbf{a}' + \mathbf{b}' = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)$

Subtraction: $\mathbf{a} - \mathbf{b} = (a_1 - b_1, a_2 - b_2, \ldots, a_n - b_n)'$

Inner Product: $\mathbf{a}'\mathbf{b} =< \mathbf{a}, \mathbf{b} >= \sum_{i=1}^n a_i b_i$

Let $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ denote matrices with dimensions $m \times n$ and $p \times q$ respectively.

Matrix Addition: If $m = p$ and $n = q$ then $\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij})$.

Matrix Subtraction: If $m = p$ and $n = q$ then $\mathbf{A} - \mathbf{B} = (a_{ij} - b_{ij})$.

Matrix Multiplication: If $n = p$ then $\mathbf{AB} = (\sum_{k=1}^n a_{ik} b_{kj})$.

Matrix Transpose: $\mathbf{A}' = (a_{ji})$; i.e. the (i, j)th element of $\mathbf{A}'$ is a_{ji}.

Outer Product of Two Vectors: $\mathbf{ab}' = (a_i b_j)$. Note that $\mathbf{a}$ and $\mathbf{b}$ do not have to have the same dimension!

1.2 Special Matrices

1-vector: $\mathbf{1}_n$, a n-dimensional vector with elements all equal to one. Note that $\mathbf{1}_n'\mathbf{1}_n = n$

J-matrix: $\mathbf{J}$, a matrix with all elements equal to one. The outer-product of two $\mathbf{1}$-vectors; $\mathbf{J}_{n\times m} = \mathbf{1}_n\mathbf{1}_m'$.

$\bar{\mathbf{J}}$-matrix: $\bar{\mathbf{J}}_n = \frac{1}{n}\mathbf{J}$.

Sample Sum and Mean: Let $\mathbf{y}$ denote a data vector of length n. Then

$$\mathbf{1}'\mathbf{y} = \sum_{i=1}^{n} y_i \quad \text{and} \quad \frac{1}{n}\mathbf{1}'\mathbf{y} = \bar{y}\,.$$

Square Matrix: The number of columns is the same as the number of rows.

Trace of a Square Matrix: The *trace* of a square matrix is the sum of its diagonal elements; i.e. $\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

Symmetric Matrix: $\mathbf{A}' = \mathbf{A}$.

Diagonal Matrix: A square matrix with all off-diagonal elements equal to zero; $\mathbf{A} = \text{diag}(\mathbf{a})$.

Identity Matrix: $\mathbf{I}_n = \text{diag}(\mathbf{1}_n)$

Inverse Matrix: $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Note that, if $\mathbf{A} = \text{diag}(a_i)$, and $a_i \neq 0$ for all i, then $\mathbf{A}^{-1} = \text{diag}(a_i^{-1})$.

Orthogonal Matrix: $\mathbf{A}'\mathbf{A} = \mathbf{I}$ or equivalently $\mathbf{A}^{-1} = \mathbf{A}'$. If $\mathbf{a}_i$ denotes that ith column of $\mathbf{A}$. Then $\mathbf{a}_i'\mathbf{a}_i = 1$ and $\mathbf{a}_i'\mathbf{a}_j = 0$ if $i \neq j$.

Idempotent Matrix: $\mathbf{A}^2 = \mathbf{A}\mathbf{A} = \mathbf{A}$. Note that $\mathbf{I}_n$ and $\bar{\mathbf{J}}_n$ are both idempotent.

Centering Matrix: $\mathbf{C} = \mathbf{I} - \bar{\mathbf{J}}$. Pre-multiplication by the matrix $\mathbf{C}$ centers a data vector $\mathbf{y}$.

$$\mathbf{C}\mathbf{y} = (\mathbf{I} - \bar{\mathbf{J}})\mathbf{y} = \mathbf{y} - \mathbf{1}\bar{y}\,.$$

1.3 Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be a real square matrix of dimension n. Then $(d, \mathbf{e})$ is an eigenvalue-vector pair for the a square matrix $\mathbf{A}$ if $\mathbf{e} \neq \mathbf{0}$ and

$$\mathbf{A}\mathbf{e} = d\mathbf{e} \quad \text{or equivalently} \quad (\mathbf{A} - d\mathbf{I})\mathbf{e} = \mathbf{0}\,.$$

Note that the eigenvector can be rescaled so that it has unit length, $\mathbf{e}'\mathbf{e} = 1$

Determinant: The determinant of a square matrix is the product of its eigenvalues, $|\mathbf{A}| = \prod_{i=1}^{n} d_i$.

Determinant of a Product: If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices, then $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$. Note that this implies that the determinant of an orthogonal matrix is either 1 or -1.

Non-singular Matrix: A square matrix has an inverse if and only if $|\mathbf{A}| \neq 0$, in which case it is said to be non-singular.

Characteristic Equation: The eigenvalues of a square matrix $\mathbf{A}$ are the solutions of the characteristic equation

$$f(x) = |\mathbf{A} - x\mathbf{I}| = 0\,.$$

Spectral Decomposition of a Symmetric Matrix: Any real symmetric matrix of dimension n can be decomposed in terms of it eigenvalue-vector pairs as follows

$$\mathbf{A} = \sum_{i=1}^{n} d_i \mathbf{e}_i \mathbf{e}_i' = \mathbf{E}\mathbf{D}\mathbf{E}'\,.$$

where $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_n)$ is a real orthogonal matrix and $\mathbf{D} = \text{diag}(d_i)$ is real diagonal matrix.

Inverse of a Symmetric Matrix: If $d_i \neq 0$ for all i then $\mathbf{A}^{-1} = \mathbf{E}\mathbf{D}^{-1}\mathbf{E}'$. Why?

Polynomial in a Symmetric Matrix: If $f(x) = \sum_{r=0}^{p} c_i x^r$ is a polynomial in x, and $\mathbf{A}$ is a symmetric matrix, then $f(\mathbf{A}) = \mathbf{E}f(\mathbf{D})\mathbf{E}'$, where $f(\mathbf{D}) = \text{diag}(f(d_i))$. In particular, $\mathbf{A}^k = \mathbf{E}\mathbf{D}^k\mathbf{E}'$ for any integer $k \geq 0$.

Eigenvalues of an Idempotent Matrix: If $\mathbf{A}$ is symmetric and idempotent, then its eigenvalues are all either 0 or 1. Why?

1.4 Quadratic Forms

Let $\mathbf{A}$ be a square matrix of dimension n and let $\mathbf{x}$ be an n-dimensional variable.

Quadratic Form: The function

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$

is called a quadratic form in $\mathbf{x}$. Without loss of generality we may assume that $\mathbf{A}$ is symmetric. Why?

Gradient Vector: The vector of first derivatives of a quadratic form is given by

$$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \mathbf{x}'\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}\,,$$

(Compare with the first derivative of the function $f(x) = ax^2$.)

Hessian Matrix: The Hessian (second derivative) matrix of a quadratic form is given by

$$\mathbf{H} = \frac{\partial^2}{\partial \mathbf{x}\partial \mathbf{x}'} f(\mathbf{x}) = 2\mathbf{A}\,.$$

(Compare with the second derivative of the function $f(x) = ax^2$.)

Positive Definite: A square matrix, $\mathbf{A}$, is said to be positive definite if the quadratic from, $f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, is positive for all values of $\mathbf{x}$, and positive semi-definite if $f(\mathbf{x}) \geq 0$. Note that the eigenvalues of a positive definite matrix are all positive. Why?

Sum of Squared Deviations: Recall that $\mathbf{C}^2 = \mathbf{C}$ and $\mathbf{C}' = \mathbf{C}$. So

$$\mathbf{y}'\mathbf{C}\mathbf{y} = \mathbf{y}'\mathbf{C}'\mathbf{C}\mathbf{y} = (\mathbf{y} - \mathbf{1}\bar{y})'(\mathbf{y} - \mathbf{1}\bar{y}) = \sum_{i=1}^{n}(y_i - \bar{y})^2 \,.$$

SYY $= \mathbf{y}'\mathbf{C}\mathbf{y}$ is referred to as the total (corrected) sum of squares.

Decomposition of the Total Sum of Squares: For any data vector $\mathbf{y}$ of dimension n the total (uncorrected) sum of squares is defined as $\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} y_i^2$. Notice that

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} = \mathbf{y}'(\bar{\mathbf{J}} + \mathbf{C})\mathbf{y} = n\bar{y}^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 \,.$$

The matrices $\bar{\mathbf{J}}$ and $\mathbf{C}$ in this decomposition are idempotent and have the property that $\bar{\mathbf{J}}\mathbf{C} = \mathbf{0}$.

1.5 Linear Dependence and Rank

Linearly Independent Vectors: The columns of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ are linearly independent if

$$\mathbf{X}\mathbf{a} = \sum_{i=1}^{k} a_i \mathbf{x}_i \neq \mathbf{0}$$

for all non-zero vectors $\mathbf{a}$. Otherwise the columns of $\mathbf{X}$ are said to be linearly dependent.

Rank of a Matrix: The rank of a matrix is its number of linearly independent columns (or rows). The rank of a diagonal matrix is equal to its number of non-zero diagonal elements.

Column Space of a Matrix: The column space of $\mathbf{X}$ $(n \times p)$ is the set $C(\mathbf{X}) = \{\mathbf{b} | \mathbf{b} = \mathbf{X}\mathbf{a}\,,\ \mathbf{a} \in R^p\}$. $C(\mathbf{X})$ is also referred to as the space spanned by the columns of $\mathbf{X}$.

Rank of a Product: $\mathrm{rk}(\mathbf{XA}) \leq \min\{\mathrm{rk}(\mathbf{X}), \mathrm{rk}(\mathbf{A})\}$ (Why?)

Multiplication by a Nonsingular Matrix: Multiplication by a nonsingular matrix does not affect the rank or row/column space.

Rank of a Symmetric Matrix: The rank of a (symmetric) matrix is equal to its number of non-zero eigenvalues.

Rank of an Idempotent Matrix: If $\mathbf{A}$ is symmetric and idempotent, then its rank is equal to its trace; i.e. rank($\mathbf{A}$)=tr($\mathbf{A}$).

1.6 Partitioned Matrices

Let $\mathbf{A} = (\mathbf{A}_{ij})$ and $\mathbf{B} = (\mathbf{B}_{ij})$ be partitioned matrices. Then, provided the products are conformable to multiplication, $\mathbf{AB} = (\sum_k \mathbf{A}_{ik}\mathbf{B}_{kj})$; i.e. $\mathbf{AB}$ is a partitioned matrix with (i, j)th element equal to $\sum_k \mathbf{A}_{ik}\mathbf{B}_{kj}$. For example, if

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} \\ \mathbf{B}_{21} \end{bmatrix}$$

then

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} \end{bmatrix} .$$

Consider the matrix $\mathbf{X}$ with columns $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Then, it follows that

$$\mathbf{XX}' = [\mathbf{x}_1, \ldots, \mathbf{x}_k] \begin{bmatrix} \mathbf{x}_1' \\ \cdots \\ \mathbf{x}_k' \end{bmatrix} = \sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i'$$

1.7 Kronecker Products

Let $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ be matrices with dimensions $n \times m$ and $r \times s$ respectively. Then their *Kronecker product*, defined as

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B})\,,$$

has dimension $nr \times ms$. Kronecker products have the following properties (assuming conformability of matrix additions and multiplications):

(i) $(\mathbf{A} \otimes \mathbf{B}) + (\mathbf{C} \otimes \mathbf{D}) = (a_{ij}\mathbf{B} + c_{ij}\mathbf{D})$

(ii) $(\mathbf{A} \otimes \mathbf{B}) + (\mathbf{A} \otimes \mathbf{D}) = (a_{ij}\mathbf{B} + a_{ij}\mathbf{D}) = \mathbf{A} \otimes (\mathbf{B} + \mathbf{D})$

(iii) $(\mathbf{A} \otimes \mathbf{B}) + (\mathbf{C} \otimes \mathbf{B}) = ((a_{ij} + c_{ij})\mathbf{B}) = (\mathbf{A} + \mathbf{C}) \otimes \mathbf{B}$

(iv) $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$

(v) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$

(vi) $\mathrm{rank}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{rank}(\mathbf{A}) \times \mathrm{rank}(\mathbf{B})$

(vii) If $\mathbf{A}$ and $\mathbf{B}$ are square then $\mathrm{tr}(\mathbf{A} \otimes \mathbf{B}) = \mathrm{tr}(\mathbf{A}) \times \mathrm{tr}(\mathbf{B})$

Some examples involving $\mathbf{I}$, $\mathbf{J}$ and $\mathbf{C}$ matrices are

$\mathbf{1}_k \otimes \mathbf{1}_n = \mathbf{1}_{kn}$

$\mathbf{I}_k \otimes \mathbf{I}_n = \mathbf{I}_{kn}$

$\mathbf{J}_k \otimes \mathbf{J}_n = \mathbf{J}_{kn}$

$\bar{\mathbf{J}}_k \otimes \bar{\mathbf{J}}_n = \bar{\mathbf{J}}_{kn}$

$\mathbf{C}_{kn} = \mathbf{I}_k \otimes \mathbf{I}_n - \bar{\mathbf{J}}_k \otimes \bar{\mathbf{J}}_n$

$\mathbf{C}_{kn}(\mathbf{I}_k \otimes \mathbf{1}_n) = \mathbf{C}_k \otimes \mathbf{1}_n$

The last identity is derived as follows:

$$\begin{aligned} \mathbf{C}_{kn}(\mathbf{I}_k \otimes \mathbf{1}_n) &= (\mathbf{I}_k \otimes \mathbf{I}_n - \bar{\mathbf{J}}_k \otimes \bar{\mathbf{J}}_n)(\mathbf{I}_k \otimes \mathbf{1}_n) \\ &= (\mathbf{I}_k \otimes \mathbf{1}_n - \bar{\mathbf{J}}_k \otimes \mathbf{1}_n) \\ &= \mathbf{C}_k \otimes \mathbf{1}_n \end{aligned}$$

2 Linear Regression

2.1 Scalar Formulation of the SLR Model

Let (x_i, y_i), $i = 1, \ldots, n$, denote a "random sample" of n (x, y) pairs. The *simple linear regression* (SLR) model relating the variables x and y is

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \ldots, n\,,$$

where e_i denotes the deviation from linearity or *error* for the ith observation. The y-variable is referred to as the *dependent variable* or *response*, and the x-variable is referred to as the *independent variable* or the *predictor* or *explanatory variable.*

Standard assumptions concerning the errors are

(i) they have mean zero;

(ii) they are uncorrelated; and

(iii) their variances do not depend on x.

The last assumption, $\text{var}(e_i|x_i) = \sigma^2$, say, is referred to as *homoscedasticity.*

Least Squares Estimates: Estimates of the *intercept* and *slope* parameters can be obtained by minimizing the sum of squared deviations

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to the *regression coefficients*, β_0 and β_1. These estimates are given by

$$\hat{\beta}_1 = \frac{\mathrm{SXY}}{\mathrm{SXX}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathbf{x}'\mathbf{C}\mathbf{y}}{\mathbf{x}'\mathbf{C}\mathbf{x}}$$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Prediction Equation and Fitted Values: The predicted response for a generic value of x is given by $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. The fitted value corresponding to the ith observation is $\hat{y}_i = \hat{y}(x_i)$.

Residuals: The ith residual is $\hat{e}_i = y_i - \hat{y}_i$.

Sum of Squared Errors: SSE $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. This quantity is also referred to as the *residual sum of squares*.

Mean Squared Error: An *unbiased* estimate of the error (or residual) variance is given by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{\mathrm{SSE}}{n-2} = \mathrm{MSE}\,.$$

The divisor $n - 2$ accounts for the fact that the two regression coefficients have been estimated.

Decomposition of the Total Sum of Squares: The total (uncorrected) sum of squares can be decomposed into variation explained by the intercept, the slope and error:

$$\sum_{i=1}^{n} y_i^2 = n\bar{y}^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

It can be shown that the second term equals $\hat{\beta}_1^2$SXX.

2.2 Matrix Formulation of the SLR Model

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ denote the *response vector*, and $\mathbf{x} = (x_1, \ldots, x_n)'$ the vector of explanatory variables (or *covariate vector*). Then the SLR model can be written in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}\,,$$

where, $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$ is the *model or design matrix*, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is the vector of *regression coefficients*, and $\mathbf{e} = (e_1, \ldots, e_n)'$ is the *error vector*. The standard assumptions concerning the errors can be restated as $\mathrm{E}(\mathbf{y}|\mathbf{x}) = \mathbf{X}\boldsymbol{\beta}$; and $\mathrm{var}(\mathbf{y}|\mathbf{x}) = \sigma^2\mathbf{I}$.

