Appendix A
Vector Spaces

This appendix reviews some of the basic definitions and properties of vector spaces. It is presumed that, with the possible exception of Theorem A.14, all of the material presented here is familiar to the reader.

Definition A.1. A set $\mathscr{M}$ is a *vector space* if, for any $x, y, z \in \mathscr{M}$ and scalars α, β, operations of vector addition and scalar multiplication are defined such that:

(1) $(x+y)+z = x+(y+z)$.
(2) $x+y = y+x$.
(3) There exists a vector $0 \in \mathscr{M}$ such that $x+0 = x = 0+x$ for any $x \in \mathscr{M}$.
(4) For any $x \in \mathscr{M}$, there exists $y \equiv -x$ such that $x+y = 0 = y+x$.
(5) $\alpha(x+y) = \alpha x + \alpha y$.
(6) $(\alpha+\beta)x = \alpha x + \beta x$.
(7) $(\alpha\beta)x = \alpha(\beta x)$.
(8) There exists a scalar ξ such that $\xi x = x$. (Typically, $\xi = 1$.)

In nearly all of our applications, we assume $\mathscr{M} \subset \mathbf{R}^n$.

Definition A.2. Let $\mathscr{M}$ be a vector space, and let $\mathscr{N}$ be a set with $\mathscr{N} \subset \mathscr{M}$. $\mathscr{N}$ is a *subspace* of $\mathscr{M}$ if and only if $\mathscr{N}$ is a vector space.

Vectors in $\mathbf{R}^n$ will be considered as $n \times 1$ matrices. The 0 vector referred to in Definition A.1 is just an $n \times 1$ matrix of zeros. Think of vectors in three dimensions as $(x, y, z)'$, where w' denotes the *transpose* of a matrix w. The subspace consisting of the z axis is

$$\left\{ \begin{pmatrix} 0 \\ 0 \\ z \end{pmatrix} \middle| \, z \in \mathbf{R} \right\}.$$

R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer Texts in Statistics, DOI 10.1007/978-1-4419-9816-3,

The subspace consisting of the x, y plane is

$$\left\{ \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \,\middle|\, x, y \in \mathbf{R} \right\}.$$

The subspace consisting of the plane that is perpendicular to the line $x = y$ in the x, y plane is

$$\left\{ \begin{pmatrix} x \\ -x \\ z \end{pmatrix} \,\middle|\, x, z \in \mathbf{R} \right\}.$$

Theorem A.3. Let $\mathscr{M}$ be a vector space, and let $\mathscr{N}$ be a nonempty subset of $\mathscr{M}$. If $\mathscr{N}$ is closed under vector addition and scalar multiplication, then $\mathscr{N}$ is a subspace of $\mathscr{M}$.

Theorem A.4. Let $\mathscr{M}$ be a vector space, and let $x_1, \ldots, x_r$ be in $\mathscr{M}$. The set of all linear combinations of $x_1, \ldots, x_r$, i.e., $\{v \,|\, v = \alpha_1 x_1 + \cdots + \alpha_r x_r, \alpha_i \in \mathbf{R}\}$, is a subspace of $\mathscr{M}$.

Definition A.5. The set of all linear combinations of $x_1, \ldots, x_r$ is called the *space spanned by* $x_1, \ldots, x_r$. If $\mathscr{N}$ is a subspace of $\mathscr{M}$, and $\mathscr{N}$ equals the space spanned by $x_1, \ldots, x_r$, then $\{x_1, \ldots, x_r\}$ is called a *spanning set* for $\mathscr{N}$.

For example, the space spanned by the vectors

$$x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

consists of all vectors of the form $(a, b, b)'$, where a and b are any real numbers.

Let A be an $n \times p$ matrix. Each column of A is a vector in $\mathbf{R}^n$. The space spanned by the columns of A is called the *column space* of A and written $C(A)$. (Some people refer to $C(A)$ as the *range space* of A and write it $R(A)$.) If B is an $n \times r$ matrix, then $C(A, B)$ is the space spanned by the $p + r$ columns of A and B.

Definition A.6. Let $x_1, \ldots, x_r$ be vectors in $\mathscr{M}$. If there exist scalars $\alpha_1, \ldots, \alpha_r$ not all zero so that $\sum \alpha_i x_i = 0$, then $x_1, \ldots, x_r$ are *linearly dependent*. If such α_is do not exist, $x_1, \ldots, x_r$ are *linearly independent*.

Definition A.7. If $\mathscr{N}$ is a subspace of $\mathscr{M}$ and if $\{x_1, \ldots, x_r\}$ is a linearly independent spanning set for $\mathscr{N}$, then $\{x_1, \ldots, x_r\}$ is called a *basis* for $\mathscr{N}$.

Theorem A.8. If $\mathscr{N}$ is a subspace of $\mathscr{M}$, all bases for $\mathscr{N}$ have the same number of vectors.

Theorem A.9. If $v_1,\ldots,v_r$ is a basis for $\mathscr{N}$, and $x\in\mathscr{N}$, then the characterization $x=\sum_{i=1}^r \alpha_i v_i$ is unique.

PROOF. Suppose $x=\sum_{i=1}^r \alpha_i v_i$ and $x=\sum_{i=1}^r \beta_i v_i$. Then $0=\sum_{i=1}^r(\alpha_i-\beta_i)v_i$. Since the vectors v_i are linearly independent, $\alpha_i-\beta_i=0$ for all i. □

Definition A.10. The *rank* of a subspace $\mathscr{N}$ is the number of elements in a basis for $\mathscr{N}$. The rank is written $r(\mathscr{N})$. If A is a matrix, the rank of $C(A)$ is called the rank of A and is written $r(A)$.

The vectors

$$x_1=\begin{pmatrix}1\\1\\1\end{pmatrix},\quad x_2=\begin{pmatrix}1\\0\\0\end{pmatrix},\quad x_3=\begin{pmatrix}2\\3\\3\end{pmatrix}$$

are linearly dependent because $0=3x_1-x_2-x_3$. Any two of x_1,x_2,x_3 form a basis for the space of vectors with the form $(a,b,b)'$. This space has rank 2.

Definition A.11. The (Euclidean) *inner product* between two vectors x and y in $\mathbf{R}^n$ is $x'y$. Two vectors x and y are *orthogonal* (written $x\perp y$) if $x'y=0$. Two subspaces $\mathscr{N}_1$ and $\mathscr{N}_2$ are orthogonal if $x\in\mathscr{N}_1$ and $y\in\mathscr{N}_2$ imply that $x'y=0$. $\{x_1,\ldots,x_r\}$ is an *orthogonal basis* for a space $\mathscr{N}$ if $\{x_1,\ldots,x_r\}$ is a basis for $\mathscr{N}$ and for $i\neq j$, $x_i'x_j=0$. $\{x_1,\ldots,x_r\}$ is an *orthonormal basis* for $\mathscr{N}$ if $\{x_1,\ldots,x_r\}$ is an orthogonal basis and $x_i'x_i=1$ for $i=1,\ldots,r$. *The terms orthogonal and perpendicular are used interchangeably*. The *length* of a vector x is $\|x\|\equiv\sqrt{x'x}$. The *distance* between two vectors x and y is the length of their difference, i.e., $\|x-y\|$.

The lengths of the vectors given earlier are

$$\|x_1\|=\sqrt{1^2+1^2+1^2}=\sqrt{3},\quad \|x_2\|=1,\quad \|x_3\|=\sqrt{22}\doteq 4.7.$$

Also, if $x=(2,1)'$, its length is $\|x\|=\sqrt{2^2+1^2}=\sqrt{5}$. If $y=(3,2)'$, the distance between x and y is the length of $x-y=(2,1)'-(3,2)'=(-1,-1)'$, which is $\|x-y\|=\sqrt{(-1)^2+(-1)^2}=\sqrt{2}$.

Just prior to Section B.4 and in Sections 2.7 and 6.3 we discuss more general versions of the concepts of inner product and length. In particular, a more general version of Definition A.11 is given in Subsection 6.3.5. The remaining results and definitions in this appendix are easily extended to general inner products.

Our emphasis on orthogonality and our need to find orthogonal projection matrices make both the following theorem and its proof fundamental tools in linear model theory:

Theorem A.12. *The Gram–Schmidt Theorem.*
Let $\mathcal{N}$ be a space with basis $\{x_1,\ldots,x_r\}$. There exists an orthonormal basis for $\mathcal{N}$, say $\{y_1,\ldots,y_r\}$, with y_s in the space spanned by $x_1,\ldots,x_s$, $s=1,\ldots,r$.

PROOF. Define the y_is inductively:

$$\begin{aligned} y_1 &= x_1 \big/ \sqrt{x_1' x_1}\,, \\ w_s &= x_s - \sum_{i=1}^{s-1} (x_s' y_i) y_i\,, \\ y_s &= w_s \big/ \sqrt{w_s' w_s}\,. \end{aligned}$$

See Exercise A.1. □

The vectors

$$x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

are a basis for the space of vectors with the form $(a,b,b)'$. To orthonormalize this basis, take $y_1 = x_1/\sqrt{3}$. Then take

$$w_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{\sqrt{3}} \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix} = \begin{pmatrix} 2/3 \\ -1/3 \\ -1/3 \end{pmatrix}.$$

Finally, normalize w_2 to give

$$y_2 = w_2 \Big/ \sqrt{6/9} = (2/\sqrt{6}, -1/\sqrt{6}, -1/\sqrt{6})'.$$

Note that another orthonormal basis for this space consists of the vectors

$$z_1 = \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \quad z_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Definition A.13. For $\mathcal{N}$ a subspace of $\mathcal{M}$, let $\mathcal{N}_{\mathcal{M}}^{\perp} \equiv \{y \in \mathcal{M} \,|\, y \perp \mathcal{N}\}$. $\mathcal{N}_{\mathcal{M}}^{\perp}$ is called the *orthogonal complement* of $\mathcal{N}$ with respect to $\mathcal{M}$. If $\mathcal{M}$ is taken as $\mathbf{R}^n$, then $\mathcal{N}^{\perp} \equiv \mathcal{N}_{\mathcal{M}}^{\perp}$ is simply referred to as the orthogonal complement of $\mathcal{N}$.