Least squares estimate: The sum of squared deviations can be expressed in matrix form as

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,.$$

Differentiating with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,.$$

Equating to zero reveals the *least squares equations*,

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}\,.$$

Thus the least squares estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,,$$

provided the matrix $\mathbf{X}'\mathbf{X}$ is non-singular. Note that, if $\mathbf{x} = \mathbf{1}c$ for some constant c, then the columns of $\mathbf{X}$ are linearly dependent and $\mathbf{X}'\mathbf{X}$ is singular. Also note that SXX$\equiv 0$ in this case and so $\hat{\beta}_1$ is undefined.

The effect of centering $\mathbf{x}$: The model says that the expected responses are linearly related to x; that is

$$\mathrm{E}(y) = \beta_0 + \beta_1 x = (\beta_0 + \beta_1 \bar{x}) + \beta_1 (x - \bar{x})\,.$$

Hence, regression on $x - \bar{x}$ rather than x simply relocates the intercept parameter.

Centering $\mathbf{x}$ changes the model matrix to $\mathbf{X}_c = [\mathbf{1}, \mathbf{Cx}]$. In this case

$$\mathbf{X}_c'\mathbf{X}_c = \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}'\mathbf{C} \end{bmatrix} [\mathbf{1} \quad \mathbf{Cx}] = \begin{bmatrix} n & 0 \\ 0 & \mathbf{x}'\mathbf{Cx} \end{bmatrix}\,,$$

and

$$\mathbf{X}_c'\mathbf{y} = \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}'\mathbf{C} \end{bmatrix} \mathbf{y} = \begin{bmatrix} n\bar{y} \\ \mathbf{x}'\mathbf{Cy} \end{bmatrix}\,.$$

It follows that the least squares estimate is

$$(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y} = \begin{bmatrix} \bar{y} \\ \mathbf{x}'\mathbf{Cy}/\mathbf{x}'\mathbf{Cx} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 \end{bmatrix}\,.$$

Fitted values and the hat-matrix: The vector of fitted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Hy}\,,$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the *hat-matrix*. Note that $\hat{\mathbf{y}} = \mathbf{1}\hat{\beta}_0 + \mathbf{x}\hat{\beta}_1$ is a linear combination of the columns of $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$. That is, $\hat{\mathbf{y}}$ is in the *column space* of $\mathbf{X}$. The vector of fitted values is the *projection* of $\mathbf{y}$ into the column space of $\mathbf{X}$. In particular, $\mathbf{HX} = \mathbf{X}$ and therefore $\mathbf{HXa} = \mathbf{Xa}$, so $\mathbf{H}$ has no effect on vectors that are already in the column space of $\mathbf{X}$. It follows that $\hat{\mathbf{y}}$ is the same if $\mathbf{x}$ is centered because $\mathbf{X}$ and $\mathbf{X}_c$ have the same column space.

Notice that

$$\mathbf{H} = \mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c' = [\mathbf{1} \quad \mathbf{Cx}] \begin{bmatrix} 1/n & 0 \\ 0 & 1/\mathbf{x}'\mathbf{Cx} \end{bmatrix} \begin{bmatrix} \mathbf{1}' \\ \mathbf{x}'\mathbf{C} \end{bmatrix} = \bar{\mathbf{J}}_n + \frac{\mathbf{Cxx}'\mathbf{C}}{\mathbf{x}'\mathbf{Cx}} \,.$$

Vector of residuals: The residual vector is given by

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \,.$$

Note that the vectors $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are orthogonal since $\hat{\mathbf{e}}'\hat{\mathbf{y}} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{H}\mathbf{y} = 0$ because $\mathbf{H}$ is idempotent. Also, the residual sum of squares is

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} \,.$$

ANOVA decomposition: Note that

$$\mathbf{I} = \bar{\mathbf{J}} + (\mathbf{H} - \bar{\mathbf{J}}) + (\mathbf{I} - \mathbf{H})$$

and that the three matrices on the right side are all symmetric and idempotent. Moreover the product of any pair of these matrices is zero. Hence, the total (uncorrected) sum of squares for the response vector can be decomposed as follows:

Source	SS	DF
mean	$\mathbf{y}'\bar{\mathbf{J}}\mathbf{y}$	1
regression on $\mathbf{x}$	$(\mathbf{x}'\mathbf{Cy})^2/\mathbf{x}'\mathbf{Cx}$	1
residual/error	$\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$	$n-2$
total (uncorrected)	$\mathbf{y}'\mathbf{y}$	n

The degrees of freedom for each component is equal to the rank of the matrix in the associated quadratic form.

2.3 Multiple Linear Regression

Let $x_{i1}, \ldots, x_{ip}$ denote a set of covariates (predictors, explanatory variables) associated with a response variable, y_i, for $i = 1, \ldots, n$. A multiple linear regression (MLR) model relating y to $x_1, \ldots, x_p$ is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i \,,$$

where it is usually assumed that $\mathrm{E}(e_i) = 0$, $\mathrm{cov}(e_i, e_j) = 0$ and $\mathrm{var}(e_i) = \sigma^2$. If we define $\mathbf{X}$ to the the $n \times (p+1)$ matrix with columns, $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, the MLR model can be written in the same matrix form as the SLR model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}$. Similarly, the least squares estimate of the regression parameter vector is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ provided that the matrix $\mathbf{X}'\mathbf{X}$ is non-singular, or equivalently the columns of $\mathbf{X}$ are linearly independent.

ANOVA decomposition: Since $\mathbf{H} = \bar{\mathbf{J}} + \mathbf{HCH}$ the ANOVA decomposition of the total (uncorrected) sum of squares is given by

Source	SS	DF
mean	$\mathbf{y}'\bar{\mathbf{J}}\mathbf{y}$	1
regression	$\mathbf{y}'\mathbf{HCHy}$	p
residual/error	$\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$	$n - p - 1$
total (uncorrected)	$\mathbf{y}'\mathbf{y}$	n

Note that the matrices $\bar{\mathbf{J}}$, $\mathbf{HCH}$ and $\mathbf{I} - \mathbf{H}$ are all symmetric and idempotent, and the product of any pair of them is a zero matrix. Also,

$$\mathrm{rank}(\mathbf{H}) = \mathrm{tr}(\mathbf{H}) = \mathrm{tr}\left[(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\right] = \mathrm{tr}(\mathbf{I}_{p+1}) = p + 1 \,.$$

Also, since $\mathrm{rank}(\bar{\mathbf{J}}) = \mathrm{tr}(\bar{\mathbf{J}}) = 1$, it follows that $\mathrm{rank}(\mathbf{HCH}) = \mathrm{rank}(\mathbf{H} - \bar{\mathbf{J}}) = \mathrm{tr}(\mathbf{H}) - \mathrm{tr}(\bar{\mathbf{J}}) = p$, and similarly $\mathrm{rank}(\mathbf{I} - \mathbf{H}) = n - p - 1$.

Mean Squared Error: As in SLR an *unbiased* estimate of the error (or residual) variance is given by the mean error,

$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n - p - 1} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}}{n - p - 1} \,.$$

The divisor $n-p-1$ accounts for the fact that $p+1$ regression coefficients have to be estimated before estimating the variance. Hence the effective sample size for estimating the error variance is less than n.

R-squared: Define R to be the sample correlation between the responses, $\mathbf{y}$ and the fitted values, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$; that is,

$$R = \frac{\mathbf{y}'\mathbf{C}\hat{\mathbf{y}}}{\sqrt{\mathbf{y}'\mathbf{C}\mathbf{y} \cdot \hat{\mathbf{y}}'\mathbf{C}\hat{\mathbf{y}}}} = \sqrt{\frac{\hat{\mathbf{y}}'\mathbf{C}\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{C}\mathbf{y}}}\,.$$

The second equality follows from the identity, $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$, and the fact that $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ are uncorrelated. Hence

$$R^2 = \frac{\hat{\mathbf{y}}'\mathbf{C}\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{C}\mathbf{y}} = \frac{\mathrm{SSR}}{\mathrm{SYY}}\,,$$

where SSR is the regression sum of squares. Thus R^2 can be interpreted as the proportion of the total (corrected) variation in y that is explained by regression on $x_1, \ldots, x_p$.

2.4 Sequential Decomposition of the Regression Sum of Squares

Recall that $\mathrm{SSR} = \mathbf{y}'(\mathbf{H} - \bar{\mathbf{J}})\mathbf{y}$. Let $\mathbf{X}_j$ be the $n \times (j+1)$ matrix consisting of the first $j+1$ columns of $\mathbf{X}$, and let $\mathbf{H}_j$ be the corresponding hat-matrix; that is,

$$\mathbf{H}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j' \quad j = 0, 1, \ldots, p\,.$$

Then

$$\mathbf{H} - \bar{\mathbf{J}} = \mathbf{H}_p - \mathbf{H}_0 = \sum_{j=1}^{p}(\mathbf{H}_j - \mathbf{H}_{j-1})\,.$$

It follows that

$$\text{SSR} = \sum_{j=1}^{p} \mathbf{y}'(\mathbf{H}_j - \mathbf{H}_{j-1})\mathbf{y} = \sum_{j=1}^{p} \text{SS}(\mathbf{x}_j|\mathbf{X}_{j-1})$$

where $\text{SS}(\mathbf{x}_j|\mathbf{X}_{j-1})$ is the sum of squares explained by regression on $\mathbf{x}_j$ over and above that explained by $\mathbf{x}_0, \dots, \mathbf{x}_{j-1}$, where $\mathbf{x}_0 = \mathbf{1}_n$.

Note that, since $\mathbf{H}_j$ is a projection matrix into the column space of $\mathbf{X}_j$

$$\mathbf{H}_j\mathbf{H}_k = \begin{cases} \mathbf{H}_j & j \leq k \\ \mathbf{H}_k & j > k \end{cases}$$

and that if $j \neq k$

$$(\mathbf{H}_j - \mathbf{H}_{j-1})(\mathbf{H}_k - \mathbf{H}_{k-1}) = \mathbf{0}\,.$$

Also,

$$\begin{aligned} \mathbf{H}_j &= \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j' \\ &= (\mathbf{X}_{j-1}, \mathbf{x}_j)\left[(\mathbf{X}_{j-1}, \mathbf{x}_j)'(\mathbf{X}_{j-1}, \mathbf{x}_j)\right]^{-1}(\mathbf{X}_{j-1}, \mathbf{x}_j)' \\ &= (\mathbf{X}_{j-1}, \mathbf{x}_j)\begin{bmatrix} \mathbf{X}_{j-1}'\mathbf{X}_{j-1} & \mathbf{X}_{j-1}'\mathbf{x}_j \\ \mathbf{x}_j'\mathbf{X}_{j-1} & \mathbf{x}_j'\mathbf{x}_j \end{bmatrix}^{-1}(\mathbf{X}_{j-1}, \mathbf{x}_j)'\,. \end{aligned}$$

Thus, if the predictors are orthogonal ($\mathbf{x}_j'\mathbf{x}_k = 0$ for all $j \neq k$),

$$\begin{aligned} \mathbf{H}_j &= \mathbf{X}_{j-1}(\mathbf{X}_{j-1}'\mathbf{X}_{j-1})^{-1}\mathbf{X}_{j-1}' + \mathbf{x}_j(\mathbf{x}_j'\mathbf{x}_j)^{-1}\mathbf{x}_j' \\ &= \mathbf{H}_{j-1} + \frac{\mathbf{x}_j\mathbf{x}_j'}{\mathbf{x}_j'\mathbf{x}_j}\,. \end{aligned}$$

In this case the sequential decomposition of SSR is independent of the ordering of the predictors.

3 Properties of Least Squares Estimates

3.1 Mean and Variance

For a random vector $\mathbf{y}$ $(n\times 1)$ and constant matrices $\mathbf{A}$ $(k\times n)$ and $\mathbf{B}$ $(m\times n)$, and constant vector $\mathbf{b}$ $(k \times 1)$

(i) $\mathrm{E}(\mathbf{Ay}+\mathbf{b}) = \mathbf{A}\mathrm{E}(\mathbf{y})+\mathbf{b}$;

(ii) $\mathrm{var}(\mathbf{Ay}+\mathbf{b}) = \mathbf{A}\mathrm{var}(\mathbf{y})\mathbf{A}'$; and

(iii) $\mathrm{cov}(\mathbf{Ay},\mathbf{By}) = \mathbf{A}\mathrm{var}(\mathbf{y})\mathbf{B}'$.

Under the assumptions of the MLR model, $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}$.

Regression coefficients: Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}\,,$$

which implies that the least squares estimate is *unbiased* in repeated sampling. Also,

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\,,$$

and so the variance of the jth estimated regression coefficient $\hat{\beta}_j$ is given by

$$\sigma^2_{\hat{\beta}_j} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}\,.$$

(In practice σ^2 is estimated by the *mean squared error* from the model fit.)

Recall that in SLR,

$$\mathbf{X}_c'\mathbf{X}_c = \begin{bmatrix} n & 0 \\ 0 & SXX \end{bmatrix}\,,$$

which implies that $\mathrm{var}(\hat{\beta}_1) = \sigma^2/\mathrm{SXX}$.

More generally, for an MLR it can be shown that

$$\sigma^2_{\hat{\beta}_j} = \sigma^2\left(\frac{1}{\mathrm{SXX}_j(1-R^2_{x_j\cdot x_{(j)}})}\right) = \frac{\sigma^2}{\mathrm{SXX}_j}\mathrm{VIF}\,,$$

where VIF denotes the *variance inflaction factor* due to correlation of x_j with the other predictors.

Fitted values and residuals: Since $\hat{\mathbf{y}} = \mathbf{Hy}$ and $\mathbf{HX} = \mathbf{X}$, $\mathrm{E}(\hat{\mathbf{y}}) = \mathbf{HX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{HH}' = \sigma^2\mathbf{H}$. Similarly, since $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, $\mathrm{E}(\hat{\mathbf{e}}) = \mathbf{0}$ and $\mathrm{var}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I} - \mathbf{H})$. It follows that

$$\mathrm{cov}(\hat{\mathbf{y}}, \hat{\mathbf{e}}) = \mathrm{cov}\left[\mathbf{Hy}, (\mathbf{I} - \mathbf{H})\mathbf{y}\right] = \mathbf{H}\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H}) = \mathbf{0}\,.$$

Standardized residuals: The ith *standardized residual* is defined as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}\,,$$

where h_{ii} the ith diagonal element of the hat-matrix, or ith *leverage value.*

3.2 Univariate Normal Distribution

A random variable Y is said to have a normal distribution with mean μ and variance σ^2 if its probability density function (pdf) is given by

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}} \qquad -\infty < y < \infty\,.$$

We write $Y \sim N(\mu, \sigma^2)$.

Linear transformation of a normal variable: If $Y \sim N(\mu, \sigma^2)$, and a and b are scalar constants, then

$$aY + b \sim\ N(a\mu + b, a^2\sigma^2)\,.$$

Standard normal variable: If $Y \sim N(\mu, \sigma^2)$, then

$$Z = \frac{Y - \mu}{\sigma} \sim\ N(0, 1)$$

has a *standard normal distribution*. The standard normal density is denoted by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \,.$$

Standard normal CDF and quantiles: The *cumulative distribution function* of the standard normal distribution is

$$\Phi(z) = \int_{-\infty}^{z} \phi(u) du \,.$$

The α-quantile of the standard normal distribution is the value z_α such that $\Phi(z_\alpha) = \alpha$; e.g. $z_{0.95} = 1.645$ and $z_{0.975} = 1.96$.

```
> pnorm(1.96); qnorm(0.975)

[1] 0.9750021

[1] 1.959964
```

3.3 Multivariate Normal Distribution

A random vector $\mathbf{y}$ has a *multivariate normal distribution* if and only if $\mathbf{a}'\mathbf{y}$ is univariate normal for any constant vector $\mathbf{a}$.

Multivariate normal density: If $\mathbf{y}$ is an n-dimensional multivariate normal variable with mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, we write $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma}$ is positive definite then $\mathbf{y}$ has pdf

$$f(\mathbf{y}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\} \quad \mathbf{y} \in R^p \,.$$

Linear transformation of a normal variable: If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}$ and $\mathbf{b}$ are $k \times n$ and $k \times 1$ matrices respectively, then

$$\mathbf{Ay} + \mathbf{b} \sim N_k(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}') \,.$$

That is, *linear transformations of normals are normal* !

Independence of normal variables: Suppose that $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\mathbf{y}' = (\mathbf{y}_1', \mathbf{y}_2')$ be a partition of $\mathbf{y}$ and suppose that the corresponding partition of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right] .$$

That is $\boldsymbol{\Sigma}_{jj}$ is the covariance matrix for $\mathbf{y}_j$, $j = 1, 2$, and the elements of $\boldsymbol{\Sigma}_{12}$ are the covariances between the componets of $\mathbf{y}_1$ and $\mathbf{y}_2$. (Note that $\mathbf{y}_1$ and $\mathbf{y}_2$ are clearly normal variables. Why?) Then, $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$. That is, *normal variables are independent if and only if they are uncorrelated.*

Indepedence of least squared estimates and MSE: It follows from the fact that $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$ that $\text{cov}(\hat{\mathbf{e}}, \hat{\boldsymbol{\beta}}) = 0$. Hence, under the normality assumptions on the errors in an MLR model, $\hat{\boldsymbol{\beta}}$ is independent of the residual vector and hence of the MSE.