Theorem A.14. Let $\mathcal{M}$ be a vector space, and let $\mathcal{N}$ be a subspace of $\mathcal{M}$. The orthogonal complement of $\mathcal{N}$ with respect to $\mathcal{M}$ is a subspace of $\mathcal{M}$; and if $x \in \mathcal{M}$, x can be written uniquely as $x = x_0 + x_1$ with $x_0 \in \mathcal{N}$ and $x_1 \in \mathcal{N}_{\mathcal{M}}^{\perp}$. The ranks of these spaces satisfy the relation $r(\mathcal{M}) = r(\mathcal{N}) + r(\mathcal{N}_{\mathcal{M}}^{\perp})$.

For example, let $\mathscr{M} = \mathbf{R}^3$ and let $\mathscr{N}$ be the space of vectors with the form $(a,b,b)'$. It is not difficult to see that the orthogonal complement of $\mathscr{N}$ consists of vectors of the form $(0,c,-c)'$. Any vector $(x,y,z)'$ can be written uniquely as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ (y+z)/2 \\ (y+z)/2 \end{pmatrix} + \begin{pmatrix} 0 \\ (y-z)/2 \\ -(y-z)/2 \end{pmatrix}.$$

The space of vectors with form $(a,b,b)'$ has rank 2, and the space $(0,c,-c)'$ has rank 1.

For additional examples, let

$$X_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

In this case,

$$C(X_0)^\perp = C\left(\begin{bmatrix} -1 & 1 \\ 0 & -2 \\ 1 & 1 \end{bmatrix}\right), \quad C(X_0)^\perp_{C(X)} = C\left(\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}\right),$$

and

$$C(X)^\perp = C\left(\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}\right).$$

PROOF OF THEOREM A.14. It is easily seen that $\mathscr{N}^\perp_{\mathscr{M}}$ is a subspace by checking Theorem A.3. Let $r(\mathscr{M}) = n$ and $r(\mathscr{N}) = r$. Let $v_1, \ldots, v_r$ be a basis for $\mathscr{N}$ and extend this with $w_1, \ldots, w_{n-r}$ to a basis for $\mathscr{M}$. Apply Gram–Schmidt to get $v_1^*, \ldots, v_r^*$, $w_1^*, \ldots, w_{n-r}^*$ an orthonormal basis for $\mathscr{M}$ with $v_1^*, \ldots, v_r^*$ an orthonormal basis for $\mathscr{N}$.

If $x \in \mathscr{M}$, then

$$x = \sum_{i=1}^{r} \alpha_i v_i^* + \sum_{j=1}^{n-r} \beta_j w_j^*.$$

Let $x_0 = \sum_{i=1}^r \alpha_i v_i^*$ and $x_1 = \sum_{j=1}^{n-r} \beta_j w_j^*$. Then $x_0 \in \mathscr{N}$, $x_1 \in \mathscr{N}^\perp_{\mathscr{M}}$, and $x = x_0 + x_1$.

To establish the uniqueness of the representation and the rank relationship, we need to establish that $\{w_1^*, \ldots, w_{n-r}^*\}$ is a basis for $\mathscr{N}^\perp_{\mathscr{M}}$. Since, by construction, the w_j^*s are linearly independent and $w_j^* \in \mathscr{N}^\perp_{\mathscr{M}}$, $j = 1, \ldots, n-r$, it suffices to show that $\{w_1^*, \ldots, w_{n-r}^*\}$ is a spanning set for $\mathscr{N}^\perp_{\mathscr{M}}$. If $x \in \mathscr{N}^\perp_{\mathscr{M}}$, write

$$x = \sum_{i=1}^{r} \alpha_i v_i^* + \sum_{j=1}^{n-r} \beta_j w_j^*.$$

However, since $x \in \mathscr{N}^\perp_{\mathscr{M}}$ and $v_k^* \in \mathscr{N}$ for $k = 1, \ldots, r$,

$$
\begin{aligned}
0 = x'v_k^* &= \left(\sum_{i=1}^{r} \alpha_i v_i^* + \sum_{j=1}^{n-r} \beta_j w_j^*\right)' v_k^* \\
&= \sum_{i=1}^{r} \alpha_i v_i^{*\prime} v_k^* + \sum_{j=1}^{n-r} \beta_j w_j^{*\prime} v_k^* \\
&= \alpha_k v_k^{*\prime} v_k^* = \alpha_k
\end{aligned}
$$

for $k = 1, \ldots, r$. Thus $x = \sum_{j=1}^{n-r} \beta_j w_j^*$, implying that $\{w_1^*, \ldots, w_{n-r}^*\}$ is a spanning set and a basis for $\mathscr{N}_{\mathscr{M}}^{\perp}$.

To establish uniqueness, let $x = y_0 + y_1$ with $y_0 \in \mathscr{N}$ and $y_1 \in \mathscr{N}_{\mathscr{M}}^{\perp}$. Then $y_0 = \sum_{i=1}^{r} \gamma_i v_i^*$ and $y_1 = \sum_{j=1}^{n-r} \delta_j w_j^*$; so $x = \sum_{i=1}^{r} \gamma_i v_i^* + \sum_{j=1}^{n-r} \delta_j w_j^*$. By the uniqueness of the representation of x under any basis, $\gamma_i = \alpha_i$ and $\beta_j = \delta_j$ for all i and j; thus $x_0 = y_0$ and $x_1 = y_1$.

Since a basis has been found for each of $\mathscr{M}$, $\mathscr{N}$, and $\mathscr{N}_{\mathscr{M}}^{\perp}$, we have $r(\mathscr{M}) = n$, $r(\mathscr{N}) = r$, and $r(\mathscr{N}_{\mathscr{M}}^{\perp}) = n - r$. Thus, $r(\mathscr{M}) = r(\mathscr{N}) + r(\mathscr{N}_{\mathscr{M}}^{\perp})$. □

Definition A.15. Let $\mathscr{N}_1$ and $\mathscr{N}_2$ be vector subspaces. Then the sum of $\mathscr{N}_1$ and $\mathscr{N}_2$ is $\mathscr{N}_1 + \mathscr{N}_2 = \{x | x = x_1 + x_2, x_1 \in \mathscr{N}_1, x_2 \in \mathscr{N}_2\}$.

Theorem A.16. $\mathscr{N}_1 + \mathscr{N}_2$ is a vector space and $C(A, B) = C(A) + C(B)$.

Exercises

Exercise A.1 Give a detailed proof of the Gram–Schmidt theorem.

Questions A.2 through A.13 involve the following matrices:

$$
A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad
B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad
D = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 2 & 5 \\ 0 & 0 \end{bmatrix}, \quad
E = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 2 & 7 \\ 0 & 0 \end{bmatrix},
$$

$$
F = \begin{bmatrix} 1 & 5 & 6 \\ 1 & 5 & 6 \\ 0 & 7 & 2 \\ 0 & 0 & 9 \end{bmatrix}, \quad
G = \begin{bmatrix} 1 & 0 & 5 & 2 \\ 1 & 0 & 5 & 2 \\ 2 & 5 & 7 & 9 \\ 0 & 0 & 0 & 3 \end{bmatrix}, \quad
H = \begin{bmatrix} 1 & 0 & 2 & 2 & 6 \\ 1 & 0 & 2 & 2 & 6 \\ 7 & 9 & 3 & 9 & -1 \\ 0 & 0 & 0 & 3 & -7 \end{bmatrix},
$$

$$K = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad L = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad N = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

Exercise A.2 Is the space spanned by the columns of A the same as the space spanned by the columns of B? How about the spaces spanned by the columns of K, L, F, D, and G?

Exercise A.3 Give a matrix whose column space contains $C(A)$.

Exercise A.4 Give two matrices whose column spaces contain $C(B)$.

Exercise A.5 Which of the following equalities are valid: $C(A) = C(A,D)$, $C(D) = C(A,B)$, $C(A,N) = C(A)$, $C(N) = C(A)$, $C(A) = C(F)$, $C(A) = C(G)$, $C(A) = C(H)$, $C(A) = C(D)$?

Exercise A.6 Which of the following matrices have linearly independent columns: A, B, D, N, F, H, G?

Exercise A.7 Give a basis for the space spanned by the columns of each of the following matrices: A, B, D, N, F, H, G.

Exercise A.8 Give the ranks of A, B, D, E, F, G, H, K, L, N.

Exercise A.9 Which of the following matrices have columns that are mutually orthogonal: B, A, D?

Exercise A.10 Give an orthogonal basis for the space spanned by the columns of each of the following matrices: A, D, N, K, H, G.

Exercise A.11 Find $C(A)^{\perp}$ and $C(B)^{\perp}$ (with respect to $\mathbf{R}^4$).

Exercise A.12 Find two linearly independent vectors in the orthogonal complement of $C(D)$ (with respect to $\mathbf{R}^4$).

Exercise A.13 Find a vector in the orthogonal complement of $C(D)$ with respect to $C(A)$.

Exercise A.14 Find an orthogonal basis for the space spanned by the columns of

$$X = \begin{bmatrix} 1 & 1 & 4 \\ 1 & 2 & 1 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 1 \\ 1 & 6 & 4 \end{bmatrix}.$$

Exercise A.15 For X as above, find two linearly independent vectors in the orthogonal complement of $C(X)$ (with respect to $\mathbf{R}^6$).

Exercise A.16 Let X be an $n \times p$ matrix. Prove or disprove the following statement: Every vector in $\mathbf{R}^n$ is in either $C(X)$ or $C(X)^\perp$ or both.

Exercise A.17 For any matrix A, prove that $C(A)$ and the null space of A' are orthogonal complements. Note: The null space is defined in Definition B.11.

Appendix B
Matrix Results

This appendix reviews standard ideas in matrix theory with emphasis given to important results that are less commonly taught in a junior/senior level linear algebra course. The appendix begins with basic definitions and results. A section devoted to eigenvalues and their applications follows. This section contains a number of standard definitions, but it also contains a number of very specific results that are unlikely to be familiar to people with only an undergraduate background in linear algebra. The third section is devoted to an intense (brief but detailed) examination of projections and their properties. The appendix closes with some miscellaneous results, some results on Kronecker products and Vec operators, and an introduction to tensors.

B.1 Basic Ideas

Definition B.1. Any matrix with the same number of rows and columns is called a *square matrix*.

Definition B.2. Let $A = [a_{ij}]$ be a matrix. The *transpose* of A, written A', is the matrix $A' = [b_{ij}]$, where $b_{ij} = a_{ji}$.

Definition B.3. If $A = A'$, then A is called *symmetric*. Note that only square matrices can be symmetric.

Definition B.4. If $A = [a_{ij}]$ is a square matrix and $a_{ij} = 0$ for $i \neq j$, then A is a *diagonal matrix*. If $\lambda_1, \ldots, \lambda_n$ are scalars, then $D(\lambda_j)$ and $\mathrm{Diag}(\lambda_j)$ are used to indicate an $n \times n$ matrix $D = [d_{ij}]$ with $d_{ij} = 0$, $i \neq j$, and $d_{ii} = \lambda_i$. If $\lambda \equiv (\lambda_1, \ldots, \lambda_n)'$, then

$D(\lambda) \equiv D(\lambda_j)$. A diagonal matrix with all 1s on the diagonal is called an *identity matrix* and is denoted I. Occasionally, I_n is used to denote an $n \times n$ identity matrix.

If $A = [a_{ij}]$ is $n \times p$ and $B = [b_{ij}]$ is $n \times q$, we can write an $n \times (p+q)$ matrix $C = [A,B]$, where $c_{ij} = a_{ij}$, $i = 1,\ldots,n$, $j = 1,\ldots,p$, and $c_{ij} = b_{i,j-p}$, $i = 1,\ldots,n$, $j = p+1,\ldots,p+q$. This notation can be extended in obvious ways, e.g., $C' = \begin{bmatrix} A' \\ B' \end{bmatrix}$.

Definition B.5. Let $A = [a_{ij}]$ be an $r \times c$ matrix and $B = [b_{ij}]$ be an $s \times d$ matrix. The *Kronecker product* of A and B, written $A \otimes B$, is an $r \times c$ matrix of $s \times d$ matrices. The matrix in the ith row and jth column is $a_{ij}B$. In total, $A \otimes B$ is an $rs \times cd$ matrix.

Definition B.6. Let A be an $r \times c$ matrix. Write $A = [A_1, A_2, \ldots, A_c]$, where A_i is the ith column of A. Then the *Vec* operator stacks the columns of A into an $rc \times 1$ vector; thus,

$$[\mathrm{Vec}(A)]' = [A_1', A_2', \ldots, A_c'].$$

EXAMPLE B.7.

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 3 \\ 0 & 4 \end{bmatrix},$$

$$A \otimes B = \begin{bmatrix} 1\begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix} & 4\begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix} \\ 2\begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix} & 5\begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix} \end{bmatrix} = \begin{bmatrix} 1 & 3 & 4 & 12 \\ 0 & 4 & 0 & 16 \\ 2 & 6 & 5 & 15 \\ 0 & 8 & 0 & 20 \end{bmatrix},$$

$$\mathrm{Vec}(A) = [1,2,4,5]'.$$

Definition B.8. Let A be an $n \times n$ matrix. A is *nonsingular* if there exists a matrix A^{-1} such that $A^{-1}A = I = AA^{-1}$. If no such matrix exists, then A is singular. If A^{-1} exists, it is called the *inverse* of A.

Theorem B.9. An $n \times n$ matrix A is nonsingular if and only if $r(A) = n$, i.e., the columns of A form a basis for $\mathbf{R}^n$.

Corollary B.10. $A_{n\times n}$ is singular if and only if there exists $x \neq 0$ such that $Ax = 0$.

For any matrix A, the set of all x such that $Ax = 0$ is easily seen to be a vector space.

Definition B.11. The set of all x such that $Ax = 0$ is called the *null space* of A.

Theorem B.12. If A is $n \times n$ and $r(A) = r$, then the null space of A has rank $n - r$.

B.2 Eigenvalues and Related Results

The material in this section deals with eigenvalues and eigenvectors either in the statements of the results or in their proofs. Again, this is meant to be a brief review of important concepts; but, in addition, there are a number of specific results that may be unfamiliar.

Definition B.13. The scalar λ is an *eigenvalue* of $A_{n\times n}$ if $A - \lambda I$ is singular. λ is an eigenvalue of *multiplicity s* if the rank of the null space of $A - \lambda I$ is s. A nonzero vector x is an *eigenvector* of A corresponding to the eigenvalue λ if x is in the null space of $A - \lambda I$, i.e., if $Ax = \lambda x$. Eigenvalues are also called *singular values* and *characteristic roots*.

For example,

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 1 \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Combining the two equations gives

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that if $\lambda \neq 0$ is an eigenvalue of A, the eigenvectors corresponding to λ (along with the vector 0) form a subspace of $C(A)$. For example, if $Ax_1 = \lambda x_1$ and $Ax_2 = \lambda x_2$, then $A(x_1 + x_2) = \lambda(x_1 + x_2)$, so the set of eigenvectors is closed under vector addition. Similarly, it is closed under scalar multiplication, so it forms a subspace (except that eigenvectors cannot be 0 and every subspace contains 0). If $\lambda = 0$, the subspace is the null space of A.

If A is a symmetric matrix, and γ and λ are distinct eigenvalues, then the eigenvectors corresponding to λ and γ are orthogonal. To see this, let x be an eigenvector for λ and y an eigenvector for γ. Then $\lambda x'y = x'Ay = \gamma x'y$, which can happen only if $\lambda = \gamma$ or if $x'y = 0$. Since λ and γ are distinct, we have $x'y = 0$.

Let $\lambda_1, \ldots, \lambda_r$ be the distinct nonzero eigenvalues of a symmetric matrix A with respective multiplicities $s(1), \ldots, s(r)$. Let $v_{i1}, \ldots, v_{is(i)}$ be a basis for the space of eigenvectors of λ_i. We want to show that $v_{11}, v_{12}, \ldots, v_{rs(r)}$ is a basis for $C(A)$. Suppose $v_{11}, v_{12}, \ldots, v_{rs(r)}$ is not a basis. Since $v_{ij} \in C(A)$ and the v_{ij}s are linearly inde-

pendent, we can pick $x \in C(A)$ with $x \perp v_{ij}$ for all i and j. Note that since $Av_{ij} = \lambda_i v_{ij}$, we have $(A)^p v_{ij} = (\lambda_i)^p v_{ij}$. In particular, $x'(A)^p v_{ij} = x'(\lambda_i)^p v_{ij} = (\lambda_i)^p x' v_{ij} = 0$, so $A^p x \perp v_{ij}$ for any i, j, and p. The vectors $x, Ax, A^2x, \ldots$ cannot all be linearly independent, so there exists a smallest value $k \leq n$ such that

$$A^k x + b_{k-1}A^{k-1}x + \cdots + b_0 x = 0.$$

Since there is a solution to this, for some real number μ we can write the equation as

$$(A - \mu I)\left(A^{k-1}x + \gamma_{k-2}A^{k-2}x + \cdots + \gamma_0 x\right) = 0,$$

and μ is an eigenvalue. (See Exercise B.1.) An eigenvector for μ is $y = A^{k-1}x + \gamma_{k-2}A^{k-2}x + \cdots + \gamma_0 x$. Clearly, $y \perp v_{ij}$ for any i and j. Since k was chosen as the smallest value to get linear dependence, we have $y \neq 0$. If $\mu \neq 0$, y is an eigenvector that does not correspond to any of $\lambda_1, \ldots, \lambda_r$, a contradiction. If $\mu = 0$, we have $Ay = 0$; and since A is symmetric, y is a vector in $C(A)$ that is orthogonal to every other vector in $C(A)$, i.e., $y'y = 0$ but $y \neq 0$, a contradiction. We have proven

Theorem B.14. If A is a symmetric matrix, then there exists a basis for $C(A)$ consisting of eigenvectors of nonzero eigenvalues. If λ is a nonzero eigenvalue of multiplicity s, then the basis will contain s eigenvectors for λ.