3.4 Maximum Likelihood

Consider the MLR model and suppose that the errors are normally distributed, then $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. So the pdf of $\mathbf{y}$ is given by

$$f(\mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} .$$

Likelihood function: The likelihood function is the pdf of the data regarded as a function of the parameters. The log-likelihood function is the logarithm of the likelihood. For an MLR model this is given by

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) .$$

Maximum likelihood estimates: Maximizing the log-likelihod with respect to $\boldsymbol{\beta}$ is equivalent to minimizing the least squares criterion. Hence the MLE for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Since $\hat{\boldsymbol{\beta}}$ is linear in $\mathbf{y}$ we have

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right] .$$

The MLE for σ^2 satisfies the ML equation

$$-\frac{n}{2\sigma^2} + \frac{\text{SSE}}{2\sigma^4} = 0 \,,$$

and so $\hat{\sigma}^2 = \frac{1}{n}\text{SSE}$. Note that the divisor is n and not $n - p - 1$.

4 Distribution Theory

4.1 Chisquared Distribution

Central chi-squared variable: A scalar random variable X has a *central chi-squared distribution with k degrees of freedom* if it has the same distribution as the sum of k squared independent standard normal variables; i.e.

$$X \sim \sum_{i=1}^{k} Z_i^2 = \mathbf{z}'\mathbf{z}$$

where $\mathbf{z} \sim N_k(\mathbf{0}, \mathbf{I})$. Write $X \sim \chi^2_k$. Note that $\mathrm{E}(X) = k$ and $\mathrm{var}(X) = 2k$.

Non-central chi-squared variable: If $\mathbf{z} \sim N_k(\boldsymbol{\mu}, \mathbf{I})$ then $X = \mathbf{z}'\mathbf{z}$ is said to have a *non-central chi-squared distribution* with k degrees of freedom and *non-centrality parameter* $\lambda = \boldsymbol{\mu}'\boldsymbol{\mu}/2$. Write $X \sim \chi^2_k(\lambda)$. Note that $\mathrm{E}(X) = k + 2\lambda$ and $\mathrm{var}(X) = 2k + 8\lambda$ (Why?)

Sum of Independent chi-squared variables: If $X_1 \sim \chi^2_{k_1}(\lambda_1)$ and $X_2 \sim \chi^2_{k_2}(\lambda_2)$ are independent, then $X_1 + X_2 \sim \chi^2_{k_1+k_2}(\lambda_1 + \lambda_2)$.

Distribution of the sum of squared errors: Consider the MLR model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$. The sum of squared errors (or residual sum of squares) is given by SSE= $\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}$. Let $\mathbf{EDE}'$ denote the spectral decomposition of $\mathbf{I}-\mathbf{H}$ and note that $(\mathbf{I}-\mathbf{H})\mathbf{X} = \mathbf{0}$ and that $\mathbf{E}'\mathbf{e}$ has the same distribution as $\mathbf{e}$. Then,

$$\begin{aligned}\text{SSE} = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} &= \mathbf{e}'(\mathbf{I}-\mathbf{H})\mathbf{e} \\ &= \mathbf{e}'\mathbf{E}'\mathbf{DEe} \\ &\sim (\sigma\mathbf{z})'\mathbf{D}(\mathbf{z}\sigma)\end{aligned}$$

where $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I})$. But, since diag($\mathbf{D}$) consists of $n-p-1$ ones and $p+1$ zeros, it follows that

$$\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} \sim \sigma^2 \sum_{i=1}^{n-p-1} Z_i^2 \sim \sigma^2\chi^2_{n-p-1}\,.$$

Null distribution of regression sum of squares: Now let $\mathbf{EDE}'$ denote the spectral decomposition of $\mathbf{HCH}$. Note that, because $\mathbf{CH1} = \mathbf{0}$, if $\beta_1 = \cdots = \beta_p = 0$, then

$$\text{SSR} = \mathbf{y}'\mathbf{HCHy} = \mathbf{e}'\mathbf{EDE}'\mathbf{e}\,.$$

and

$$\mathbf{z} = \frac{1}{\sigma}\mathbf{E}'\mathbf{e} \sim N_n(\mathbf{0}, \mathbf{I})\,.$$

It follows that, under $H_0 : \beta_1 = \cdots = \beta_p = 0$,

$$\mathbf{y}'\mathbf{HCHy} \sim (\sigma\mathbf{z})'\mathbf{D}(\sigma\mathbf{z}) \sim \sigma^2\chi^2_p\,.$$

Furthermore, since $\hat{\mathbf{y}}$ is independent of $\hat{\mathbf{e}}$, SSR is independent of SSE.

4.2 Student's t-Distribution

T-variable: A scalar random variable T has a *central t-distribution with k degrees of freedom* if

$$T \sim \frac{Z}{\sqrt{X/k}}\,,$$

where $Z \sim N(0,1)$ independently of $X \sim \chi^2_k$. Write $T \sim t_k$. If $Z \sim N(\mu,1)$ then T has a *non-central t-distribution* with non-centrality parameter μ.

Studentized regression coefficents: Recall that σ^2 is estimated by the MSE; i.e. $\hat{\sigma}^2 = \text{SSE}/(n-p-1)$. Consider the *Studentized* regression coefficient,

$$\begin{aligned} T &= \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \\ &= \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}_{jj}}} \\ &= \frac{(\hat{\beta}_j - \beta_j)/\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}_{jj}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} \\ &\sim \frac{Z}{\sqrt{X/(n-p-1)}} \\ &\sim t_{n-p-1}\,. \end{aligned}$$

Wald statistic: The Wald statistic for testing $H_0 : \beta_j = \beta_{j0}$ is

$$W(\beta_{j0}) = T(\beta_{j0})^2 = \left(\frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}_{\hat{\beta}_j}} \right)^2 .$$

Under H_0, $W \sim t^2_{n-p-1}$. H_0 is rejected at significance level α if $W(\beta_{j0}) > t^2_{1-\alpha/2,n-p-1}$, or equivalently if $|T(\beta_{j0})| > t_{1-\alpha/2,n-p-1}$.

Confidence interval: A $100(1-\alpha)\%$ confidence interval for β_j is the set of values not rejected at significance level α; that is,

$$\begin{aligned} 100(1-\alpha)\% \text{ CI} &= \left\{\beta_j \| T(\beta_j)| < t_{1-\alpha/2,n-p-1}\right\} \\ &= \left(\hat{\beta}_j - t_{1-\alpha/2,n-p-1}\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{1-\alpha/2,n-p-1}\hat{\sigma}_{\hat{\beta}_j}\right) . \end{aligned}$$

Pointwise confidence bands for the expected response: The expected response for a given set of predictors $\mathbf{x} = (1, x_1, \ldots, x_p)'$ is estimated by

$$\hat{y}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}} = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim N\left(\mathbf{x}'\boldsymbol{\beta}, \sigma^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right) .$$

A $100(1-\alpha)\%$ confidence interval for $\mathbf{x}'\boldsymbol{\beta}$ is therefore given by

$$\hat{y}(\mathbf{x}) \pm t_{1-\alpha/2}\hat{\sigma}_{\hat{y}} ,$$

where $\hat{\sigma}_{\hat{y}} = \hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$.

Pointwise prediction bands: What can we say about a future response, y, at a given set of predictors? Note that $\text{var}(y - \mathbf{x}'\boldsymbol{\beta}) = \sigma^2$, but

$$\text{var}(y - \mathbf{x}'\hat{\boldsymbol{\beta}}) = \sigma^2\left(1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right) = \hat{\sigma}^2_{y-\hat{y}} ,$$

say. Hence a $100(1-\alpha)\%$ prediction interval for y is

$$\hat{y}(\mathbf{x}) \pm t_{1-\alpha/2}\hat{\sigma}_{y-\hat{y}} .$$

4.3 F Distribution

F-variable: A random variable F is said to have an F-distribution if

$$F \sim \frac{X_1/k_1}{X_2/k_2} ,$$

where $X_i \sim \chi^2_{k_i}$, $i = 1, 2$, independently.

Relationship between T and F distributions: If $Z \sim N(0,1)$ independently of $X \sim \chi^2_k$, then

$$T^2 \sim \frac{Z^2}{X/k} \sim F_{1,k}\,.$$

That is, the square of T-variable has an F-distribution.

Global F-test: If $\beta_1 = \cdots = \beta_p = 0$, then SSR $\sim \sigma^2\chi^2_p$ independently of SSE $\sim \sigma^2\chi^2_{n-p-1}$, and hence

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

We reject $H_0 : \beta_1 = \cdots = \beta_p = 0$ for large values of F; specifically, if $F > F_{1-\alpha,p,n-p-1}$. Note also that the global F-statistic is a function of the model's R-squared value, since

$$F = \frac{n-p-1}{p}\frac{R^2}{1-R^2}\,.$$

Comparison of Nested Models: Consider the *reduced* model matrix, $\mathbf{X}_k$, and let $\mathbf{X} = [\mathbf{X}_k, \bar{\mathbf{X}}_k]$ denote the corresponding partition of the *complete* model matrix.

Recall that, if $j > k$, $(\mathbf{H}_j - \mathbf{H}_{j-1})\mathbf{X}_k = \mathbf{0}$ and so, if $\beta_{k+1} = \cdots = \beta_p = 0$,

$$(\mathbf{H}_j - \mathbf{H}_{j-1})\mathbf{X}\boldsymbol{\beta} = (\mathbf{H}_j - \mathbf{H}_{j-1})[\mathbf{X}_k, \bar{\mathbf{X}}_k]\boldsymbol{\beta} = \mathbf{0}\,.$$

Hence, if $\beta_{k+1} = \cdots = \beta_p = 0$, the difference in the sum of squares explained by the complete and reduced models is

$$\begin{aligned}
\text{SSR(complete)} - \text{SSR(reduced)} &= \mathbf{y}'\left[\sum_{k+1}^{p}(\mathbf{H}_j - \mathbf{H}_{j-1})\right]\mathbf{y} \\
&\sim \mathbf{e}'\left[\sum_{k+1}^{p}(\mathbf{H}_j - \mathbf{H}_{j-1})\right]\mathbf{e} \\
&\sim (\sigma\mathbf{z})'\left[\sum_{k+1}^{p}(\mathbf{H}_j - \mathbf{H}_{j-1})\right](\sigma\mathbf{z})
\end{aligned}$$

where $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I})$. Thus, using the facts that $\text{tr}(\mathbf{H}_j - \mathbf{H}_{j-1}) = 1$ for $j = 1, \ldots, p$, and $(\mathbf{H}_j - \mathbf{H}_{j-1})(\mathbf{I} - \mathbf{H}) = \mathbf{0}$, it follows that

$$\text{SSR(complete)} - \text{SSR(reduced)} \sim \sigma^2 \chi^2_{p-k}$$

independently of the error sum of squares for the complete model, and so the statistic

$$F = \frac{[\text{SSR(complete)} - \text{SSR(reduced)}]/(p-k)}{\text{SSE(complete)}/(n-p-1)}$$

has an $F_{p-k,n-p-1}$ distribution if $\beta_{k+1} = \cdots = \beta_p = 0$.

4.4 Distribution of Quadratic Forms

Expected value of a quadratic form in a normal variable: Suppose that $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is positive definite. Then, for any symmetric $n \times n$ matrix $\mathbf{A}$,

$$E(\mathbf{y}'\mathbf{A}\mathbf{y}) = \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\,.$$

Proof: This result follows from the fact that $\text{tr}(AB) = \text{tr}(BA)$. Specifically,

$$\mathbf{y}'\mathbf{A}\mathbf{y} = (\mathbf{y} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{y} - \boldsymbol{\mu}) + 2\boldsymbol{\mu}'\mathbf{A}\mathbf{y} - \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\,,$$

and hence

$$\begin{aligned} \text{E}(\mathbf{y}'\mathbf{A}\mathbf{y}) &= \text{E}[(\mathbf{y} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{y} - \boldsymbol{\mu})] + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \\ &= \text{E}[\text{tr}(\mathbf{A}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})')] + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \\ &= \text{tr}[\mathbf{A}\text{E}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})')] + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \\ &= \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\,. \end{aligned}$$

Distribution of a quadratic form in a normal variable: Suppose in addition that

1. $\text{rank}(\mathbf{A}) = r$; and

2. $\mathbf{A\Sigma}$ is idempotent.

Then

$$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2_r(\lambda)\,,$$

where $\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

Note that this implies that under MLR assumptions, $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, if $\mathbf{A}$ is idempotent

$$\frac{\mathbf{y}'\mathbf{A}\mathbf{y}}{\sigma^2} \sim \chi^2_r(\lambda)\,, \quad \text{where} \quad \lambda = \frac{1}{2\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}\,.$$

Proof: Let $\mathbf{\Sigma}^{1/2}$ denote the symmetric square root of $\mathbf{\Sigma}$. Note that $\mathbf{A\Sigma}$ idempotent implies $\mathbf{A\Sigma A} = \mathbf{A}$ which in turn implies $\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$ is idempotent. It follows that $\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2} = \mathbf{E}_1\mathbf{E}_1'$, where $\mathbf{E}_1$ is $n \times r$ with orthonormal columns. Hence

$$\begin{aligned} \mathbf{y}'\mathbf{A}\mathbf{y} &= \mathbf{y}'\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1/2}\mathbf{y} \\ &= \mathbf{y}'\mathbf{\Sigma}^{-1/2}\mathbf{E}_1\mathbf{E}_1'\mathbf{\Sigma}^{-1/2}\mathbf{y} \\ &\sim \mathbf{z}'\mathbf{z}\,, \end{aligned}$$

where $\mathbf{z} \sim N(\mathbf{E}_1'\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu}, \mathbf{I}_r)$. That is $\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2_r(\lambda)$, where

$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{\Sigma}^{-1/2}\mathbf{E}_1'\mathbf{E}_1\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\,.$$

Regression sum of squares in MLR: Consider the MLR setting in which $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, and

$$\frac{1}{\sigma^2}\text{SSR} = \frac{1}{\sigma^2}\mathbf{y}'\mathbf{HCH}\mathbf{y}\,.$$

Furthermore,

$$\mathbf{A\Sigma} = \frac{1}{\sigma^2}\mathbf{HCH} \times \sigma^2\mathbf{I} = \mathbf{HCH}$$

is idempotent with rank p. Hence

$$\frac{1}{\sigma^2}\text{SSR} \sim \chi^2_p(\lambda) \quad \text{or equivalently} \quad \text{SSR} \sim \sigma^2\chi^2_p(\lambda)\,,$$

where

$$\lambda = \frac{1}{2\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{HCHX}\boldsymbol{\beta} = \frac{1}{2\sigma^2}\boldsymbol{\beta}\mathbf{X}'\mathbf{CX}\boldsymbol{\beta}\,.$$

Notice that

$$\text{E(SSR)} = \sigma^2\left(p + \frac{1}{\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{CX}\boldsymbol{\beta}\right)$$

and so

$$\text{E(MSR)} = \sigma^2\left(1 + \frac{1}{p\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{CX}\boldsymbol{\beta}\right) > \sigma^2\,.$$

Note also that the non-centrality parameter doesn't involve β_0 because $\mathbf{C1} = \mathbf{0}$, and that $\lambda = 0$ under the global null hypothesis, $\beta_1 = \cdots = \beta_p = 0$.

Cochran's Theorem: Suppose that $\mathbf{A}_i$ is an $n \times n$ symmetric, idempotent matrix with rank $= r_i$, for $i = 1, \ldots, k$, and that

$$\sum_{i=1}^{k} \mathbf{A}_i = \mathbf{I}_n \quad \text{and} \quad \sum_{i=1}^{k} r_i = n\,.$$

Then, under the standard MLR assumptions,

$$\mathbf{y}'\mathbf{A}_i\mathbf{y} \sim \sigma^2\chi^2_{r_i}(\lambda_i)$$

independently for $i = 1, \ldots, k$, where $\lambda_i = \frac{1}{2\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}_i\mathbf{X}\boldsymbol{\beta}$.

Proof: Using the Spectral Decomposition Theorem and the fact that $\mathbf{A}_i$ is idempotent of rank r_i we have the identity $\mathbf{A}_i = \mathbf{E}_i\mathbf{E}_i'$, where $\mathbf{E}_i$ is an $n \times r_i$ matrix with orthnormal columns. Let $\mathbf{E} = [\mathbf{E}_1, \dots, \mathbf{E}_k]$. Then, the identity

$$\mathbf{I} = \sum_{i=1}^{k} \mathbf{A}_i = \sum_{i=1}^{k} \mathbf{E}_i\mathbf{E}_i' = \mathbf{E}'\mathbf{E}$$

implies that $\mathbf{E}$ is an orthogonal matrix and so $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$ if $i \neq j$, and hence that $\mathbf{y}'\mathbf{A}_i\mathbf{y}$, $i = 1, \dots, k$ are independent.