If λ is an eigenvalue of A with multiplicity s, then we can think of λ as being an eigenvalue s times. With this convention, the rank of A is the number of nonzero eigenvalues. The total number of eigenvalues is n if A is an $n \times n$ matrix.

For a symmetric matrix A, if we use eigenvectors corresponding to the zero eigenvalue, we can get a basis for $\mathbf{R}^n$ consisting of eigenvectors. We already have a basis for $C(A)$, and the eigenvectors of 0 are the null space of A. For A symmetric, $C(A)$ and the null space of A are orthogonal complements. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of a symmetric matrix A. Let $v_1, \ldots, v_n$ denote a basis of eigenvectors for $\mathbf{R}^n$, with v_i being an eigenvector for λ_i for any i.

Theorem B.15. If A is symmetric, there exists an orthonormal basis for $\mathbf{R}^n$ consisting of eigenvectors of A.

PROOF. Assume $\lambda_{i1} = \cdots = \lambda_{ik}$ are all the λ_is equal to any particular value λ, and let $v_{i1}, \ldots, v_{ik}$ be a basis for the space of eigenvectors for λ. By Gram–Schmidt there exists an orthonormal basis $w_{i1}, \ldots, w_{ik}$ for the space of eigenvectors corresponding to λ. If we do this for each distinct eigenvalue, we get a collection of orthonormal sets that form a basis for $\mathbf{R}^n$. Since, as we have seen, for $\lambda_i \neq \lambda_j$, any eigenvector for λ_i is orthogonal to any eigenvector for λ_j, the basis is orthonormal. □

Definition B.16. A square matrix P is *orthogonal* if $P' = P^{-1}$. Note that if P is orthogonal, so is P'.

Some examples of orthogonal matrices are

$$P_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{2} & -\sqrt{3} & 1 \\ \sqrt{2} & 0 & -2 \\ \sqrt{2} & \sqrt{3} & 1 \end{bmatrix}, \quad P_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Theorem B.17. $P_{n\times n}$ is orthogonal if and only if the columns of P form an orthonormal basis for $\mathbf{R}^n$.

PROOF. $\Leftarrow$ It is clear that if the columns of P form an orthonormal basis for $\mathbf{R}^n$, then $P'P = I$.

$\Rightarrow$ Since P is nonsingular, the columns of P form a basis for $\mathbf{R}^n$. Since $P'P = I$, the basis is orthonormal. □

Corollary B.18. $P_{n\times n}$ is orthogonal if and only if the rows of P form an orthonormal basis for $\mathbf{R}^n$.

PROOF. P is orthogonal if and only if P' is orthogonal if and only if the columns of P' are an orthonormal basis if and only if the rows of P are an orthonormal basis. □

Theorem B.19. If A is an $n \times n$ symmetric matrix, then there exists an orthogonal matrix P such that $P'AP = \text{Diag}(\lambda_i)$, where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A.

PROOF. Let $v_1, v_2, \ldots, v_n$ be an orthonormal set of eigenvectors of A corresponding, respectively, to $\lambda_1, \lambda_2, \ldots, \lambda_n$. Let $P = [v_1, \ldots, v_n]$. Then

$$\begin{aligned} P'AP &= \begin{bmatrix} v_1' \\ \vdots \\ v_n' \end{bmatrix} [Av_1, \ldots, Av_n] \\ &= \begin{bmatrix} v_1' \\ \vdots \\ v_n' \end{bmatrix} [\lambda_1 v_1, \ldots, \lambda_n v_n] \\ &= \begin{bmatrix} \lambda_1 v_1'v_1 & \ldots & \lambda_n v_1'v_n \\ \vdots & \ddots & \vdots \\ \lambda_1 v_n'v_1 & \ldots & \lambda_n v_n'v_n \end{bmatrix} \\ &= \text{Diag}(\lambda_i). \end{aligned}$$

□

The *singular value decomposition* for a symmetric matrix is given by the following corollary.

Corollary B.20. $A = PD(\lambda_i)P'$.

For example, using results illustrated earlier,

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

Definition B.21. A symmetric matrix A is *positive (nonnegative) definite* if, for any nonzero vector $v \in \mathbf{R}^n$, $v'Av$ is positive (nonnegative).

Theorem B.22. A is nonnegative definite if and only if there exists a square matrix Q such that $A = QQ'$.

PROOF. $\Rightarrow$ We know that there exists P orthogonal with $P'AP = \text{Diag}(\lambda_i)$. The λ_is must all be nonnegative, because if $e_j' = (0,\ldots,0,1,0,\ldots,0)$ with the 1 in the jth place and we let $v = Pe_j$, then $0 \le v'Av = e_j'\text{Diag}(\lambda_i)e_j = \lambda_j$. Let $Q = P\text{Diag}\left(\sqrt{\lambda_i}\right)$. Then, since $P\text{Diag}(\lambda_i)P' = A$, we have

$$QQ' = P\text{Diag}(\lambda_i)P' = A.$$

$\Leftarrow$ If $A = QQ'$, then $v'Av = (Q'v)'(Q'v) \ge 0$. □

Corollary B.23. A is positive definite if and only if Q is nonsingular for any choice of Q.

PROOF. There exists $v \ne 0$ such that $v'Av = 0$ if and only if there exists $v \ne 0$ such that $Q'v = 0$, which occurs if and only if Q' is singular. The contrapositive of this is that $v'Av > 0$ for all $v \ne 0$ if and only if Q' is nonsingular. □

Theorem B.24. If A is an $n \times n$ nonnegative definite matrix with nonzero eigenvalues $\lambda_1,\ldots,\lambda_r$, then there exists an $n \times r$ matrix $Q = Q_1Q_2^{-1}$ such that Q_1 has orthonormal columns, $C(Q_1) = C(A)$, Q_2 is diagonal and nonsingular, and $Q'AQ = I$.

PROOF. Let $v_1,\ldots,v_n$ be an orthonormal basis of eigenvectors with $v_1,\ldots,v_r$ corresponding to $\lambda_1,\ldots,\lambda_r$. Let $Q_1 = [v_1,\ldots,v_r]$. By an argument similar to that in the proof of Theorem B.19, $Q_1'AQ_1 = \text{Diag}(\lambda_i)$, $i = 1,\ldots,r$. Now take $Q_2 = \text{Diag}(\sqrt{\lambda_i})$ and $Q = Q_1Q_2^{-1}$. □

Corollary B.25. Let $W = Q_1Q_2$. Then $WW' = A$.

PROOF. Since $Q'AQ = Q_2^{-1}Q_1'AQ_1Q_2^{-1} = I$ and Q_2 is symmetric, $Q_1'AQ_1 = Q_2Q_2'$. Multiplying gives

$$Q_1Q_1'AQ_1Q_1' = (Q_1Q_2)(Q_2'Q_1') = WW'.$$

But Q_1Q_1' is a perpendicular projection matrix onto $C(A)$, so $Q_1Q_1'AQ_1Q_1' = A$ (cf. Definition B.31 and Theorem B.35). □

Corollary B.26. $AQQ'A = A$ and $QQ'AQQ' = QQ'$.

PROOF. $AQQ'A = WW'QQ'WW' = WQ_2Q_1'Q_1Q_2^{-1}Q_2^{-1}Q_1'Q_1Q_2W' = A$. Moreover, $QQ'AQQ' = QQ'WW'QQ' = QQ_2^{-1}Q_1'Q_1Q_2Q_2Q_1'Q_1Q_2^{-1}Q' = QQ'$. □

Definition B.27. Let $A = [a_{ij}]$ be an $n \times n$ matrix. The *trace* of A is $\text{tr}(A) = \sum_{i=1}^n a_{ii}$.

Theorem B.28. For matrices $A_{r\times s}$ and $B_{s\times r}$, $\text{tr}(AB) = \text{tr}(BA)$.

PROOF. See Exercise B.8. □

Theorem B.29. If A is a symmetric matrix, $\text{tr}(A) = \sum_{i=1}^n \lambda_i$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

PROOF. $A = PD(\lambda_i)P'$ with P orthogonal

$$\begin{aligned}\text{tr}(A) &= \text{tr}(PD(\lambda_i)P') = \text{tr}(D(\lambda_i)P'P)\\ &= \text{tr}(D(\lambda_i)) = \sum_{i=1}^n \lambda_i.\end{aligned}$$ □

To illustrate, we saw earlier that the matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ had eigenvalues of 3 and 1. In fact, a stronger result than Theorem B.29 is true. We give it without proof.

Theorem B.30. $\text{tr}(A) = \sum_{i=1}^n \lambda_i$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. Moreover, the determinant of A is $\det(V) = \prod_{i=1}^n \lambda_i$.

B.3 Projections

This section is devoted primarily to a discussion of perpendicular projection operators. It begins with their definition, some basic properties, and two important characterizations: Theorems B.33 and B.35. A third important characterization, Theorem B.44, involves generalized inverses. Generalized inverses are defined, briefly

studied, and applied to projection operators. The section continues with the examination of the relationships between two perpendicular projection operators and closes with discussions of the Gram–Schmidt theorem, eigenvalues of projection operators, and oblique (nonperpendicular) projection operators.

We begin by defining a *perpendicular projection operator (ppo)* onto an arbitrary space. To be consistent with later usage, we denote the arbitrary space $C(X)$ for some matrix X.

Definition B.31. M is a perpendicular projection operator (matrix) onto $C(X)$ if and only if

(i) $v \in C(X)$ implies $Mv = v$ (projection),
(ii) $w \perp C(X)$ implies $Mw = 0$ (perpendicularity).

For example, consider the subspace of $\mathbf{R}^2$ determined by vectors of the form $(2a, a)'$. It is not difficult to see that the orthogonal complement of this subspace consists of vectors of the form $(b, -2b)'$. The perpendicular projection operator onto the $(2a, a)'$ subspace is

$$M = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix}.$$

To verify this note that

$$M \begin{pmatrix} 2a \\ a \end{pmatrix} = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix} \begin{pmatrix} 2a \\ a \end{pmatrix} = \begin{pmatrix} (0.8)2a + 0.4a \\ (0.4)2a + 0.2a \end{pmatrix} = \begin{pmatrix} 2a \\ a \end{pmatrix}$$

and

$$M \begin{pmatrix} b \\ -2b \end{pmatrix} = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix} \begin{pmatrix} b \\ -2b \end{pmatrix} = \begin{pmatrix} 0.8b + 0.4(-2b) \\ 0.4b + 0.2(-2b) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Notationally, M is used to indicate the ppo onto $C(X)$. If A is another matrix, M_A denotes the ppo onto $C(A)$. Thus, $M \equiv M_X$. When using X with a subscript, say 0, write the ppo onto $C(X_0)$ as $M_0 \equiv M_{X_0}$.

Proposition B.32. If M is a perpendicular projection operator onto $C(X)$, then $C(M) = C(X)$.

PROOF. See Exercise B.2. □

Note that both columns of

$$M = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix}$$

have the form $(2a, a)'$.

Theorem B.33. M is a perpendicular projection operator on $C(M)$ if and only if $MM = M$ and $M' = M$.

PROOF. $\Rightarrow$ Write $v = v_1 + v_2$, where $v_1 \in C(M)$ and $v_2 \perp C(M)$, and let $w = w_1 + w_2$ with $w_1 \in C(M)$ and $w_2 \perp C(M)$. Since $(I-M)v = (I-M)v_2 = v_2$ and $Mw = Mw_1 = w_1$, we get

$$w'M'(I-M)v = w_1'M'(I-M)v_2 = w_1'v_2 = 0.$$

This is true for any v and w, so we have $M'(I-M) = 0$ or $M' = M'M$. Since $M'M$ is symmetric, M' must also be symmetric, and this implies that $M = MM$.

$\Leftarrow$ If $M^2 = M$ and $v \in C(M)$, then since $v = Mb$ we have $Mv = MMb = Mb = v$. If $M' = M$ and $w \perp C(M)$, then $Mw = M'w = 0$ because the columns of M are in $C(M)$. □

In our example,

$$MM = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix} \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix} = M$$

and

$$M = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix} = M'.$$

Proposition B.34. Perpendicular projection operators are unique.

PROOF. Let M and P be perpendicular projection operators onto some space $\mathscr{M}$. Let $v \in \mathbf{R}^n$ and write $v = v_1 + v_2$, $v_1 \in \mathscr{M}$, $v_2 \perp \mathscr{M}$. Since v is arbitrary and $Mv = v_1 = Pv$, we have $M = P$. □

For any matrix X, we will now find two ways to characterize the perpendicular projection operator onto $C(X)$. The first method depends on the Gram–Schmidt theorem; the second depends on the concept of a generalized inverse.

Theorem B.35. Let $o_1, \ldots, o_r$ be an orthonormal basis for $C(X)$, and let $O = [o_1, \ldots, o_r]$. Then $OO' = \sum_{i=1}^{r} o_i o_i'$ is the perpendicular projection operator onto $C(X)$.

PROOF. OO' is symmetric and $OO'OO' = OI_rO' = OO'$; so, by Theorem B.33, it only remains to show that $C(OO') = C(X)$. Clearly $C(OO') \subset C(O) = C(X)$. On the other hand, if $v \in C(O)$, then $v = Ob$ for some vector $b \in \mathbf{R}^r$ and $v = Ob = OI_rb = OO'Ob$; so clearly $v \in C(OO')$. □

For example, to find the perpendicular projection operator for vectors of the form $(2a,a)'$, we can find an orthonormal basis. The space has rank 1 and to normalize $(2a,a)'$, we must have

$$1 = (2a,a)' \begin{pmatrix} 2a \\ a \end{pmatrix} = 4a^2 + a^2 = 5a^2;$$

so $a^2 = 1/5$ and $a = \pm 1/\sqrt{5}$. If we take $(2/\sqrt{5}, 1/\sqrt{5})'$ as our orthonormal basis, then

$$M = \begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} (2/\sqrt{5}, 1/\sqrt{5}) = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix},$$

as was demonstrated earlier.

One use of Theorem B.35 is that, given a matrix X, one can use the Gram–Schmidt theorem to get an orthonormal basis for $C(X)$ and thus obtain the perpendicular projection operator.

We now examine properties of generalized inverses. Generalized inverses are a generalization on the concept of the inverse of a matrix. Although the most common use of generalized inverses is in solving systems of linear equations, our interest lies primarily in their relationship to projection operators. The discussion below is given for an arbitrary matrix A.

Definition B.36. A *generalized inverse* of a matrix A is any matrix G such that $AGA = A$. The notation A^- is used to indicate a generalized inverse of A.

Theorem B.37. If A is nonsingular, the unique generalized inverse of A is A^{-1}.

PROOF. $AA^{-1}A = IA = A$, so A^{-1} is a generalized inverse. If $AA^-A = A$, then $AA^- = AA^-AA^{-1} = AA^{-1} = I$; so A^- is the inverse of A. □

Theorem B.38. For any symmetric matrix A, there exists a generalized inverse of A.

PROOF. There exists P orthogonal so that $P'AP = D(\lambda_i)$ and $A = PD(\lambda_i)P'$. Let

$$\gamma_i = \begin{cases} 1/\lambda_i, & \text{if } \lambda_i \neq 0 \\ 0, & \text{if } \lambda_i = 0\,, \end{cases}$$

and $G = PD(\gamma_i)P'$. We now show that G is a generalized inverse of A. P is orthogonal, so $P'P = I$ and

$$\begin{aligned} AGA &= PD(\lambda_i)P'PD(\gamma_i)P'PD(\lambda_i)P' \\ &= PD(\lambda_i)D(\gamma_i)D(\lambda_i)P' \\ &= PD(\lambda_i)P' \\ &= A. \end{aligned}$$

□

Although this is the only existence result we really need, later we will show that generalized inverses exist for arbitrary matrices.

Theorem B.39. If G_1 and G_2 are generalized inverses of A, then so is G_1AG_2.

PROOF. $A(G_1AG_2)A = (AG_1A)G_2A = AG_2A = A$. □

For A symmetric, A^- need not be symmetric.

EXAMPLE B.40. Consider the matrix

$$\begin{bmatrix} a & b \\ b & b^2/a \end{bmatrix}.$$

It has a generalized inverse

$$\begin{bmatrix} 1/a & -1 \\ 1 & 0 \end{bmatrix},$$

and in fact, by considering the equation

$$\begin{bmatrix} a & b \\ b & b^2/a \end{bmatrix} \begin{bmatrix} r & s \\ t & u \end{bmatrix} \begin{bmatrix} a & b \\ b & b^2/a \end{bmatrix} = \begin{bmatrix} a & b \\ b & b^2/a \end{bmatrix},$$

it can be shown that if $r = 1/a$, then any solution of $at + as + bu = 0$ gives a generalized inverse.

Corollary B.41. For a symmetric matrix A, there exists A^- such that $A^-AA^- = A^-$ and $(A^-)' = A^-$.

PROOF. Take A^- as the generalized inverse in the proof of Theorem B.38. Clearly, $A^- = PD(\gamma_i)P'$ is symmetric and

$$A^-AA^- = PD(\gamma_i)P'PD(\lambda_i)P'PD(\gamma_i)P' = PD(\gamma_i)D(\lambda_i)D(\gamma_i)P' = PD(\gamma_i)P' = A^-.$$

□

Definition B.42. A generalized inverse A^- for a matrix A that has the property $A^-AA^- = A^-$ is said to be *reflexive*.

Corollary B.41 establishes the existence of a reflexive generalized inverse for any symmetric matrix. Note that Corollary B.26 previously established the existence of a reflexive generalized inverse for any nonnegative definite matrix.

Generalized inverses are of interest in that they provide an alternative to the characterization of perpendicular projection matrices given in Theorem B.35. The two results immediately below characterize the perpendicular projection matrix onto $C(X)$.

Lemma B.43. If G and H are generalized inverses of $(X'X)$, then

(i) $XGX'X = XHX'X = X$,
(ii) $XGX' = XHX'$.

PROOF. For $v \in \mathbf{R}^n$, let $v = v_1 + v_2$ with $v_1 \in C(X)$ and $v_2 \perp C(X)$. Also let $v_1 = Xb$ for some vector b. Then

$$v'XGX'X = v_1'XGX'X = b'(X'X)G(X'X) = b'(X'X) = v'X.$$

Since v and G are arbitrary, we have shown (i).