Application to MLR: Recall that the matrices $\bar{\mathbf{J}}$, **HCH** and $\mathbf{I} - \mathbf{H}$, with ranks 1, p and $n - p - 1$ respectively, are all idempotent and sum to the identity. Furthermore

$$\mathbf{HCH} = \sum_{i=1}^{p} (\mathbf{H}_j - \mathbf{H}_{j-1}) \,,$$

where the matrices $\mathbf{H}_j - \mathbf{H}_{j-1}$, $j = 1, \dots, p$, are all idempotent and with ranks equal to 1.

5 Model Diagnostics

5.1 Residuals

Under MLR assumptions, $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, the residual vector, $\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, has variance-covariance matrix, $\sigma^2(\mathbf{I} - \mathbf{H})$. The ith Studentized residual is defined as

$$r_{i,std} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \,.$$

An alternative, *externally Studentized residual* is defined as

$$r_{i,ext} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \,,$$

where $\hat{\sigma}^2_{(i)}$ is the MSE for the MLR fit with the ith observation omitted, or *ith deletion variance*. Note that $\hat{\sigma}^2_{(i)}$ can be computed without refitting the model as

$$\hat{\sigma}^2_{(i)} = \frac{(n-2)\hat{\sigma}^2 - \hat{e}_i/(1-h_{ii})}{n-p-2} .$$

5.2 Cook's distance

Let $\hat{\boldsymbol{\beta}}_{(i)}$ denote the LS estimate of $\boldsymbol{\beta}$ if the ith observation is omitted. Then

$$\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

is the effect of omitting the ith observation on the vector of fitted values. Cook's distance is defined as the summary measure

$$C_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(p+1)\hat{\sigma}^2} .$$

The identity

$$C_i = \frac{h_{ii} r^2_{i,std}}{(p+1)(1-h_{ii})} .$$

implies that Cook's distance can be computed for all observations based on a single MLR fit to the complete data.

What values of Cook's distance indicate that the ith observation is influential? One answer is based on the amount by which a confidence set for $\boldsymbol{\beta}$ changes when an observation is omitted. A $100(1-\alpha)\%$ confidence set for $\boldsymbol{\beta}$ consists of those values not rejected at significance level α; i.e.

$$\left\{ \boldsymbol{\beta} \,|\, \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(p+1)\hat{\sigma}^2} < F_{1-\alpha}(p+1, n-p-1) \right\} .$$

For example, the ith observation might be considered influential if it moves the estimate outside of a 50% confidence set, or $C_i > F_{0.5}(p+1, n-p-1)$. Another rule-of-thumb is $C_i > 4/n$.

5.3 DFBETA and DFFIT

Recall that the t-statistic for testing $H_0 : \beta_j = \beta_{j0}$ is

$$T_j = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}} \,.$$

A measure of the influence of the ith observation on $\hat{\beta}_j$ is

$$DFBETA_j = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}} \,,$$

with a common cutoff being $|DFBETA_j| > 2/\sqrt{n}$.

A measure of the influence of a particular observation on its fitted value is

$$DFFIT_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} \,.$$

Suggested cutoffs for DFFIT values are $2\sqrt{(p+1)/n}$ and $2\sqrt{(p+1)/(n-p-1)}$.

5.4 COVRATIO

5.5 Added Variable Plots

Added variable plots allow one to graphically assess the linearity assumption for an individual predictor when it is added to the model in the same way that linearity can be assessed in SLR. Recall that the response vector, $\mathbf{y}$, is the sum of the fitted values, $\hat{\mathbf{y}}$, or explained variation, and the residuals, $\hat{\mathbf{e}}$, or unexplained variation. Write

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}(\mathbf{X}) \quad \text{and} \quad \hat{\mathbf{e}} = \hat{\mathbf{e}}(\mathbf{y}|\mathbf{X}) \,.$$

With this notation, if $\mathbf{X}_{-j}$ is the set of predictors with $\mathbf{x}_j$ omitted, $\hat{\mathbf{e}}(\mathbf{y}|\mathbf{X}_{-j})$ is the residual vector from regression of $\mathbf{y}$ on $\mathbf{X}_{-j}$, and $\hat{\mathbf{e}}(\mathbf{x}_j|\mathbf{X}_{-j})$ is the residual vector from the regression of $\mathbf{x}_j$ on $\mathbf{X}_{-j}$. Then, $\hat{\beta}_j$, the partial slope for the jth predictor from regressing $\mathbf{y}$ on $\mathbf{X}$ is the same as the slope from a regression of $\hat{\mathbf{e}}(\mathbf{y}|\mathbf{X}_{-j})$ on $\hat{\mathbf{e}}(\mathbf{x}_j|\mathbf{X}_{-j})$.

5.6 Box-Cox Transformation

The Box-Cox transformation is method for modifying the response with the goal of inducing normally distributed errors. The transformations consider are in a power family indexed by a paramter λ,

$$y(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda & \text{if} \quad \lambda \neq 0 \\ \log(y) & \text{if} \quad \lambda = 0 \end{cases}$$

It is assumed that there exists a value of λ such the the MLR assumptions hold. It is possible to constuct a confidence interval for λ by fitting the MLR over a range of values of the modified power transformation,

$$z(\lambda) = y(\lambda) \times \text{gm}(\mathbf{y})^{1-\lambda}.$$

where $\text{gm}(\mathbf{y})$ denotes the geometric mean of the response vector. See Cook & Weisberg (1999, Chapter 13) for further details.

6 Model Selection

Suppose there are p predictors, $x_1, \ldots, x_p$, available for explaining the variation in a response, y. Then, there are 2^p potential first-order MLR models, $2^{p+\binom{p}{2}}$ models if all two-factor interactions are considered, and $2^{2p+\binom{p}{2}}$ second-order models. In modern applications such as finance and genetics, the value of p can be very large. For example, genetic association studies may involve $p = O(10^3)$ or more SNPs.

In what follows, p refers to the total number of predictors in the model, including interactions and polynomial terms. We consider only *well-formed* *models* which always include lower order terms for predictors involved in higher-order interactions and polynomial terms. For example, the model, $E(y) = \beta_0 + \beta_1 x^2$, which implies that the response mean is maximized or minimized at $x = 0$, is not allowed. In particular, all models considered will contain an intercept parameter.

6.1 Model selection criteria

Rsquared: The R^2 value for an MLR model fit is given by

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} = \frac{\text{SSR}}{\text{TSS}}\,,$$

where TSS $= \sum_{i=1}^{n}(y_i - \bar{y})^2$. R^2 has the property that it cannot decrease when a new predictor is added to the model. Thus, the R^2 criterion does not penalize model complexity.

In contrast, the *adjusted* R^2 criterion,

$$R^2_{adj} = 1 - \frac{\text{SSE}/(n-p)}{\text{TSS}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-p}\,,$$

can decrease when predictors are added.

Mallow's C_k: Let MSE_p denote the MSE for the model containing all p predictors under consideration (plus the intercept), and let MSE_k denote the MSE for a reduced model with k predictors. Then Mallow's criterion is defined as

$$C_k = (n-k)\frac{\text{MSE}_k}{\text{MSE}_p} - (n - 2k)\,.$$

A reduced model provides a good fit if $\text{MSE}_k \approx \text{MSE}_p$ and hence $C_k \approx k$.

PRESS statistic: The *predicted residual sum of squares.*

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2\,,$$

measures how well a model can predict each response value using the remaining data. It can be shown that

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2 .$$

Akaike Information Criterion: Suppose that $y_1, \ldots, y_n$ are independent and identically distributed observations from a parametric model, $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-dimensional parameter. For example, if $y_i \sim N(\mu, \sigma^2)$, then $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $p = 2$. An alternative model is $y_i \sim G(\alpha, \beta)$. Consider selecting the model that best predicts a future observation, y, as measured by

$$E\left\{\log f(y|\boldsymbol{\theta})\right\} .$$

An unbiased estimate of this quantity is the sample average,

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}) = \frac{1}{n}\log\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) = \frac{1}{n}\log L(\boldsymbol{\theta}) ,$$

which is the log-likelhood as a function of $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}$ is unknown we replace it by the ML estimate $\hat{\boldsymbol{\theta}}$, resulting in

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) = \frac{1}{n}\log L(\hat{\boldsymbol{\theta}}) .$$

Akaike showed that this *plug-in* estimator is biased by and amount p/n (approximately for large n). This suggests selecting the model that maximizes the criterion

$$\frac{1}{n}\log L(\hat{\boldsymbol{\theta}}) - \frac{p}{n} ,$$

or equivalently mimimizes

$$\text{AIC} = -2\log\hat{L} + 2p .$$

A related statistic is the *Bayesian Information Criterion*,

$$\mathrm{BIC} = -2 \log \hat{L} + p \log n\,.$$

AIC for the MLR model: Consider the MLR model $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. The log-likelihood function is

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,.$$

Evaluating the log-likelihood at the ML estimates, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$, gives

$$\log \hat{L} = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}$$

and

$$\mathrm{AIC} = n\log(2\pi\hat{\sigma}^2) + n + 2(p+2)\,.$$

Comparison with log-normal MLR: Suppose that $\mathbf{y}_* = \log \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Then

$$\begin{aligned} f_Y(\mathbf{y}) &= f_{Y_*}(\mathbf{y}_*)\left|\frac{\partial \mathbf{y}_*}{\partial \mathbf{y}}\right| \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_*-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}_*-\mathbf{X}\boldsymbol{\beta})\right\}\prod_{i=1}^{n} y_i^{-1}\,. \end{aligned}$$

Hence

$$\mathrm{AIC}(\mathbf{y}) = n\log(2\pi\hat{\sigma}_*^2) + n + 2\sum_{i=1}^{n}\log y_i + 2(p+2) = \mathrm{AIC}(\mathbf{y}_*) + 2\sum_{i=1}^{n}\log y_i\,.$$

6.2 Least Absolute Shrinkage and Selection Operator

For $p > 0$ the L_p-norm of a vector $\mathbf{x}$ is defined as

$$||\mathbf{x}||_p = \left(\sum |x_i|^p\right)^{1/p}\,.$$

In particular, $||\mathbf{x}||_1 = \sum |x_i|$ and $||\mathbf{x}||_2 = \sqrt{\sum x_i^2}$. Hence, the least squares estimator is given by

$$\boldsymbol{\beta}_{ls} = \underset{\boldsymbol{\beta}}{\arg\min}\, ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2\,,$$

which equals $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$ provided $\mathbf{X}'\mathbf{X}$ is nonsingular.

The *ridge regression* estimator is obtained by least squares with a constraint. Specifically,

$$\boldsymbol{\beta}_{ridge} = \underset{\boldsymbol{\beta}}{\arg\min}\, ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \text{ subject to } ||\boldsymbol{\beta}||_2^2 \leq t$$

where $t > 0$. This is equivalent, for some $\lambda > 0$, to

$$\begin{aligned} \boldsymbol{\beta}_{ridge} &= \underset{\boldsymbol{\beta}}{\arg\min} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2 \right\} \\ &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}\,. \end{aligned}$$

(This optimization problem is related to random effects models, to be discussed later.)

The *LASSO* estimator (Tibshirani, 1996) is defined by

$$\boldsymbol{\beta}_{lasso} = \underset{\boldsymbol{\beta}}{\arg\min}\, ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \text{ subject to } ||\boldsymbol{\beta}||_1 \leq t$$

where $t > 0$, which is equivalent, for some $\lambda > 0$, to

$$\boldsymbol{\beta}_{lasso} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\}$$

Note both the ridge and LASSO estimators are defined even when $p > n$.

7 Categorical Predictors

Suppose that a categorical predictor or *factor* has k distinct levels or categories. Let

$$I_{ij} = \begin{cases} 1 & \text{if subject } i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases}$$

be an indicator for whether subject i is at level j of the factor. Then, for example, if $k = 4$ then the rows of the model matrix are

$$\begin{array}{ccccc r} 1 & 1 & 0 & 0 & 0 & \text{level/category } 1 \\ 1 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 1 & 0 & 3 \\ 1 & 0 & 0 & 0 & 1 & 4 \end{array}$$

The linear model

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j I_{ij} + e_i$$

implies the expected response in category j is

$$\mu_j = \beta_0 + \beta_j \, .$$

This model is *not identifiable* because there are $k+1$ parameters determining k means. More specifically, the columns of $\mathbf{X}$ are linearly dependent and so there isn't a unique solution to the least squares equations. We can make the model identifiable by constraining the parameters.

Mean model: If we set $\beta_0 = 0$, or equivalently, eliminate the column corresponding to the intercept in the model matrix, then

$$y_i = \sum_{j=1}^{k} \beta_j I_{ij} + e_i$$

and

$$\mu_j = \beta_j\,, \quad \text{for} \quad j = 1, \ldots, k\,.$$

Reference level: If we set $\beta_l = 0$ (typically, $l = 1$ or k), then

$$y_i = \beta_0 + \sum_{j \neq l}^{k} \beta_j I_{ij} + e_i$$

and

$$\begin{aligned} \mu_l &= \beta_0 \quad \text{and} \\ \mu_j &= \beta_0 + \beta_j \quad \text{for} \quad j \neq l\,. \end{aligned}$$

That is, category l is a reference level to which the other levels are compared, since for $j \neq l$

$$\beta_j = \mu_j - \mu_l\,.$$

This approach to dealing with categorical predictors is referred to as *dummy coding*.

Effects coding: Under the *sum constraint*, $\sum_{j=1}^{k} \beta_j = 0$,

$$y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j (I_{ij} - I_{ik}) + e_i$$

and

$$\begin{aligned} \mu_j &= \beta_0 + \beta_j \quad \text{and} \\ \frac{1}{k}\sum_{j=1}^{k} \mu_j &= \beta_0\,. \end{aligned}$$

Thus, β_j represents the deviation of level j from the mean response averaged over all categories (the *overall mean*). Notice that

$$\beta_k = -\sum_{j=1}^{k-1} \beta_j \, ,$$

so the coefficient associated with level k is a function of the others.

The rows of the model matrix when $k = 4$ under the different constraints are

$$\begin{array}{cccc|cccc|cccc|cccc}
\multicolumn{4}{c}{\beta_0 = 0} & \multicolumn{4}{c}{\beta_1 = 0} & \multicolumn{4}{c}{\beta_k = 0} & \multicolumn{4}{c}{\sum \beta_j = 0} \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & \text{-1} & \text{-1} & \text{-1}
\end{array}$$

Example: Consider the experiment to compare four diets involving beef or cereal as protein sources in either low or high amounts (Snedecor & Cochran, 1967). Forty rats were randomly assigned, 10 to each of the 4 diets. The response is weight gain.

```
> rats=read.table("../data/weight.txt",sep="",header=FALSE)
> colnames(rats)=c("BL","BH","CL","CH")
> apply(rats,2,mean)

   BL    BH    CL    CH
 79.2 100.0  83.9  85.9

> weight=scan("../data/weight.txt")
> treatment=factor(rep(c("BL","BH","CL","CH"),10))
```

The default in R is to use the first level as the reference category. This is the *contr.treatment* contrasts option.

```
> options(contrasts=c("contr.treatment","contr.poly"))
> lmfit.trt=lm(weight~treatment)
> summary(lmfit.trt)$coef
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept)    100.0   4.728577 21.148009 6.842420e-22
treatmentBL    -20.8   6.687218 -3.110411 3.644273e-03
treatmentCH    -14.1   6.687218 -2.108500 4.201233e-02
treatmentCL    -16.1   6.687218 -2.407578 2.130926e-02
```

Notice that, since the levels are ordered alphabetically, the first level is BH. The fitted model is

$$\hat{y} = 100.0 - 20.8 I_{BL} - 14.1 I_{CH} - 16.1 I_{CL}\,.$$

Use the *contr.SAS* option to make the last category the reference.

```
> options(contrasts=c("contr.SAS","contr.poly"))
> lmfit.sas=lm(weight~treatment)
> summary(lmfit.sas)$coef
```

```
            Estimate Std. Error    t value     Pr(>|t|)
(Intercept)     83.9   4.728577 17.7431799 2.212604e-19
treatmentBH     16.1   6.687218  2.4075780 2.130926e-02
treatmentBL     -4.7   6.687218 -0.7028333 4.866800e-01
treatmentCH      2.0   6.687218  0.2990780 7.666003e-01
```

In this case the fitted model is

$$\hat{y} = 83.9 + 16.1 I_{BH} - 4.7 I_{BL} + 2.0 I_{CH}\,.$$

Finally, use *contr.sum* to impose the sum constraints; i.e. effects coding.

```
> options(contrasts=c("contr.sum","contr.poly"))
> lmfit.sum=lm(weight~treatment)
> summary(lmfit.sum)$coef
```

```
            Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    87.25   2.364289 36.9032765 3.321099e-30
treatment1     12.75   4.095068  3.1135013 3.614434e-03
treatment2     -8.05   4.095068 -1.9657792 5.707375e-02
treatment3     -1.35   4.095068 -0.3296648 7.435638e-01
```

The fitted model using effects coding is

$$\hat{y} = 87.25 + 12.75(I_{BH} - I_{CL}) - 8.05(I_{BL} - I_{CL}) - 1.35(I_{CH} - I_{CL})\,.$$

```
> anova(lmfit.sum)
```

```
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value  Pr(>F)
treatment  3 2404.1  801.37   3.584 0.02297
Residuals 36 8049.4  223.59
```

Global F-test: The hypothesis, $H_0 : \beta_1 = \cdots = \beta_k = 0$, of *no difference between the treatment means* can be assessed using the global F-test which has $k-1$ and $N-k$ degrees of freedom if N is the total sample size.