To see (ii), observe that for the arbitrary vector v above,

$$XGX'v = XGX'Xb = XHX'Xb = XHX'v. \qquad \square$$

Since $X'X$ is symmetric, there exists a generalized inverse $(X'X)^-$ that is symmetric. For this generalized inverse, $X(X'X)^-X'$ is symmetric; so, by the above lemma, $X(X'X)^-X'$ must be symmetric for any choice of $(X'X)^-$.

Theorem B.44. $X(X'X)^-X'$ is the perpendicular projection operator onto $C(X)$.

PROOF. We need to establish conditions (i) and (ii) of Definition B.31. (i) For $v \in C(X)$, write $v = Xb$, so by Lemma B.43, $X(X'X)^-X'v = X(X'X)^-X'Xb = Xb = v$. (ii) If $w \perp C(X)$, $X(X'X)^-X'w = 0$. $\square$

For example, one spanning set for the subspace of vectors with the form $(2a,a)'$ is $(2,1)'$. It follows that

$$M = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \left[(2,1) \begin{pmatrix} 2 \\ 1 \end{pmatrix} \right]^{-1} (2,1) = \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix},$$

as was shown earlier.

The next five results examine the relationships between two perpendicular projection matrices.

Theorem B.45. Let M_1 and M_2 be perpendicular projection matrices on $\mathbf{R}^n$. $(M_1 + M_2)$ is the perpendicular projection matrix onto $C(M_1, M_2)$ if and only if $C(M_1) \perp C(M_2)$.

PROOF. $\Leftarrow$ If $C(M_1) \perp C(M_2)$, then $M_1M_2 = M_2M_1 = 0$. Because

$$(M_1 + M_2)^2 = M_1^2 + M_2^2 + M_1M_2 + M_2M_1 = M_1^2 + M_2^2 = M_1 + M_2$$

and

$$(M_1+M_2)' = M_1' + M_2' = M_1 + M_2,$$

M_1+M_2 is the perpendicular projection matrix onto $C(M_1+M_2)$. Clearly $C(M_1+M_2) \subset C(M_1,M_2)$. To see that $C(M_1,M_2) \subset C(M_1+M_2)$, write $v = M_1b_1 + M_2b_2$. Then, because $M_1M_2 = M_2M_1 = 0$, $(M_1+M_2)v = v$. Thus, $C(M_1,M_2) = C(M_1+M_2)$.

$\Rightarrow$ If M_1+M_2 is a perpendicular projection matrix, then

$$\begin{aligned}(M_1+M_2) &= (M_1+M_2)^2 = M_1^2 + M_2^2 + M_1M_2 + M_2M_1 \\ &= M_1 + M_2 + M_1M_2 + M_2M_1.\end{aligned}$$

Thus, $M_1M_2 + M_2M_1 = 0$.

Multiplying by M_1 gives $0 = M_1^2M_2 + M_1M_2M_1 = M_1M_2 + M_1M_2M_1$ and thus $-M_1M_2M_1 = M_1M_2$. Since $-M_1M_2M_1$ is symmetric, so is M_1M_2. This gives $M_1M_2 = (M_1M_2)' = M_2M_1$, so the condition $M_1M_2 + M_2M_1 = 0$ becomes $2(M_1M_2) = 0$ or $M_1M_2 = 0$. By symmetry, this says that the columns of M_1 are orthogonal to the columns of M_2. □

Theorem B.46. If M_1 and M_2 are symmetric, $C(M_1) \perp C(M_2)$, and (M_1+M_2) is a perpendicular projection matrix, then M_1 and M_2 are perpendicular projection matrices.

PROOF.

$$(M_1+M_2) = (M_1+M_2)^2 = M_1^2 + M_2^2 + M_1M_2 + M_2M_1.$$

Since M_1 and M_2 are symmetric with $C(M_1) \perp C(M_2)$, we have $M_1M_2 + M_2M_1 = 0$ and $M_1+M_2 = M_1^2 + M_2^2$. Rearranging gives $M_2 - M_2^2 = M_1^2 - M_1$, so $C(M_2 - M_2^2) = C(M_1^2 - M_1)$. Now $C(M_2 - M_2^2) \subset C(M_2)$ and $C(M_1^2 - M_1) \subset C(M_1)$, so $C(M_2 - M_2^2) \perp C(M_1^2 - M_1)$. The only way a vector space can be orthogonal to itself is if it consists only of the zero vector. Thus, $M_2 - M_2^2 = M_1^2 - M_1 = 0$, and $M_2 = M_2^2$ and $M_1 = M_1^2$. □

Theorem B.47. Let M and M_0 be perpendicular projection matrices with $C(M_0) \subset C(M)$. Then $M - M_0$ is a perpendicular projection matrix.

PROOF. Since $C(M_0) \subset C(M)$, $MM_0 = M_0$ and, by symmetry, $M_0M = M_0$. Checking the conditions of Theorem B.33, we see that $(M-M_0)^2 = M^2 - MM_0 - M_0M + M_0^2 = M - M_0 - M_0 + M_0 = M - M_0$, and $(M-M_0)' = M - M_0$. □

Theorem B.48. Let M and M_0 be perpendicular projection matrices with $C(M_0) \subset C(M)$. Then $C(M-M_0)$ is the orthogonal complement of $C(M_0)$ with respect to $C(M)$, i.e., $C(M-M_0) = C(M_0)^{\perp}_{C(M)}$.

PROOF. $C(M-M_0) \perp C(M_0)$ because $(M-M_0)M_0 = MM_0 - M_0^2 = M_0 - M_0 = 0$. Thus, $C(M-M_0)$ is contained in the orthogonal complement of $C(M_0)$ with respect to $C(M)$. If $x \in C(M)$ and $x \perp C(M_0)$, then $x = Mx = (M-M_0)x + M_0x = (M-M_0)x$. Thus, $x \in C(M-M_0)$, so the orthogonal complement of $C(M_0)$ with respect to $C(M)$ is contained in $C(M-M_0)$. □

Corollary B.49. $r(M) = r(M_0) + r(M-M_0)$.

One particular application of these results involves I, the perpendicular projection operator onto $\mathbf{R}^n$. For any other perpendicular projection operator M, $I-M$ is the perpendicular projection operator onto the orthogonal complement of $C(M)$ with respect to $\mathbf{R}^n$.

For example, the subspace of vectors with the form $(2a, a)'$ has an orthogonal complement consisting of vectors with the form $(b, -2b)'$. With M as given earlier,

$$I - M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.8 & 0.4 \\ 0.4 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.2 & -0.4 \\ -0.4 & 0.8 \end{bmatrix}.$$

Note that

$$(I-M)\begin{pmatrix} b \\ -2b \end{pmatrix} = \begin{pmatrix} b \\ -2b \end{pmatrix} \quad \text{and} \quad (I-M)\begin{pmatrix} 2a \\ a \end{pmatrix} = 0;$$

so by definition $I-M$ is the perpendicular projection operator onto the space of vectors with the form $(b, -2b)'$.

At this point, we examine the relationship between perpendicular projection operations and the Gram–Schmidt theorem (Theorem A.12). Recall that in the Gram–Schmidt theorem, $x_1, \ldots, x_r$ denotes the original basis and $y_1, \ldots, y_r$ denotes the orthonormal basis. Let

$$M_s = \sum_{i=1}^{s} y_i y_i'.$$

Applying Theorem B.35, M_s is the ppo onto $C(x_1, \ldots, x_s)$. Now define

$$w_{s+1} = (I - M_s)x_{s+1}.$$

Thus, w_{s+1} is the projection of x_{s+1} onto the orthogonal complement of $C(x_1, \ldots, x_s)$. Finally, y_{s+1} is just w_{s+1} normalized.

Consider the eigenvalues of a perpendicular projection operator M. Let $v_1, \ldots, v_r$ be a basis for $C(M)$. Then $Mv_i = v_i$, so v_i is an eigenvector of M with eigenvalue 1. In fact, 1 is an eigenvalue of M with multiplicity r. Now, let $w_1, \ldots, w_{n-r}$ be a basis for $C(M)^{\perp}$. $Mw_j = 0$, so 0 is an eigenvalue of M with multiplicity $n-r$. We have completely characterized the n eigenvalues of M. Since $\operatorname{tr}(M)$ equals the sum of the eigenvalues, we have $\operatorname{tr}(M) = r(M)$.

In fact, if A is an $n \times n$ matrix with $A^2 = A$, any basis for $C(A)$ is a basis for the space of eigenvectors for the eigenvalue 1. The null space of A is the space of eigenvectors for the eigenvalue 0. The rank of A and the rank of the null space of A

add to n, and A has n eigenvalues, so all the eigenvalues are accounted for. Again, $\mathrm{tr}(A) = r(A)$.

Definition B.50.

(a) If A is a square matrix with $A^2 = A$, then A is called *idempotent*.
(b) Let $\mathscr{N}$ and $\mathscr{M}$ be two spaces with $\mathscr{N} \cap \mathscr{M} = \{0\}$ and $r(\mathscr{N}) + r(\mathscr{M}) = n$. The square matrix A is a *projection operator* onto $\mathscr{N}$ along $\mathscr{M}$ if 1) $Av = v$ for any $v \in \mathscr{N}$, and 2) $Aw = 0$ for any $w \in \mathscr{M}$.

If the square matrix A has the property that $Av = v$ for any $v \in C(A)$, then A is the projection operator (matrix) onto $C(A)$ along $C(A')^{\perp}$. (Note that $C(A')^{\perp}$ is the null space of A.) It follows immediately that if A is idempotent, then A is a projection operator onto $C(A)$ along $\mathscr{N}(A) = C(A')^{\perp}$.