Analysis as a 2×2 ***factorial experiment***: The four diets are formed as combinations of two factors, protein source (B or C) and protein amount (H or L), with 10 rats assigned to each combination. Thus, an equivalent model using the first level of each factor as the reference is

$$\mu = \beta_0 + \beta_1 I_C + \beta_2 I_L + \beta_3 I_C I_L \,.$$

In SAS the model is

$$\mu = \beta_0 + \beta_1 I_B + \beta_2 I_H + \beta_3 I_B I_H \,,$$

and under sum constraints it is

$$\mu = \beta_0 + \beta_1 (I_B - I_C) + \beta_2 (I_H - I_L) + \beta_3 (I_B - I_C)(I_H - I_L) \,.$$

The means as a function of the paramters under the different coding schemes are given in the following table.

Treatment	R	SAS	Sum
BH	β_0	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
BL	$\beta_0 + \beta_2$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 - \beta_2 - \beta_3$
CH	$\beta_0 + \beta_1$	$\beta_0 + \beta_2$	$\beta_0 - \beta_1 + \beta_2 - \beta_3$
CL	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	β_0	$\beta_0 - \beta_1 - \beta_2 + \beta_3$

Model notation using factor level indexing: Consider a single factor experiment. Let y_{ij} denote the jth response at level i. Then, we can write the model as

$$y_{ij} = \mu + \alpha_i + e_{ij} \,,$$

where μ is the overall mean and $\sum_i \alpha_i = 0$.

If there are two (crossed) factors, as with the diet experiment, then the model is written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} \,,$$

with constraints,

$$\sum_i \alpha_i = 0\,, \quad \sum_j \beta_j = 0\,, \quad \sum_i \alpha\beta_{ij} = \sum_j \alpha\beta_{ij} = 0\,.$$

In this case the α and β coefficients represent *main* or *additive* effects of the factor levels, whereas the $\alpha\beta$ coefficients represent deviations from additivity or *interactions* between the two factors.

For example, if protein source is the α-factor and protein amount is the β-factor, then

$$\begin{aligned}
\mu &= \frac{1}{4}\left[\mu_{BH} + \mu_{BL} + \mu_{CH} + \mu_{CL}\right] \\
\alpha_1 &= \frac{1}{4}\left[\mu_{BH} + \mu_{BL} - \mu_{CH} - \mu_{CL}\right]\,, \quad \alpha_2 = -\alpha_1 \\
\beta_1 &= \frac{1}{4}\left[\mu_{BH} - \mu_{BL} + \mu_{CH} - \mu_{CL}\right]\,, \quad \beta_2 = -\beta_1 \\
\alpha\beta_{11} &= \frac{1}{4}\left[\mu_{BH} - \mu_{BL} - \mu_{CH} + \mu_{CL}\right]\,, \quad \alpha\beta_{12} = \alpha\beta_{21} = -\alpha\beta_{11} = -\alpha\beta_{22}
\end{aligned}$$

```
> protein=factor(rep(c("B","C"),10,each=2))
> amount=factor(rep(c("L","H"),20))
> lmfit.2b2=lm(weight~protein*amount)
> summary(lmfit.2b2)$coef

                 Estimate Std. Error    t value     Pr(>|t|)
(Intercept)         87.25   2.364289 36.9032765 3.321099e-30
protein1             2.35   2.364289  0.9939564 3.268783e-01
amount1              5.70   2.364289  2.4108731 2.114491e-02
protein1:amount1     4.70   2.364289  1.9879129 5.446757e-02
```

Linear contrasts among treatment means: A linear combination of treatment means

$$L = \sum_{i=1}^{k} c_i \mu_i\,,$$

is called a linear contrast if $\sum_1^k c_i = 0$. An unbiased estimate of L is given by the corresponding contrast among sample means,

$$\hat{L} = \sum_{i=1}^{k} c_i \bar{y}_i \,.$$

Under MLR assumptions, the variance of $\hat{L}$ is given by

$$\text{var}(\hat{L}) = \sum_{i=1}^{k} c_i^2 \frac{\sigma^2}{n_i} = \sigma^2 \sum_{i=1}^{k} \frac{c_i^2}{n_i} \,.$$

In the rat diet study for example, $n_{ij} = 10$ for all four source/amount combinations. Hence

$$\text{var}(\hat{\alpha}_1) = \sigma^2 \sum_{ij} \frac{(1/4)^2}{10} = \frac{\sigma^2}{40} \,.$$

Since $\hat{\sigma}^2 = 223.59$ the estimated variance is

$$\hat{\text{var}}(\hat{\alpha}_1) = \frac{223.59}{40}$$

and the estimated standard error is

$$\hat{\text{se}}(\hat{\alpha}_1) = \sqrt{\frac{223.59}{40}} = 2.364 \,.$$

Models with both categorical and numerical predictors: Consider a situation with one categorical predictor T and one numerical predictor, X. The model

$$y_i = \alpha_0 + \sum_{j=1}^{k} \alpha_j I_{ij} + \beta_0 x_i + e_i \,,$$

where $\sum_1^k \alpha_j = 0$, implies that differences between the levels of T are independent of the value of X. Conversely, the effect of changing the value of X

is the same at all levels of T. In this situation it is conventional to compare the levels of T at the average value of X. The estimated response means at the average value of X are referred to as *least squares means* or *lsmeans* or sometimes *adjusted means.* These are given by

$$\hat{\mu}_j = \hat{\alpha}_0 + \hat{\alpha}_j + \hat{\beta}_0 \bar{x} = \bar{y}_j - \hat{\beta}_0(\bar{x}_j - \bar{x})\,.$$

Consider the following extension of the previous model,

$$y_i = \alpha_0 + \sum_{j=1}^{k} \alpha_j I_{ij} + \beta_0 x_i + \sum_{j=1}^{k} \beta_j I_{ij} x_i + e_i\,.$$

where $\sum_1^k \alpha_j = 0$ and $\sum_1^k \beta_j = 0$. In this case the difference between levels of T (in terms of the expected response) depends on the value of X. That is, there is *interaction* - the relationship between Y and T depends on X. We can test the hypothesis of no interaction, $H_0 : \beta_1 = \cdots = \beta_k$, by comparing the fits of the two models.

The same model using factor level indexing is given by

$$y_{ij} = (\alpha_0 + \alpha_i) + (\beta_0 + \beta_i)x_{ij} + e_{ij}\,,$$

where $\sum_1^k \alpha_i = 0$ and $\sum_1^k \beta_i = 0$. The indicators are not needed with this notation because the indexes on the responses indicate the factor level.

8 Balanced single factor design

Consider a balanced single factor design:

level	1	2	$\cdots$	n
1	y_{11}	y_{12}	$\cdots$	y_{1n}
2				
$\vdots$				
i	y_{i1}	y_{i2}	$\cdots$	y_{in}
$\vdots$				
k	y_{k1}	y_{k2}	$\cdots$	y_{kn}

where y_{ij} is the jth response at level i.

8.1 Fixed effects model

If the k factor levels in the experiment are the only ones of interest, then the *fixed effects* model

$$y_{ij} = \mu + \alpha_i + e_{ij}\,,$$

where $\sum_1^k \alpha_i = 0$ and $e_{ij} \sim N(0, \sigma_e^2)$ independently, is appropriate. Note that this model has a total of $k+1$ parameters (including the error variance).

Let

$$\mathbf{y} = (y_{11}, \ldots, y_{1n}, y_{21}, \ldots, y_{2n}, \ldots, y_{k1}, \ldots, y_{kn})',$$

let $\mu_i = \mu + \alpha_i$ denote the expected response at level i, and let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$. Then, in matrix notation the model is

$$\begin{aligned}\mathbf{y} &= (\mathbf{I}_k \otimes \mathbf{1}_n)\boldsymbol{\mu} + (\mathbf{I}_k \otimes \mathbf{I}_n)\mathbf{e} \\ &= (\mathbf{I}_k \otimes \mathbf{1}_n)\mathbf{M}^{-1}\boldsymbol{\beta} + (\mathbf{I}_k \otimes \mathbf{I}_n)\mathbf{e}\,,\end{aligned}$$

where $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_k \otimes \mathbf{I}_n)$, $\boldsymbol{\beta} = \mathbf{M}\boldsymbol{\mu} = (\mu, \alpha_1, \ldots, \alpha_{k-1})'$, and

$$\mathbf{M} = \left[\begin{array}{ccccc} 1/k & 1/k & \cdots & 1/k & 1/k \\ 1-1/k & -1/k & \cdots & -1/k & -1/k \\ & & \cdots & & \\ -1/k & -1/k & \cdots & 1-1/k & -1/k \end{array}\right].$$

8.2 Random effects model

Now suppose that the levels of the factor occurring in the experiment represent a random sample from a (large or potentially infinite) population. For example, genetic varieties in a study of crop yield. In such cases it makes sense to model the factor level effects as random as opposed to fixed. (If the experiment was repeated, it would involve a different sample of factor levels.) Consider the model

$$y_{ij} = \mu + a_i + e_{ij} \, ,$$

where $a_i \sim N(0, \sigma_a^2)$ independently of $e_{ij} \sim N(0, \sigma_e^2)$. Here we have assumed that the factor levels in the experiment are a random sample from a normal distribution. The parameter μ represents the expected response averaged over the (perhaps hyperthetical) population of factor levels. Note that this model has only three parameters regardless of the value of k.

In matrix notation the model is

$$\mathbf{y} = (\mathbf{1}_k \otimes \mathbf{1}_n)\mu + (\mathbf{I}_k \otimes \mathbf{1}_n)\mathbf{a} + (\mathbf{I}_k \otimes \mathbf{I}_n)\mathbf{e}$$

where $\mathbf{a} \sim N(\mathbf{0}_k, \sigma_a^2 \mathbf{I}_k)$ and $\mathbf{e} \sim N(\mathbf{0}_{kn}, \sigma_e^2 \mathbf{I}_k \otimes \mathbf{I}_n)$ independently. Because $\mathbf{a}$ and $\mathbf{e}$ are independent, the variance-covariance matrix for the response vector is

$$\boldsymbol{\Sigma} = \sigma_a^2(\mathbf{I}_k \otimes \mathbf{J}_n) + \sigma_e^2(\mathbf{I}_k \otimes \mathbf{I}_n) = \mathbf{I}_k \otimes (\sigma_a^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n) \, .$$

The model imposes a covariance structure on the responses at the same level. Specifically, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})'$, then

$$\mathbf{y}_i = \mathbf{1}_n \mu + \mathbf{1}_n a_i + \mathbf{e}_i \, ,$$

and $\mathbf{y}_i \sim N(\mathbf{1}_n \mu, \sigma_a^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n)$ independently for $i = 1, \ldots, n$. In particular, the model implies that the correlation between two responses at the same level is given by

$$\mathrm{cor}(y_{ij}, y_{ik}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \, .$$

8.3 Estimation

ML estimation for a one-way random effects model: The likelihood function based on all the data is

$$\begin{aligned} L(\mu,\sigma_a^2,\sigma_e^2) &= \prod_{i=1}^{k} |2\pi(\sigma_a^2\mathbf{J}_n+\sigma_e^2\mathbf{I}_n)|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}(\mathbf{y}_i-\mathbf{1}_n\mu)'(\sigma_a^2\mathbf{J}_n+\sigma_e^2\mathbf{I}_n)^{-1}(\mathbf{y}_i-\mathbf{1}_n\mu)\right\}\,, \end{aligned}$$

and so the log-likelihood is

$$\begin{aligned} l(\mu,\sigma_a^2,\sigma_e^2) &= -\frac{1}{2}\sum_{i=1}^{k}\ln|2\pi(\sigma_a^2\mathbf{J}_n+\sigma_e^2\mathbf{I}_n)| \\ &\quad -\frac{1}{2}\sum_{i=1}^{k}(\mathbf{y}_i-\mathbf{1}_n\mu)'(\sigma_a^2\mathbf{J}_n+\sigma_e^2\mathbf{I}_n)^{-1}(\mathbf{y}_i-\mathbf{1}_n\mu) \end{aligned}$$

Maximizing the log-likelihood with respect to the three model parameters results in the estimates

$$\begin{aligned} \hat{\mu} &= \bar{y}_{\cdot\cdot} \\ \hat{\sigma}_e^2 &= \frac{1}{k(n-1)}W \\ \hat{\sigma}_a^2 &= \frac{1}{nk}B-\frac{1}{nk(n-1)}W\,, \end{aligned}$$

provided $\hat{\sigma}_a^2 > 0$, where

$$W=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2=\mathbf{y}'(\mathbf{I}_k\otimes\mathbf{C}_n)\mathbf{y}$$

and

$$B=\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2=\mathbf{y}'(\mathbf{C}_k\otimes\bar{\mathbf{J}}_n)\mathbf{y}$$

are the *within* and *between* sums of squares respectively.

Restricted/Residual maximum likelihood: The modified data vector, $\mathbf{y}^* = \mathbf{C}_{kn}\mathbf{y}$, has zero mean since $\mathbf{C}_{kn}\mathbf{1}_{kn}\mu = \mathbf{0}_{kn}$, and variance-covariance matrix, $\boldsymbol{\Sigma}^* = \mathbf{C}_{kn}\boldsymbol{\Sigma}\mathbf{C}_{kn}$, which only depends on the variance components, σ_a^2 and σ_e^2. Hence, we can estimate the variance components independently of μ by maximizing the likelihood based on the modified data vector. The REML estimates of the variance components are

$$\begin{aligned}\hat{\sigma}_e^2 &= \frac{1}{k(n-1)}W\\ \hat{\sigma}_a^2 &= \frac{1}{n(k-1)}B - \frac{1}{nk(n-1)}W\,,\end{aligned}$$

provided $\hat{\sigma}_a^2 > 0$.

Method of moments: Note that

$$W = \mathbf{y}'(\mathbf{I}_k \otimes \mathbf{C}_n)\mathbf{y} = \sum_{i=1}^{k} \mathbf{y}_i'\mathbf{C}_n\mathbf{y}_i\,.$$

and $\mathbf{y}_i \sim N(\mathbf{1}_n\mu, \sigma_a^2\mathbf{J}_n + \sigma_e^2\mathbf{I}_n)$ independently. Hence

$$\begin{aligned}\mathrm{E}(W) &= k\,\mathrm{tr}\left[\mathbf{C}_n(\sigma_a^2\mathbf{J}_n + \sigma_e^2\mathbf{I}_n)\right] + k\,\mu\mathbf{1}_n'\mathbf{C}_n\mathbf{1}_n\mu\\ &= k(n-1)\sigma_e^2\,.\end{aligned}$$

It follows that

$$\hat{\sigma}_e^2 = \frac{1}{k(n-1)}W$$

is an unbiased estimator of σ_e^2. Similarly,

$$\begin{aligned}\mathrm{E}(B) &= E\left[\mathbf{y}'(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)\mathbf{y}\right]\\ &= \mathrm{tr}\left\{(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)[\sigma_a^2(\mathbf{I}_k \otimes \mathbf{J}_n) + \sigma_e^2(\mathbf{I}_k \otimes \mathbf{I}_n)]\right\}\\ &\quad +\mu(\mathbf{1}_k \otimes \mathbf{1}_n)'(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)(\mathbf{1}_k \otimes \mathbf{1}_n)\mu\\ &= (n\sigma_a^2 + \sigma_e^2)\mathrm{tr}(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)\\ &= \sigma_a^2 n(k-1) + \sigma_e^2(k-1)\,.\end{aligned}$$

Hence

$$\frac{1}{n(k-1)}B - \frac{1}{nk(n-1)}W$$

is an unbiased estimate of σ_a^2.