The uniqueness of projection operators can be established like it was for perpendicular projection operators. Note that $x \in \mathbf{R}^n$ can be written uniquely as $x = v + w$ for $v \in \mathscr{N}$ and $w \in \mathscr{M}$. To see this, take basis matrices for the two spaces, say N and M, respectively. The result follows from observing that $[N, M]$ is a basis matrix for $\mathbf{R}^n$. Because of the rank conditions, $[N, M]$ is an $n \times n$ matrix. It is enough to show that the columns of $[N, M]$ must be linearly independent.

$$0 = [N, M] \begin{bmatrix} b \\ c \end{bmatrix} = Nb + Mc$$

implies $Nb = M(-c)$ which, since $\mathscr{N} \cap \mathscr{M} = \{0\}$, can only happen when $Nb = 0 = M(-c)$, which, because they are basis matrices, can only happen when $b = 0 = (-c)$, which implies that $\begin{bmatrix} b \\ c \end{bmatrix} = 0$, and we are done.

Any projection operator that is not a perpendicular projection is referred to as an *oblique projection operator*.

To show that a matrix A is a projection operator onto an arbitrary space, say $C(X)$, it is necessary to show that $C(A) = C(X)$ and that for $x \in C(X)$, $Ax = x$. A typical proof runs in the following pattern. First, show that $Ax = x$ for any $x \in C(X)$. This also establishes that $C(X) \subset C(A)$. To finish the proof, it suffices to show that $Av \in C(X)$ for any $v \in \mathbf{R}^n$ because this implies that $C(A) \subset C(X)$.

In this book, our use of the word "perpendicular" is based on the standard inner product, that defines Euclidean distance. In other words, for two vectors x and y, their inner product is $x'y$. By definition, the vectors x and y are orthogonal if their inner product is 0. In fact, for any two vectors x and y, let θ be the angle between x and y. Then $x'y = \sqrt{x'x}\sqrt{y'y}\cos\theta$. The length of a vector x is defined as the square root of the inner product of x with itself, i.e., $\|x\| \equiv \sqrt{x'x}$. The distance between two vectors x and y is the length of their difference, i.e., $\|x - y\|$.

These concepts can be generalized. For a positive definite matrix B, we can define an inner product between x and y as $x'By$. As before, x and y are orthogonal if their inner product is 0 and the length of x is the square root of its inner product with

itself (now $\|x\|_B \equiv \sqrt{x'Bx}$). As argued above, any idempotent matrix is always a projection operator, but which one is the perpendicular projection operator depends on the inner product. As can be seen from Proposition 2.7.2 and Exercise 2.5, the matrix $X(X'BX)^{-}X'B$ is an oblique projection onto $C(X)$ for the standard inner product; but it is the perpendicular projection operator onto $C(X)$ with the inner product defined using the matrix B.

B.4 Miscellaneous Results

Proposition B.51. For any matrix X, $C(XX') = C(X)$.

PROOF. Clearly $C(XX') \subset C(X)$, so we need to show that $C(X) \subset C(XX')$. Let $x \in C(X)$. Then $x = Xb$ for some b. Write $b = b_0 + b_1$, where $b_0 \in C(X')$ and $b_1 \perp C(X')$. Clearly, $Xb_1 = 0$, so we have $x = Xb_0$. But $b_0 = X'd$ for some d; so $x = Xb_0 = XX'd$ and $x \in C(XX')$. □

Corollary B.52. For any matrix X, $r(XX') = r(X)$.

PROOF. See Exercise B.4. □

Corollary B.53. If $X_{n\times p}$ has $r(X) = p$, then the $p \times p$ matrix $X'X$ is nonsingular.

PROOF. See Exercise B.5. □

Proposition B.54. If B is nonsingular, $C(XB) = C(X)$.

PROOF. Clearly, $C(XB) \subset C(X)$. To see that $C(X) \subset C(XB)$, take $x \in C(X)$. It follows that for some vector b, $x = Xb$; so $x = XB(B^{-1}b) \in C(XB)$. □

It follows immediately from Proposition B.54 that the perpendicular projection operators onto $C(XB)$ and $C(X)$ are identical.

We now show that generalized inverses always exist.

Theorem B.55. For any matrix X, there exists a generalized inverse X^{-}.

PROOF. We know that $(X'X)^{-}$ exists. Set $X^{-} = (X'X)^{-}X'$. Then $XX^{-}X = X(X'X)^{-}X'X = X$ because $X(X'X)^{-}X'$ is a projection matrix onto $C(X)$. □

Note that for any X^-, the matrix XX^- is idempotent and hence a projection operator.

Proposition B.56. When all inverses exist,

$$[A+BCD]^{-1} = A^{-1} - A^{-1}B\left[C^{-1}+DA^{-1}B\right]^{-1}DA^{-1}.$$

PROOF.

$$\begin{aligned}
&[A+BCD]\left[A^{-1} - A^{-1}B\left[C^{-1}+DA^{-1}B\right]^{-1}DA^{-1}\right] \\
&= I - B\left[C^{-1}+DA^{-1}B\right]^{-1}DA^{-1} + BCDA^{-1} \\
&\quad - BCDA^{-1}B\left[C^{-1}+DA^{-1}B\right]^{-1}DA^{-1} \\
&= I - B\left[I+CDA^{-1}B\right]\left[C^{-1}+DA^{-1}B\right]^{-1}DA^{-1} + BCDA^{-1} \\
&= I - BC\left[C^{-1}+DA^{-1}B\right]\left[C^{-1}+DA^{-1}B\right]^{-1}DA^{-1} + BCDA^{-1} \\
&= I - BCDA^{-1} + BCDA^{-1} = I.
\end{aligned}$$

□

When we study linear models, we frequently need to refer to matrices and vectors that consist entirely of 1s. Such matrices are denoted by the letter J with various subscripts and superscripts to specify their dimensions. J_r^c is an $r \times c$ matrix of 1s. The subscript indicates the number of rows and the superscript indicates the number of columns. If there is only one column, the superscript may be suppressed, e.g., $J_r = J_r^1$. In a context where we are dealing with vectors in $\mathbf{R}^n$, the subscript may also be suppressed, e.g., $J = J_n = J_n^1$.

A matrix of 0s is always denoted by 0.

B.5 Properties of Kronecker Products and Vec Operators

Kronecker products and Vec operators are extremely useful in multivariate analysis and some approaches to variance component estimation. They are also often used in writing balanced ANOVA models. We now present their basic algebraic properties.

1. If the matrices are of conformable sizes, $[A \otimes (B+C)] = [A \otimes B] + [A \otimes C]$.
2. If the matrices are of conformable sizes, $[(A+B) \otimes C] = [A \otimes C] + [B \otimes C]$.
3. If a and b are scalars, $ab[A \otimes B] = [aA \otimes bB]$.
4. If the matrices are of conformable sizes, $[A \otimes B][C \otimes D] = [AC \otimes BD]$.
5. The transpose of a Kronecker product matrix is $[A \otimes B]' = [A' \otimes B']$.
6. The generalized inverse of a Kronecker product matrix is $[A \otimes B]^- = [A^- \otimes B^-]$.
7. For two vectors v and w, $\mathrm{Vec}(vw') = w \otimes v$.
8. For a matrix W and conformable matrices A and B, $\mathrm{Vec}(AWB') = [B \otimes A]\mathrm{Vec}(W)$.

9. For conformable matrices A and B, $\mathrm{Vec}(A)'\mathrm{Vec}(B) = \mathrm{tr}(A'B)$.
10. The Vec operator commutes with any matrix operation that is performed elementwise. For example, $\mathrm{E}\{\mathrm{Vec}(W)\} = \mathrm{Vec}\{\mathrm{E}(W)\}$ when W is a random matrix. Similarly, for conformable matrices A and B and scalar ϕ, $\mathrm{Vec}(A+B) = \mathrm{Vec}(A) + \mathrm{Vec}(B)$ and $\mathrm{Vec}(\phi A) = \phi\,\mathrm{Vec}(A)$.
11. If A and B are positive definite, then $A \otimes B$ is positive definite.

Most of these are well-known facts and easy to establish. Two of them are somewhat more unusual, and we present proofs.

ITEM 8. We show that for a matrix W and conformable matrices A and B, $\mathrm{Vec}(AWB') = [B \otimes A]\mathrm{Vec}(W)$. First note that if $\mathrm{Vec}(AW) = [I \otimes A]\mathrm{Vec}(W)$ and $\mathrm{Vec}(WB') = [B \otimes I]\mathrm{Vec}(W)$, then $\mathrm{Vec}(AWB') = [I \otimes A]\mathrm{Vec}(WB') = [I \otimes A][B \otimes I]\mathrm{Vec}(W) = [B \otimes A]\mathrm{Vec}(W)$.