Distribution of within and between sums of squares: Note that

$$(\mathbf{I}_k \otimes \mathbf{C}_n)\mathbf{\Sigma} = (\mathbf{I}_k \otimes \mathbf{C}_n)\left[\sigma_a^2(\mathbf{I}_k \otimes \mathbf{J}_n) + \sigma_e^2(\mathbf{I}_k \otimes \mathbf{I}_n)\right] = \sigma_e^2(\mathbf{I}_k \otimes \mathbf{C}_n)$$

which implies that $(\mathbf{I}_k \otimes \mathbf{C}_n)\mathbf{\Sigma}/\sigma_e^2$ is idempotent, and

$$\mu(\mathbf{1}_k \otimes \mathbf{1}_n)'(\mathbf{I}_k \otimes \mathbf{C}_n)(\mathbf{1}_k \otimes \mathbf{1}_n)\mu = 0\,.$$

Since $\text{rank}(\mathbf{I}_k \otimes \mathbf{C}_n) = k(n-1)$, it follows that

$$W = \mathbf{y}'(\mathbf{I}_k \otimes \mathbf{C}_n)\mathbf{y} \sim \sigma_e^2\chi^2_{k(n-1)}\,.$$

Similarly, since

$$\begin{aligned}(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)\mathbf{\Sigma} &= (\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)\left[\sigma_a^2(\mathbf{I}_k \otimes \mathbf{J}_n) + \sigma_e^2(\mathbf{I}_k \otimes \mathbf{I}_n)\right] \\ &= \sigma_a^2(\mathbf{C}_k \otimes \mathbf{J}_n) + \sigma_e^2(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n) \\ &= (n\sigma_a^2 + \sigma_e^2)(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)\end{aligned}$$

and

$$\mu(\mathbf{1}_k \otimes \mathbf{1}_n)'(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)(\mathbf{1}_k \otimes \mathbf{1}_n)\mu = 0$$

and $\text{rank}(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n) = k-1$,

$$B = \mathbf{y}'(\mathbf{C}_k \otimes \bar{\mathbf{J}}_n)\mathbf{y} \sim (n\sigma_a^2 + \sigma_e^2)\chi^2_{k-1}\,.$$

Also, since W and B are independent (why?),

$$\frac{B/(k-1)}{W/k(n-1)} \sim \frac{n\sigma_a^2 + \sigma_e^2}{\sigma_e^2} \times F_{k-1,k(n-1)}\,.$$

8.4 Prediction

Conditional distributions of normal subvectors: Suppose that $\mathbf{y} = (y_1, y_2)'$ has a bivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and variance-covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ such that $|\boldsymbol{\Sigma}| > 0$. The conditional distribution of y_2 given y_1 is obtained as

$$f_{2|1}(y_2|y_1) = \frac{f(y_1, y_2)}{f_1(y_1)} \, ,$$

where

$$f(y_1, y_2) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\}$$

is the joint distribution of $\mathbf{y}$ and

$$f_1(y_1) = \frac{1}{\sqrt{2\pi\sigma_{11}}} \exp\left\{-\frac{(y_1 - \mu_1)^2}{2\sigma_{11}}\right\} \, .$$

After some algebra it can be shown that

$$f_{2|1}(y_2|y_1) = \frac{1}{\sqrt{2\pi\sigma_{2|1}}} \exp\left\{-\frac{(y_2 - \mu_{2|1})^2}{2\sigma_{2|1}}\right\} \, ,$$

where

$$\begin{aligned} \mu_{2|1} &= \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(y_1 - \mu_1) \\ \sigma_{2|1} &= \sigma_{22} - \frac{\sigma_{21}\sigma_{12}}{\sigma_{11}} \, . \end{aligned}$$

That is, the conditional distribution of y_2 given y_1 is $N(\mu_{2|1}, \sigma_{2|1})$.

The generalization of this result to the full multivariate normal settings is as follows. Suppose that $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (with $|\boldsymbol{\Sigma}| > 0$), and $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ is a partition of $\mathbf{y}$ with corresponding mean partition $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and covariance partition

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{22} \end{array}\right] .$$

Then the condition distribution of $\mathbf{y}_2$ given $\mathbf{y}_1$ is normal with mean and covariance

$$\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$$

and

$$\boldsymbol{\Sigma}_{2|1} \quad = \quad \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

respectively.

Prediction of random effects in a balance one-way experiment: Averaging over the replicate responses at a given factor level results in the model,

$$\bar{y}_{i\cdot} = \mu + a_i + \bar{e}_{i\cdot}\,.$$

Hence,

$$\left[\begin{array}{c} \bar{y}_{i\cdot} \\ a_i \end{array}\right] \sim N\left(\left[\begin{array}{c} \mu \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma_a^2 + \sigma_e^2/n & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 \end{array}\right]\right),$$

and

$$a_i|\bar{y}_{i\cdot} \sim N\left[\frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/n}(\bar{y}_i - \mu)\,, \frac{\sigma_a^2\sigma_e^2/n}{\sigma_a^2 + \sigma_e^2/n}\right]$$

Replacing μ by its MLE in the conditional mean results in the *best linear unbiased predictor* (BLUP) of the random effect a_i; i.e.

$$\text{BLUP}(a_i) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\,.$$

9 Balanced Two-Factor crossed designs

Consider a design with two *crossed* factors A and B with a and b levels respectively. A crossed design is one in which every combination of the factors occurs in the experiment. Assume for now that the design is balanced in the sense that there are exactly r replicate response measurements at each of the ab *treatment* combinations. Let y_{ijk} denote the kth replicate response at treatment combination (i, j). Note that, if replicate (R) is considered as a third factor, then R is *nested* within the AB combinations. Let $\mathbf{y}$ denote the full response vector ordered according to the convention by which subscripts on the right complete a cycle before subscripts to their left are updated.

9.1 Single replicate case

If $r = 1$ then the third subscript is redundant and so

$$\mathbf{y} = (y_{11}, \ldots, y_{1b}, y_{21}, \ldots, y_{2b}, \ldots, y_{a1}, \ldots, y_{ab})' .$$

The ANOVA decomposition for this design is given by

Source	Sum of Squares		DF
Intercept	$\sum_{i=1}^{a} \sum_{j=1}^{b} \bar{y}_{\cdot\cdot}^2 =$	$\mathbf{y}'(\bar{\mathbf{J}}_a \otimes \bar{\mathbf{J}}_b)\mathbf{y}$	1
A	$\sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 =$	$\mathbf{y}'(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b)\mathbf{y}$	$a - 1$
B	$\sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 =$	$\mathbf{y}'(\bar{\mathbf{J}}_a \otimes \mathbf{C}_b)\mathbf{y}$	$b - 1$
Error	$\sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2 =$	$\mathbf{y}'(\mathbf{C}_a \otimes \mathbf{C}_b)\mathbf{y}$	$(a - 1)(b - 1)$
Total		$\mathbf{y}'(\mathbf{I}_a \otimes \mathbf{I}_b)\mathbf{y}$	ab

Note that this decomposition is determined by the design and is not related to any distributional assumptions about the factors or error terms.

Both factors fixed: If both A and B are fixed factors a potential statistical model is

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} ,$$

where $\sum_1^a \alpha_i = \sum_1^b \beta_j = 0$, and where $e_{ij} \sim N(0, \sigma_e^2)$ independently. Note that the model implies that the effects of the two factors are additive.

The matrix form of this model is

$$\mathbf{y} = (\mathbf{1}_a \otimes \mathbf{1}_b)\mu + (\mathbf{I}_a \otimes \mathbf{1}_b)\boldsymbol{\alpha} + (\mathbf{1}_a \otimes \mathbf{I}_b)\boldsymbol{\beta} + \mathbf{e}\,,$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_a)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_b)'$ satisfy $\mathbf{1}_a'\boldsymbol{\alpha} = 0$ and $\mathbf{1}_b'\boldsymbol{\beta} = 0$, and $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_a \otimes \mathbf{I}_b)$.

Let $\mathbf{A}_1 = \bar{\mathbf{J}}_a \otimes \bar{\mathbf{J}}_b$, $\mathbf{A}_2 = \mathbf{C}_a \otimes \bar{\mathbf{J}}_b$, $\mathbf{A}_3 = \bar{\mathbf{J}}_a \otimes \mathbf{C}_b$, and $\mathbf{A}_4 = \mathbf{C}_a \otimes \mathbf{C}_b$. Note $\mathbf{A}_i$ is symmetric and idempotent with rank r_i, $i = 1, \ldots, 4$, such that

$$\mathbf{I}_{ab} = \mathbf{I}_a \otimes \mathbf{I}_b = \sum_{i=1}^{4} \mathbf{A}_i \quad \text{and} \quad \sum_{i=1}^{4} r_i = ab\,.$$

It follows by Cochran's theorem that

$$\mathbf{y}'\mathbf{A}_i\mathbf{y} \sim \sigma_e^2 \chi^2_{r_i}(\lambda_i)$$

independently for $i = 1, \ldots, 4$, where

$$\lambda_i = \frac{1}{2\sigma_e^2}\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}\,.$$

For example, because $\mathbf{C}_a\mathbf{1}_a = \mathbf{0}_a$ and $\mathbf{C}_b\mathbf{1}_b = \mathbf{0}_b$,

$$\lambda_4 = \frac{1}{2\sigma_e^2}\boldsymbol{\mu}'(\mathbf{C}_a \otimes \mathbf{C}_b)\boldsymbol{\mu} = 0,$$

and so

$$\mathbf{y}'(\mathbf{C}_a \otimes \mathbf{C}_b)\mathbf{y} \sim \sigma_e^2 \chi^2_{(a-1)(b-1)}\,.$$

Similarly,

$$\begin{aligned}
\lambda_2 = \frac{1}{2\sigma_e^2}\boldsymbol{\mu}'(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b)\boldsymbol{\mu} &= \frac{1}{2\sigma_e^2}\boldsymbol{\alpha}'(\mathbf{I}_a \otimes \mathbf{1}_b)'(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b)(\mathbf{I}_a \otimes \mathbf{1}_b)\boldsymbol{\alpha} \\
&= \frac{1}{2\sigma_e^2}\boldsymbol{\alpha}'(\mathbf{C}_a \otimes \mathbf{1}_b'\mathbf{1}_b)\boldsymbol{\alpha} \\
&= \frac{b}{2\sigma_e^2}\boldsymbol{\alpha}'\mathbf{C}_a\boldsymbol{\alpha} \\
&= \frac{b}{2\sigma_e^2}\sum_{i=1}^{a}\alpha_i^2 \\
&= \frac{b(a-1)\theta_\alpha}{2\sigma_e^2}\,,
\end{aligned}$$

and so

$$\mathbf{y}'(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b)\mathbf{y} \sim \sigma_e^2\chi^2_{a-1}(\lambda_2)\,.$$

Recall that, if $X \sim \chi_d^2(\lambda)$, then $\mathrm{E}(X) = d + 2\lambda$. Hence, the expected mean squared error is

$$\mathrm{E}(\mathrm{MSE}) = \mathrm{E}\left\{\frac{\mathbf{y}'(\mathbf{C}_a \otimes \mathbf{C}_b)\mathbf{y}}{(a-1)(b-1)}\right\} = \sigma_e^2\,,$$

whereas the expected mean square for factor A is

$$\mathrm{E}(\mathrm{MSA}) = \mathrm{E}\left\{\frac{\mathbf{y}'(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b)\mathbf{y}}{(a-1)}\right\} = \sigma_e^2 + b\theta_\alpha\,.$$

We can test the hypothesis $H_0 : \alpha_1 = \cdots = \alpha_a$ of no difference between the levels of factor A using the ratio of mean squares

$$F = \frac{\mathrm{MSA}}{\mathrm{MSE}} \sim F_{a-1,(a-1)(b-1)}(\lambda_2)\,.$$

Under H_0, $\lambda_2 = 0$ and the F-statistic has a central F-distribution. On the other hand, if H_0 is false, the expected value of the numerator is greater than that of the denominator and the F-statistic has a distribution that is stochastically larger than a central F-distribution. Thus, large values of F lead to rejection of H_0.

Randomized blocks experiment: Suppose that factor B is a random blocking factor. That is, if the experiment was repeated it would involve a different random sample of blocks. In this case the model is

$$\mathbf{y} = (\mathbf{1}_a \otimes \mathbf{1}_b)\mu + (\mathbf{I}_a \otimes \mathbf{1}_b)\boldsymbol{\alpha} + (\mathbf{1}_a \otimes \mathbf{I}_b)\mathbf{b} + \mathbf{e}\,,$$

where $\mathbf{b} \sim N(\mathbf{0}_b, \sigma_b^2\mathbf{I}_b)$ and $\mathbf{e} \sim N(\mathbf{0}_{ab}, \sigma_e^2\mathbf{I}_{ab})$. The variance-covariance matrix for the response vector is therefore,

$$\boldsymbol{\Sigma} = \sigma_b^2(\mathbf{J}_a \otimes \mathbf{I}_b) + \sigma_e^2(\mathbf{I}_a \otimes \mathbf{I}_b)\,.$$

In this case, responses in the same block are dependent, and so Cochran's theorem does not apply. However, the fact that

$$\mathbf{A}_4\boldsymbol{\Sigma} = (\mathbf{C}_a \otimes \mathbf{C}_b)\left[\sigma_b^2(\mathbf{J}_a \otimes \mathbf{I}_b) + \sigma_e^2(\mathbf{I}_a \otimes \mathbf{I}_b)\right] = \sigma_e^2(\mathbf{C}_a \otimes \mathbf{C}_b)\,,$$

and

$$\mathbf{A}_2\boldsymbol{\Sigma} = (\mathbf{C}_a \otimes \bar{\mathbf{J}}_b)\left[\sigma_b^2(\mathbf{J}_a \otimes \mathbf{I}_b) + \sigma_e^2(\mathbf{I}_a \otimes \mathbf{I}_b)\right] = \sigma_e^2(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b)$$

imply that the sums of squares for error and factor A have the same distribution as in the fixed effects case. Note that independence follows from the fact that $\mathbf{A}_2\boldsymbol{\Sigma}\mathbf{A}_4 = \mathbf{0}$.

On the other hand

$$\mathbf{A}_3\boldsymbol{\Sigma} = (\bar{\mathbf{J}}_a \otimes \mathbf{C}_b)\left[\sigma_b^2(\mathbf{J}_a \otimes \mathbf{I}_b) + \sigma_e^2(\mathbf{I}_a \otimes \mathbf{I}_b)\right] = (a\sigma_b^2 + \sigma_e^2)(\bar{\mathbf{J}}_a \otimes \mathbf{C}_b)$$

implies that

$$\text{SSB} \sim (a\sigma_b^2 + \sigma_e^2)\chi_{b-1}^2(\lambda_3)\,,$$

where

$$\lambda_3 = \frac{\boldsymbol{\mu}'\mathbf{A}_3\boldsymbol{\mu}}{2(a\sigma_b^2 + \sigma_e^2)} = 0$$

because

$$\mathbf{A}_3\boldsymbol{\mu} = (\bar{\mathbf{J}}_a \otimes \mathbf{C}_b)\left[(\mathbf{1}_a \otimes \mathbf{1}_b)\mu + (\mathbf{I}_a \otimes \mathbf{1}_b)\boldsymbol{\alpha}\right] = \mathbf{0}\,.$$

Independence of SSE and SSB follows from the fact that $\mathbf{A}_3\boldsymbol{\Sigma}\mathbf{A}_4 = \mathbf{0}$. Hence, the statistic,

$$F = \frac{\text{MSB}}{\text{MSE}} \sim \frac{a\sigma_b^2 + \sigma_e^2}{\sigma_e^2} F_{b-1,(a-1)(b-1)}\ ,$$

can be used to test the hypothesis $H_0 : \sigma_b^2 = 0$.

9.2 Balanced two-factor crossed design with replication

The ANOVA decomposition for a balanced two-factor crossed design with replication is as follows.

Source	Sum of Squares	$\mathbf{y}'\mathbf{A}_i\mathbf{y}$
Intercept	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r \bar{y}_{\cdots}^2$	$\mathbf{y}'(\bar{\mathbf{J}}_a \otimes \bar{\mathbf{J}}_b \otimes \bar{\mathbf{J}}_r)\mathbf{y}$
A	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$	$\mathbf{y}'(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b \otimes \bar{\mathbf{J}}_r)\mathbf{y}$
B	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$	$\mathbf{y}'(\bar{\mathbf{J}}_a \otimes \mathbf{C}_b \otimes \bar{\mathbf{J}}_r)\mathbf{y}$
AB	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$	$\mathbf{y}'(\mathbf{C}_a \otimes \mathbf{C}_b \otimes \bar{\mathbf{J}}_r)\mathbf{y}$
Error	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{ij\cdot})^2$	$\mathbf{y}'(\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{C}_r)\mathbf{y}$

Note that the replicate factor, R say, with levels $1, \ldots, r$, is *nested* in the treatment combinations. For example, replicate 1 in treatment combination (1,1) is not the same as replicate 1 in treatment combination (1,2).

Both factors fixed: If both A and B are fixed factors the statistical model that generates the ANOVA decomposition in the table above is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}\ ,$$

where $e_{ijk} \sim N(0, \sigma_e^2)$ independently, and where

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \alpha\beta_{ij} = \sum_{j=1}^{b} \alpha\beta_{ij} = 0\,.$$

This model allows for all ab treatment means to be different. The matrix form of the model is

$$\begin{aligned} \mathbf{y} &= (\mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_r)\mu + (\mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_r)\boldsymbol{\alpha} \\ &\quad + (\mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_r)\boldsymbol{\beta} + (\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_r)\boldsymbol{\alpha\beta} + \mathbf{e}\,, \end{aligned}$$

where $\mathbf{e} \sim N(\mathbf{0}_{abr}, \sigma_e^2 \mathbf{I}_{abr})$, and

$$\mathbf{1}_a'\boldsymbol{\alpha} = \mathbf{1}_b'\boldsymbol{\beta} = (\mathbf{1}_a \otimes \mathbf{I}_b)'\boldsymbol{\alpha\beta} = (\mathbf{I}_a \otimes \mathbf{1}_b)'\boldsymbol{\alpha\beta} = 0\,.$$

It can easily be verified that the $\mathbf{A}_i$ matrices in the ANOVA decomposition satisfy the conditions of Cochran's theorem and hence $\mathbf{y}\mathbf{A}_i\mathbf{y} \sim \sigma_e^2 \chi^2_{r_i}(\lambda_i)$ independently, where $r_i = \text{rank}(\mathbf{A}_i)$ and $\lambda_i = \boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}/(2\sigma_e^2)$. For example, $\lambda_5 = 0$ because $\mathbf{C}_r\mathbf{1}_r = \mathbf{0}$, but

$$\begin{aligned} \lambda_4 &= \frac{1}{2\sigma_e^2}\boldsymbol{\alpha\beta}'(\mathbf{C}_a \otimes \mathbf{C}_b \otimes \mathbf{1}_r'\bar{\mathbf{J}}_r\mathbf{1}_r)\boldsymbol{\alpha\beta} \\ &= \frac{r}{2\sigma_e^2}\sum_{i=1}^{a}\sum_{j=1}^{b} \alpha\beta_{ij}^2 \\ &= \frac{(a-1)(b-1)r\theta_{\alpha\beta}}{2\sigma_e^2}\,. \end{aligned}$$

The expected mean squares for the model terms are given in the following table.

Source	DF	EMS
Intercept	1	$\sigma_e^2 + abr\mu^2$
A	$a-1$	$\sigma_e^2 + br\theta_\alpha$
B	$b-1$	$\sigma_e^2 + ar\theta_\beta$
AB	$(a-1)(b-1)$	$\sigma_e^2 + r\theta_{\alpha\beta}$
Error	$ab(r-1)$	σ_e^2

For example,

$$\mathrm{E(MSAB)} = \frac{\sigma_e^2}{(a-1)(b-1)}\left[(a-1)(b-1)+2\lambda_4\right] = \sigma_e^2 + r\theta_{\alpha\beta}\,.$$

The hypothesis of *no interaction* is $\theta_{\alpha\beta} = 0$. This can be tested using the F-statistic,

$$F = \frac{\mathrm{MSAB}}{\mathrm{MSE}} \sim F_{(a-1)(b-1),ab(r-1)}(\lambda_4)\,,$$

which has a central F-distribution under H_0 but is stochastically larger if H_0 is false.

Randomized blocks with replication: Suppose that B is a random (e.g. blocking) factor, then the statistical model is

$$y_{ijk} = \mu + \alpha_i + b_j + \alpha b_{ij} + e_{ijk}\,,$$

where $b_j \sim N(0,\sigma_b^2)$, $\alpha b_{ij} \sim N(0,\sigma_{\alpha b}^2)$, and $e_{ijk} \sim N(\sigma_e^2)$ independently, and $\sum_{i=1}^a \alpha_i = 0$. (Some authors argue that the model should also include the constraint, $\sum_{i=1}^a \alpha b_{ij} = 0$, for all $j = 1,\ldots,b$.)

The matrix form of the model is

$$\begin{aligned}\mathbf{y} &= (\mathbf{1}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_r)\mu + (\mathbf{I}_a \otimes \mathbf{1}_b \otimes \mathbf{1}_r)\boldsymbol{\alpha} \\ &\quad + (\mathbf{1}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_r)\mathbf{b} + (\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{1}_r)\boldsymbol{\alpha}\mathbf{b} + \mathbf{e}\,,\end{aligned}$$

where $\mathbf{1}_a'\boldsymbol{\alpha} = 0$ and

$$\boldsymbol{\Sigma} = (\mathbf{J}_a \otimes \mathbf{I}_b \otimes \mathbf{J}_r)\sigma_b^2 + (\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{J}_r)\sigma_{\alpha b}^2 + (\mathbf{I}_a \otimes \mathbf{I}_b \otimes \mathbf{I}_r)\sigma_e^2\,.$$

The expected mean squares associated with this model are as follows.

Source	DF	EMS
Intercept	1	$\sigma_e^2 + r\sigma_{\alpha b}^2 + ar\sigma_b^2 + abr\mu^2$
A	$a-1$	$\sigma_e^2 + r\sigma_{\alpha b}^2 + br\theta_\alpha$
B	$b-1$	$\sigma_e^2 + r\sigma_{\alpha b}^2 + ar\sigma_b^2$
AB	$(a-1)(b-1)$	$\sigma_e^2 + r\sigma_{\alpha b}^2$
Error	$ab(r-1)$	σ_e^2

For example,

$$\mathbf{A}_2\mathbf{\Sigma} = (r\sigma_{\alpha b}^2 + \sigma_e^2)(\mathbf{C}_a \otimes \bar{\mathbf{J}}_b \otimes \bar{\mathbf{J}}_r)\,,$$

implies that

$$\mathbf{y}'\mathbf{A}_2\mathbf{y} \sim (r\sigma_{\alpha b}^2 + \sigma_e^2)\chi_{a-1}^2(\lambda_2)\,,$$

where

$$\lambda_2 = \frac{1}{2(r\sigma_{\alpha b}^2 + \sigma_e^2)}\boldsymbol{\mu}'\mathbf{A}_2\boldsymbol{\mu} = \frac{(a-1)br\theta_\alpha}{2(r\sigma_{\alpha b}^2 + \sigma_e^2)}$$

So

$$\mathrm{E(MSA)} = \frac{r\sigma_{\alpha b} + \sigma_e^2}{a-1}(a - 1 + 2\lambda_2) = \sigma_e^2 + r\sigma_{\alpha b}^2 + br\theta_\alpha$$

Thus, in this case, to test the hypothesis $H_0 : \theta_\alpha = 0$, we use the F-statistic,

$$F = \frac{\mathrm{MSA}}{\mathrm{MSAB}} \sim F_{(a-1),(a-1)(b-1)}(\lambda_2)\,.$$

The model implies that the sample mean for treatment i (level i of factor A) is given by

$$\bar{y}_{i\cdot\cdot} = \mu + \alpha_i + \bar{b}_{\cdot} + \bar{\alpha b}_{i\cdot} + \bar{e}_{i\cdot\cdot}\,.$$

Hence, the difference between two treatment means is

$$\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot} = (\alpha_i - \alpha_{i'}) + (\bar{\alpha b}_{i\cdot} - \bar{\alpha b}_{i'\cdot}) + (\bar{e}_{i\cdot\cdot} - \bar{e}_{i'\cdot\cdot})\,.$$

Thus,

$$\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}\ N\left[\alpha_i - \alpha_{i'}, \frac{2}{br}(\sigma_e^2 + r\sigma_{\alpha b}^2)\right]$$

Since an MSAB is independent of the treatment means and has a chisquared distirbution proportional to $\sigma_e^2 + r\sigma_{\alpha b}^2$, we can test the hypothesis $H_0: \alpha_i = \alpha_{i'}$ using the t-statistic,

$$T = \frac{\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}}{\hat{se}(\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot})},$$

where

$$\hat{se}(\bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}) = \sqrt{\frac{2}{br}\text{MSAB}}\,.$$

Alternatively, the *least significant difference* (LSD) to declare two means to be different (at, say, significance level α) is

$$\text{LSD} = |t_{\alpha/2}|\sqrt{\frac{2}{br}\text{MSAB}}\,.$$

10 Linear Mixed Models for Designed Experiments

10.1 Summary of general theory for balanced experiments

Suppose that the n-dimensional response vector, $\mathbf{y}$, has a multivariate normal distribution with mean, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$. Suppose, in addition, that the mean vector is determined by factors which are crossed with or nested within other factors according to an experimental design leading to a additive linear model. Further, the sum of squares explained by the ith term in the model is given by $\mathbf{y}'\mathbf{A}_i\mathbf{y}$, where $\mathbf{A}_i$ is an idempotent matrix with rank r_i such that

- $\sum_i^k \mathbf{A}_i = \mathbf{I}_n$ and $\sum_{i=1}^k r_i = n$;
- $\mathbf{A}_i \boldsymbol{\Sigma} \mathbf{A}_j = \mathbf{0}$ for all $i \neq j$; and
- $\mathbf{A}_i \boldsymbol{\Sigma} = c_i \mathbf{A}_i$ for $i = 1, \ldots, k$;

where c_i is a constant function of variance components. Then

$$\mathbf{y}' \mathbf{A}_i \mathbf{y} \sim c_i \chi^2_{r_i}(\lambda_i) \quad \text{with} \quad \lambda_i = \frac{\boldsymbol{\mu}' \mathbf{A}_i \boldsymbol{\mu}}{2c_i}$$

independently for $i = 1, \ldots, k$.

The expected mean square associated with the ith sum of squares is given by

$$\mathrm{E}(\mathrm{MSA}_i) = \mathrm{E}\left(\frac{\mathbf{y}' \mathbf{A}_i \mathbf{y}}{r_i} \right) = \frac{c_i(r_i + 2\lambda_i)}{r_i} = c_i + \frac{1}{r_i} \boldsymbol{\mu}' \mathbf{A}_i \boldsymbol{\mu} \, .$$

If $\lambda_i = 0$ then

$$F = \frac{\mathrm{MSA}_j}{\mathrm{MSA}_i} = \frac{\mathbf{y}' \mathbf{A}_j \mathbf{y}/r_j}{\mathbf{y}' \mathbf{A}_i \mathbf{y}/r_i} \sim \frac{c_j}{c_i} F_{r_j, r_i}(\lambda_j) \, .$$

10.2 Repeated measures designs

Consider experiments in which a response of interest is measured multiple times on each of n subjects. When analyzing such experiments it is natural to consider *subject* as a random factor. Note that *subject* could mean a person, a mouse, a plant, or even a field plot.

Single factor experiment with repeated measures: Let y_{ij} denote the measurement taken at time j on subject i, where $i = 1, \ldots, n$ and $j = 1, \ldots, t$. Of interest is whether the mean response is changing over time. Since the factors, *subject* and *time*, are crossed, but there is no replication at each subject/time combination, a natural model is

$$y_{ij} = \mu + s_i + \tau_j + e_{ij} \, ,$$

where $\sum_j^t \tau_j = 0$, $s_i \sim N(0, \sigma_s^2)$, and $e_{ij} \sim N(0, \sigma_e^2)$ independently. This model has the same structure as a randomized blocks experiment in which each treatment occurs exactly once in each block.

The covariance matrix for the response vector is therefore

$$\boldsymbol{\Sigma} = \sigma_s^2(\mathbf{I}_n \otimes \mathbf{J}_t) + \sigma_e^2(\mathbf{I}_n \otimes \mathbf{I}_t) = \text{blockdiag}(\sigma_s^2\mathbf{J}_t + \sigma_e^2\mathbf{I}_n)\,.$$

That is, the covariance matrix for the response vector, $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})'$ is

$$\text{var}(\mathbf{y}_i) = \sigma_s^2\mathbf{J}_t + \sigma_e^2\mathbf{I}_t\,,$$

but responses of different subjects are independent. In particular, the variance of any response is $\sigma^2 = \sigma_s^2 + \sigma_e^2$, and the covariance between any pair of responses from the same subject is σ_s^2. Thus, responses from the same subject are *equicorrelated*, with the correlation given by

$$\alpha = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}\,.$$

Two factors, with repeated measures on one factor: Suppose that the subjects are divided (at random) into g treatment groups. Let y_{ijk} denote the response of subject j in group i at time k. The group and time factors are crossed with the number of replicates at each group/time combination being equal to the number of subjects in the group. On the other hand the subject factor is nested within group. An appropriate additive model is

$$y_{ijk} = \mu + \alpha_i + s_{j(i)} + \tau_k + \alpha\tau_{ik} + e_{ijk}$$

with sum constraints

$$\sum_{i=1}^{g} \alpha_i = \sum_{i=1}^{g} \alpha\tau_{ik} = \sum_{k=1}^{t} \alpha\tau_{ik} = 0\,,$$

and distributional assumptions

$$s_{j(i)} \sim N(0, \sigma_s^2) \quad \text{and} \quad e_{ijk} \sim N(0, \sigma_e^2)$$

independently.

Using Kronecker product notation, the model is

$$\begin{aligned} \mathbf{y} &= (\mathbf{1}_g \otimes \mathbf{1}_n \otimes \mathbf{1}_t)\mu + (\mathbf{I}_g \otimes \mathbf{1}_n \otimes \mathbf{1}_t)\boldsymbol{\alpha} + (\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{1}_t)\mathbf{s} \\ &\quad + (\mathbf{1}_g \otimes \mathbf{1}_n \otimes \mathbf{I}_t)\boldsymbol{\tau} + (\mathbf{I}_g \otimes \mathbf{1}_n \otimes \mathbf{I}_t)\boldsymbol{\alpha\tau} + (\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{I}_t)\mathbf{e}\,, \end{aligned}$$

where $\mathbf{s} \sim N(\mathbf{0}_n, \sigma_s^2 \mathbf{I}_{gn})$ and $\mathbf{e} \sim N(\mathbf{0}_{gnt}, \sigma_e^2 \mathbf{I}_{gnt})$ independently. Equivalently, $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\mathbf{1}_g \otimes \mathbf{1}_n \otimes \mathbf{1}_t)\mu + (\mathbf{I}_g \otimes \mathbf{1}_n \otimes \mathbf{1}_t)\boldsymbol{\alpha} + (\mathbf{1}_g \otimes \mathbf{1}_n \otimes \mathbf{I}_t)\boldsymbol{\tau} + (\mathbf{I}_g \otimes \mathbf{1}_n \otimes \mathbf{I}_t)\boldsymbol{\alpha\tau}$$

and

$$\boldsymbol{\Sigma} = \sigma_s^2(\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{J}_t) + \sigma_e^2(\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{I}_t)\,.$$

The ANOVA decomposition and expected mean square table for this model is given by

Source	$\mathbf{y}'\mathbf{A}_i\mathbf{y}$	DF	EMS
1. Intercept	$\mathbf{y}'(\bar{\mathbf{J}}_g \otimes \bar{\mathbf{J}}_n \otimes \bar{\mathbf{J}}_t)\mathbf{y}$	1	$\sigma_e^2 + t\sigma_s^2 + gnt\mu^2$
2. Group	$\mathbf{y}'(\mathbf{C}_g \otimes \bar{\mathbf{J}}_n \otimes \bar{\mathbf{J}}_t)\mathbf{y}$	$g-1$	$\sigma_e^2 + t\sigma_s^2 + nt\theta_\alpha$
3. Subject	$\mathbf{y}'(\mathbf{I}_g \otimes \mathbf{C}_n \otimes \bar{\mathbf{J}}_t)\mathbf{y}$	$g(n-1)$	$\sigma_e^2 + t\sigma_s^2$
4. Time	$\mathbf{y}'(\bar{\mathbf{J}}_g \otimes \bar{\mathbf{J}}_n \otimes \mathbf{C}_t)\mathbf{y}$	$t-1$	$\sigma_e^2 + gn\theta_\tau$
5. Group×Time	$\mathbf{y}'(\mathbf{C}_g \otimes \bar{\mathbf{J}}_n \otimes \mathbf{C}_t)\mathbf{y}$	$(g-1)(t-1)$	$\sigma_e^2 + n\theta_{\alpha\tau}$
6. Error	$\mathbf{y}'(\mathbf{I}_g \otimes \mathbf{C}_n \otimes \mathbf{C}_t)\mathbf{y}$	$g(n-1)(t-1)$	σ_e^2

For example,

$$\begin{aligned} \mathbf{A}_2\boldsymbol{\Sigma} &= (\mathbf{C}_g \otimes \bar{\mathbf{J}}_n \otimes \bar{\mathbf{J}}_t)\left[\sigma_s^2(\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{J}_t) + \sigma_e^2(\mathbf{I}_g \otimes \mathbf{I}_n \otimes \mathbf{I}_t)\right] \\ &= (\sigma_e^2 + t\sigma_s^2)\mathbf{A}_2\,, \end{aligned}$$

and

$$\boldsymbol{\mu}'\mathbf{A}_2\boldsymbol{\mu} = nt\boldsymbol{\alpha}'\mathbf{C}_g\boldsymbol{\alpha} + n\boldsymbol{\alpha\tau}'(\mathbf{C}_g \otimes \bar{\mathbf{J}}_t)\boldsymbol{\alpha\tau} = (g-1)nt\theta_\alpha$$

imply that

$$\begin{aligned}
\mathrm{E(MSG)} &= \frac{1}{g-1}\left\{\mathrm{tr}(\mathbf{A}_2\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}_2\boldsymbol{\mu}\right\} \\
&= \frac{1}{g-1}\left\{(g-1)(\sigma_e^2 + t\sigma_s^2) + (g-1)nt\theta_\alpha\right\} \\
&= \sigma_e^2 + t\sigma_s^2 + nt\theta_\alpha
\end{aligned}$$

Thus, the significance of the group by time interaction can be assessed using the F-ratio, MSGT/MSE. However, if the data suggests no interaction, the hypothesis of no difference between the groups is assessed using the ratio MSG/MSS.

Two factors with repeated measure on both factors: Consider a pilot study involving a sample of $n = 5$ subjects and two drugs. The blood serum concentration of drug A is measure 1, 2, 3, and 6 hours after is taken by the subjects. After a washout period the same process is repeated for drug B. Let y_{ijk} denote the concentration of drug j at time k for subject i. Then, if subject is considered a random factor, the design allows for the following mixed model,

$$y_{ijk} = \mu + s_i + \delta_j + (s\delta)_{ij} + \tau_k + (s\tau)_{ik} + (\delta\tau)_{jk} + e_{ijk}\,,$$

where

$$\sum_j \delta_j = \sum_k \tau_k = \sum_j (\delta\tau)_{jk} = \sum_k (\delta\tau)_{jk} = 0$$

and

$$s_i \sim N(0, \sigma_s^2) \quad (s\delta)_{ij} \sim N(0, \sigma_{s\delta}) \quad (s\tau)_{ij} \sim N(0, \sigma_{s\tau}) \quad e_{ijk} \sim N(0, \sigma_e^2)\,.$$

```
> drug.df[1:10,]

   subject drug time    y
1        1    a    1 1.08
2        1    a    2 1.99
3        1    a    3 1.46
4        1    a    6 1.21
5        1    b    1 1.48
6        1    b    2 2.50
7        1    b    3 2.62
8        1    b    6 1.95
9        2    a    1 1.19
10       2    a    2 2.10

> lmfit=lm(y~subject+drug+subject:drug+time
+   +subject:time+drug:time,data=drug.df)
> anova(lmfit)[,1:3]

             Df Sum Sq Mean Sq
subject       4 4.4351 1.10878
drug          1 0.0497 0.04970
time          3 3.2716 1.09053
subject:drug  4 2.4365 0.60913
subject:time 12 1.2192 0.10160
drug:time     3 0.0988 0.03295
Residuals    12 1.1934 0.09945
```

For this model the expected mean square table is as follows.

Source	DF	EMS
Subject	4	$\sigma_e^2 + 2\sigma_{s\tau}^2 + 4\sigma_{s\delta}^2 + 8\sigma_s^2$
Drug	1	$\sigma_e^2 + 4\sigma_{s\delta}^2 + 20\theta_\delta$
Subject*Drug	4	$\sigma_e^2 + 4\sigma_{s\delta}^2$
Time	3	$\sigma_e^2 + 2\sigma_{s\tau}^2 + 10\theta_\tau$
Subject*Time	12	$\sigma_e^2 + 2\sigma_{s\tau}^2$
Drug*Time	3	$\sigma_e^2 + 5\theta_{\delta\tau}$
Error	12	σ_e^2

```
          F-statistic P-value

drug           0.0816  0.7800

time          10.7338  0.0010

drug*time      0.3313  0.8029
```

10.3 Split-Plot Experimental Design

10.4 Balanced Incomplete Block Design (BIBD)

Consider a randomized blocks experiment to compare t treatments. Suppose that each block can only accomodate $k < t$ treatments. This might be because of a desire to keep the blocks small and homogeneous, or because of physical constraints on the blocks. If each treatment is replicated r times, then the total number of units is tr. The number of units is also given by bk, where b is the number of blocks; that is,

$$tr = bk\,.$$

The design is said to be balanced if all pairs of treatments occur together in a block the same number of times, λ say. Then the total number of treatments occurring in the same block as treatment i is given by

$$\lambda(t-1) = r(k-1)\,.$$

For example, suppose that $t = 4$ and $k = 2$. Then $4r = 2b$ implies that $b = 2r$; that is, the number of blocks must be twice the number of replicates. Moreover, $3\lambda = r$ implies that the number of replicates is a multiple of three, and so the minimal design has $b = 6$ blocks.

Fixed blocks model: Let T_j denote the set of treatments that occur in block j, and let y_{ij} be the response of treatment $i \in T_j$ in block j. Then, a model with fixed block effects is

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}\,,$$

where

$$\sum_{i=1}^{t} \alpha_i = \sum_{j=1}^{b} \beta_j = 0 \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2)\,.$$

If $\mathbf{y}_j$ denotes the set of k responses in block j then

$$\mathbf{y}_j = \mathbf{1}_k\mu + \boldsymbol{\alpha}_j + \mathbf{1}_k\beta_j + \mathbf{e}_j\,,$$

where $\boldsymbol{\alpha}_j$ is the vector of treatment effects for treatments occurring in block j.

Let $\mathbf{H}_k$ denote a rank k orthogonal matrix with first column equal to $\mathbf{h}_1 = \mathbf{1}_k'/\sqrt{k}$. Note that this implies the remaining columns are orthonormal contrasts. The Helmert matrix is a special case of this kind of matrix. Consider the transformation $\mathbf{y}* = (\mathbf{H}_k \otimes \mathbf{I}_b)\mathbf{y}$, or $\mathbf{y}_j^* = \mathbf{H}_k'\mathbf{y}_j$, for $j = 1, \ldots, b$. The transformation is invertable since $\mathbf{y} = (\mathbf{H}_k \otimes \mathbf{I}_b)\mathbf{y}^*$ and so the information in $\mathbf{y}^*$ about the model parameters is equivalent to that in $\mathbf{y}$. The individual transformed observations are $y_{ij}^* = \mathbf{h}_i'\mathbf{y}_j$, for $i = 1, \ldots, k$, $j = 1, \ldots, b$. Notice that the corresponding transformed errors, e_{ij}^*, are distributed as $N(0, \sigma_e^2)$ independently.

In particular, we have

$$y^*_{1j} = \mathbf{h}'_1\mathbf{y}_j = \sqrt{k}\mu + \mathbf{h}'_1\boldsymbol{\alpha}_j + \sqrt{k}\beta_j + e^*_{1j}$$

is the model for *inter-block* variation. This data contains no information about treatment differences because the treatment effects are confounded with the fixed block effects. (There is one observation per block.)

On the other hand $y^*_{2j}, \ldots, y^*_{kj}$ represent contrasts between treatments within block j. The *intra-block* model is

$$y^*_{ij} = \mathbf{h}'_i\mathbf{y}_j = \mathbf{h}'_i\boldsymbol{\alpha}_j + e^*_{ij}\,,$$

for $i = 2, \ldots, k$ and $j = 1, \ldots, b$. Thus, this data carries no information about the block effects, and all the information about the treatment effects.

Random blocks model: If block is considered a random factor, then the model is

$$y_{ij} = \mu + \alpha_i + b_j + e_{ij}\,,$$

where

$$\sum_{i=1}^{t} \alpha_i = 0 \quad b_j \sim N(0, \sigma^2_b) \quad \text{and} \quad e_{ij} \sim N(0, \sigma^2_e)\,.$$

In this case the inter-block model,

$$y^*_{1j} = \sqrt{k}\mu + \mathbf{h}'_1\boldsymbol{\alpha}_j + (\sqrt{k}b_j + e^*_{1j})\,, \quad j = 1, \ldots, b\,,$$

does contain information about the treatment effects. Since the intra-block model is the same as in the fixed block case, it is possible to get improved estimates of the treatment effects by combining the estimates from the two models. This was referred to historically as *recovery of inter-block information.*

A “complication” with the random blocks model is that it does not lead to exact tests for comparing treatments. However, there are several methods for constructing approximate tests which boil down to different ways of approximating the denominator degrees of freedom for F-tests. A general approach to constructing approximate F-statistics is described in the next section.

Comparison of dishwashing detergents: John (1961) describes an experiment to compare dishwashing detergents using a BIBD with $t = 9$ treatments (detergents) and $k = 3$ treatments per block (washing sessions). (The limiting factor was the number of basins/operators to wash plates in). The response was the number of plates washed before the foam dissappeared. The minimal BIBD in this case requires $b = 12$ blocks resulting in $r = 4$ replicates of each treatment, and each pair of treatments occuring exactly once in together in a block.

```
> dw=read.csv("../data/dishwash.csv")
> head(dw)

  session detergent plates
1       1         1     19
2       1         2     17
3       1         3     11
4       2         4      6
5       2         5     26
6       2         6     23

> block=factor(dw$session); trt=factor(dw$detergent); y=dw$plates
> lmfit=lm(y~block+trt); anova(lmfit)

Analysis of Variance Table
```

```
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
block     11  412.75  37.523  45.533 6.028e-10
trt        8 1086.81 135.852 164.854 6.809e-14
Residuals 16   13.19   0.824
```

```
> lmefit=lmer(y~trt+(1|block)); anova(lmefit)
```

```
Analysis of Variance Table
    Df Sum Sq Mean Sq F value
trt  8 1419.3  177.42  220.57
```

The F-statistics produced using the default settings in SAS Proc Mixed and JMP are 220.57 and 195.77. SAS reports an error degrees of freedom of 16 (based on the containment method), whereas JMP reports a value of 24.95 (based on the Kenward-Rodgers method). All three packages, R lmer, SAS Proc Mixed and JMP report exactly the same REML estimates of variance components.

11 Unbalanced Data

The general form of a variance components model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{q} \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}\,,$$

where $\mathbf{b}_i \sim N(\mathbf{0}_m, \sigma_i^2 \mathbf{I}_{m_i})$, for $i = 1, \ldots, q$, and $\mathbf{e} \sim N(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n)$ independently. Here σ_0^2 denotes the residual/error variance and σ_i^2, $i = 1, \ldots, q$ are the variances of the random effects.

The variance components model described above can be rewritten in its marginal form, $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\Sigma} = \sum_{i=0}^{q} \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i' \,.$$

Let $\mathbf{m}$ be a vector containing the indices of missing responses, and let $\mathbf{M}$ denote the identity matrix of rank n with rows in $\mathbf{m}$ removed. Then $\mathbf{y}^* = \mathbf{M}\mathbf{y}$ is the observed response vector, and $\mathbf{y}^* \sim N(\mathbf{X}^*\boldsymbol{\beta}, \boldsymbol{\Sigma}^*)$, where $\mathbf{X}^* = \mathbf{M}\mathbf{X}$, and $\boldsymbol{\Sigma}^* = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}'$.

11.1 Generalized Least Squares

The log-likelihood function based on the response vector $\mathbf{y}$ is given by

$$\log L = -\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\,.$$

Maximizing with respect to $\boldsymbol{\beta}$ results in the generalized least squares (GLS) estimator,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\,.$$

which reduces to $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in the MLR case, $\boldsymbol{\Sigma} = \sigma_0^2\mathbf{I}_n$. (This is also true for the balanced (complete) designs discussed in the previous section.) Note that, since the GLS esimator is linear in $\mathbf{y}$,

$$\hat{\boldsymbol{\beta}} \sim N\left[\boldsymbol{\beta}, (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\right]\,.$$

Most statistical packages allow that variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ to be extracted from the model fit.

Consider a general linear hypothesis about the fixed effects of the form $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. For example, in an RCBD experiment involving four treatments we may wish to test the hypothesis that the treatment means are all equal. With the mean parameterization (i.e. μ_i equal to the expected response for treatment i), this is accomplished with

$$\mathbf{L} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} .$$

With the standard sum constraints parameterization we need

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .$$

so that $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ is equivalent to $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

Notice that

$$\mathbf{L}\hat{\boldsymbol{\beta}} \sim N\left[\mathbf{L}\boldsymbol{\beta}, \mathbf{L}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{L}'\right] .$$

Hence, it follows that, if $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$,

$$\hat{\boldsymbol{\beta}}'\mathbf{L}'\left[\mathbf{L}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{L}'\right]^{-1}\mathbf{L}\hat{\boldsymbol{\beta}} \sim \chi^2_r$$

where r is the number of rows in (more precisely, the rank of) $\mathbf{L}$. This result motivates the *generalized F-statistic* for testing $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, defined by

$$F = \hat{\boldsymbol{\beta}}'\mathbf{L}'\left[\mathbf{L}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{L}'\right]^{-1}\mathbf{L}\hat{\boldsymbol{\beta}}/r \, ,$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimate of $\boldsymbol{\Sigma}$ (e.g. by ML, REML or the method of moments). In the MLR case, $\boldsymbol{\Sigma} = \sigma_0^2\mathbf{I}_n$, the generalized F-statistic reduces to

$$F = \frac{\hat{\boldsymbol{\beta}}'\mathbf{L}'\left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]^{-1}\mathbf{L}\hat{\boldsymbol{\beta}}/r}{\hat{\sigma}_0^2}$$

where $\hat{\sigma}_0^2 =$MSE. In this setting, the test-statistic has an exact F-distribution if $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$; specicically, $F \sim F_{r,n-p-1}$, where $rank(\mathbf{X}) = p + 1$. This appears to be the rationale for the defining the sum of squares associated with an effect $\mathbf{L}\hat{\boldsymbol{\beta}}$ (in *lmer* anova ouput) as

$$\hat{\sigma}_0^2 \times \hat{\boldsymbol{\beta}}'\mathbf{L}'\left[\mathbf{L}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{L}'\right]^{-1}\mathbf{L}\hat{\boldsymbol{\beta}} \, .$$

11.2 Satterthwaite's Approximation

Recall that the expected mean squares associated with a fixed factor with sum of squares, $\mathbf{y}'\mathbf{A}\mathbf{y}$, has the general form

$$\mathrm{E}(MS) = c + \frac{\delta}{r}\,,$$

where $\delta = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$, $r = \mathrm{rank}(\mathbf{A})$ and c is a linear combination of variance components; i.e.

$$c = \sum_{i=0}^{q} k_i\sigma_i^2\,.$$

An F-statistic for testing $\delta = 0$ is available if there is a random component with EMS equal to c.

Suppose that no such random component is available (perhaps because the data is unbalanced), but there exists a linear combination of mean squares associated with the random factors,

$$W = \sum_{i=0}^{q} l_i MS_i\,,$$

with the property that $\mathrm{E}(W) = c$. Hence the ratio, MS/W, does not depend on c.

The sum of squares, SS_i associated with the ith random factor has the property, $\mathrm{SS}_i \sim c_i\chi^2_{r_i}(0)$, independently. Hence the statistic, W, is a linear combination of independent central chisquared variables, and so the ratio, MS/W, does not have an F-distribution.

Satterthwaite (1946) proposed approximating the distribution of W by a scaled chisquare variable with the same first two moments. In particular,

$$\mathrm{E}(MS_i) = c_i \quad \text{and} \quad \mathrm{var}(\mathrm{MS}_i) = 2c_i^2/r_i\,,$$

together imply

$$\mathrm{E}(W) = \sum_{i=0}^{q} l_i c_i = \sum_{i=0}^{q} k_i \sigma_i^2 = c$$

and

$$\mathrm{var}(W) = 2 \sum_{i=0}^{q} \frac{l_i^2 c_i^2}{r_i} \, .$$

The approximation is therefore

$$W \sim \text{approx} \quad c\chi_f^2(0)/r^*$$

where, by matching variances,

$$2\frac{c^2}{r^*} = 2\sum_{i=0}^{q} \frac{l_i^2 c_i^2}{r_i} \, ,$$

or equivalently

$$r^* = \frac{c^2}{\sum_{i=0}^{q} l_i^2 c_i^2 / r_i} \, .$$

In practice, since the c_i's and c are unknown (they are linear combinations of variance components), they must be replaced by unbiased estimates, MS_i and W respectively. Finally, an approximate F-test of the hypothesis $\delta = 0$ is formed by referring the ratio, MS/W, to an F-distribution with r and r^* degrees of freedom, where

$$r^* = \frac{W^2}{\sum_{i=0}^{q} l_i^2 MS_i^2 / r_i} \, .$$

References

Cook, R. D. & Weisberg, S. (1999). *Applied Regression including Computing and Graphics.* Wiley Inter-Science. ISBN: 0-471-31711-X.

John, P. W. M. (1961). An application of balanced incomplete block designs. *Technometrics* 3 51–54.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2 110–114.

Snedecor, G. W. & Cochran, G. C. (1967). *Statistical Methods.* Ames, Iowa: Iowa State University Press, 6th ed.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statist. Soc. B* 58 267–288.