To see that $\mathrm{Vec}(AW) = [I \otimes A]\mathrm{Vec}(W)$, let W be $r \times s$ and write W in terms of its columns $W = [w_1, \ldots, w_s]$. Then $AW = [Aw_1, \ldots, Aw_s]$ and $\mathrm{Vec}(AW)$ stacks the columns $Aw_1, \ldots, Aw_s$. On the other hand,

$$[I \otimes A]\mathrm{Vec}(W) = \begin{bmatrix} A & & 0 \\ & \ddots & \\ 0 & & A \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_s \end{bmatrix} = \begin{bmatrix} Aw_1 \\ \vdots \\ Aw_s \end{bmatrix}.$$

To see that $\mathrm{Vec}(WB') = [B \otimes I]\mathrm{Vec}(W)$, take W as above and write $B_{m\times s} = [b_{ij}]$ with *rows* $b_1', \ldots, b_m'$. First note that $WB' = [Wb_1, \ldots, Wb_m]$, so $\mathrm{Vec}(WB')$ stacks the columns $Wb_1, \ldots, Wb_m$. Now observe that

$$[B \otimes I_r]\mathrm{Vec}(W) = \begin{bmatrix} b_{11}I_r & \cdots & b_{1s}I_r \\ \vdots & \ddots & \vdots \\ b_{m1}I_r & \cdots & b_{ms}I_r \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_s \end{bmatrix} = \begin{bmatrix} Wb_1 \\ \vdots \\ Wb_m \end{bmatrix}.$$

ITEM 11. To see that if $A_{r\times r}$ and $B_{s\times s}$ are positive definite, then $A \otimes B$ is positive definite, consider the eigenvalues and eigenvectors of A and B. Recall that a symmetric matrix is positive definite if and only if all of its eigenvalues are positive. Suppose that $Av = \phi v$ and $Bw = \theta w$. We now show that all of the eigenvalues of $A \otimes B$ are positive. Observe that

$$\begin{aligned} [A \otimes B][v \otimes w] &= [Av \otimes Bw] \\ &= [\phi v \otimes \theta w] \\ &= \phi\theta[v \otimes w]. \end{aligned}$$

This shows that $[v \otimes w]$ is an eigenvector of $[A \otimes B]$ corresponding to the eigenvalue $\phi\theta$. As there are r choices for ϕ and s choices for θ, this accounts for all rs of the eigenvalues in the $rs \times rs$ matrix $[A \otimes B]$. Moreover, ϕ and θ are both positive, so all of the eigenvalues of $[A \otimes B]$ are positive.

B.6 Tensors

Tensors are simply an alternative notation for writing vectors. This notation has substantial advantages when dealing with quadratic forms and when dealing with more general concepts than quadratic forms. Our main purpose in discussing them here is simply to illustrate how flexibly subscripts can be used in writing vectors.

Consider a vector $Y = (y_1, \ldots, y_n)'$. The tensor notation for this is simply y_i. We can write another vector $a = (a_1, \ldots, a_n)'$ as a_i. When written individually, the subscript is not important. In other words, a_i is the same vector as a_j. Note that the length of these vectors needs to be understood from the context. Just as when we write Y and a in conventional vector notation, there is nothing in the notation y_i or a_i to tell us how many elements are in the vector.

If we want the inner product $a'Y$, in tensor notation we write $a_i y_i$. Here we are using something called the *summation convention*. Because the subscripts on a_i and y_i are the same, $a_i y_i$ is taken to mean $\sum_{i=1}^{n} a_i y_i$. If, on the other hand, we wrote $a_i y_j$, this means something completely different. $a_i y_j$ is an alternative notation for the Kronecker product $[a \otimes Y] = (a_1 y_1, \ldots, a_1 y_n, a_2 y_1, \ldots, a_n y_n)'$. In $[a \otimes Y] \equiv a_i y_j$, we have two subscripts identifying the rows of the vector.

Now, suppose we want to look at a quadratic form $Y'AY$, where Y is an n vector and A is $n \times n$. One way to rewrite this is

$$Y'AY = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i a_{ij} y_j = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} y_i y_j = \mathrm{Vec}(A)'[Y \otimes Y].$$

Here we have rewritten the quadratic form as a linear combination of the elements in the vector $[Y \otimes Y]$. The linear combination is determined by the elements of the vector $\mathrm{Vec}(A)$. In tensor notation, this becomes quite simple. Using the summation convention in which objects with the same subscript are summed over,

$$Y'AY = y_i a_{ij} y_j = a_{ij} y_i y_j.$$

The second term just has the summation signs removed, but the third term, which obviously gives the same sum as the second, is actually the tensor notation for $\mathrm{Vec}(A)'[Y \otimes Y]$. Again, $\mathrm{Vec}(A) = (a_{11}, a_{21}, a_{31}, \ldots, a_{nn})'$ uses two subscripts to identify rows of the vector. Obviously, if you had a need to consider things like

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ijk} y_i y_j y_k \equiv a_{ijk} y_i y_j y_k,$$

the tensor version $a_{ijk} y_i y_j y_k$ saves some work.

There is one slight complication in how we have been writing things. Suppose A is not symmetric and we have another n vector W. Then we might want to consider

$$W'AY = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i a_{ij} y_j.$$

From item 8 in the previous subsection,

$$W'AY = \text{Vec}(W'AY) = [Y' \otimes W']\text{Vec}(A).$$

Alternatively,

$$W'AY = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i a_{ij} y_j = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} y_j w_i = \text{Vec}(A)'[Y \otimes W]$$

or $W'AY = Y'A'W = \text{Vec}(A')'[W \otimes Y]$. However, with A nonsymmetric, $W'A'Y = \text{Vec}(A')'[Y \otimes W]$ is typically different from $W'AY$. The Kronecker notation requires that care be taken in specifying the order of the vectors in the Kronecker product, and whether or not to transpose A before using the Vec operator. In tensor notation, $W'AY$ is simply $w_i a_{ij} y_j$. In fact, the orders of the vectors can be permuted in any way; so, for example, $a_{ij} y_j w_i$ means the same thing. $W'A'Y$ is simply $w_i a_{ji} y_j$. The tensor notation and the matrix notation require less effort than the Kronecker notation.

For our purposes, the real moral here is simply that the subscripting of an individual vector does not matter. We can write a vector $Y = (y_1, \ldots, y_n)'$ as $Y = [y_k]$ (in tensor notation as simply y_k), or we can write the same n vector as $Y = [y_{ij}]$ (in tensor notation, simply y_{ij}), where, as long as we know the possible values that i and j can take on, the actual order in which we list the elements is not of much importance. Thus, if $i = 1, \ldots, t$ and $j = 1, \ldots, N_i$, with $n = \sum_{i=1}^{t} N_i$, it really does not matter if we write a vector Y as $(y_1, \ldots, y_n)$, or $(y_{11}, \ldots, y_{1N_1}, y_{21}, \ldots, y_{tN_t})'$ or $(y_{t1}, \ldots, y_{tN_t}, y_{t-1,1}, \ldots, y_{1N_1})'$ or in any other fashion we may choose, as long as we keep straight which row of the vector is which. Thus, a linear combination $a'Y$ can be written $\sum_{k=1}^{n} a_k y_k$ or $\sum_{i=1}^{t}\sum_{j=1}^{N_i} a_{ij} y_{ij}$. In tensor notation, the first of these is simply $a_k y_k$ and the second is $a_{ij} y_{ij}$. These ideas become very handy in examining analysis of variance models, where the standard approach is to use multiple subscripts to identify the various observations. The subscripting has no intrinsic importance; the only thing that matters is knowing which row is which in the vectors. The subscripts are an aid in this identification, but they do not create any problems. We can still put all of the observations into a vector and use standard operations on them.

B.7 Exercises

Exercise B.1

(a) Show that

$$A^k x + b_{k-1}A^{k-1}x + \cdots + b_0 x = (A - \mu I)\left(A^{k-1}x + \tau_{k-2}A^{k-2}x + \cdots + \tau_0 x\right) = 0,$$

where μ is any nonzero solution of $b_0 + b_1 w + \cdots + b_k w^k = 0$ with $b_k = 1$ and $\tau_j = -(b_0 + b_1\mu + \cdots + b_j\mu^j)/\mu^{j+1}$, $j = 0, \ldots, k$.

(b) Show that if the only root of $b_0 + b_1 w + \cdots + b_k w^k$ is zero, then the factorization in (a) still holds.

(c) The solution μ used in (a) need not be a real number, in which case μ is a complex eigenvalue and the τ_is are complex; so the eigenvector is complex. Show that with A symmetric, μ must be real because the eigenvalues of A must be real. In particular, assume that

$$A(y + iz) = (\lambda + i\gamma)(y + iz),$$

for y, z, λ, and γ real vectors and scalars, respectively, set $Ay = \lambda y - \gamma z$, $Az = \gamma y + \lambda z$, and examine $z'Ay = y'Az$.

Exercise B.2 Prove Proposition B.32.

Exercise B.3 Show that any nonzero symmetric matrix A can be written as $A = PDP'$, where $C(A) = C(P)$, $P'P = I$, and D is nonsingular.

Exercise B.4 Prove Corollary B.52.

Exercise B.5 Prove Corollary B.53.

Exercise B.6 Show $\mathrm{tr}(cI_n) = nc$.

Exercise B.7 Let a, b, c, and d be real numbers. If $ad - bc \neq 0$, find the inverse of

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Exercise B.8 Prove Theorem B.28, i.e., let A be an $r \times s$ matrix, let B be an $s \times r$ matrix, and show that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Exercise B.9 Determine whether the matrices given below are positive definite, nonnegative definite, or neither.

$$\begin{bmatrix} 3 & 2 & -2 \\ 2 & 2 & -2 \\ -2 & -2 & 10 \end{bmatrix}, \quad \begin{bmatrix} 26 & -2 & -7 \\ -2 & 4 & -6 \\ -7 & -6 & 13 \end{bmatrix}, \quad \begin{bmatrix} 26 & 2 & 13 \\ 2 & 4 & 6 \\ 13 & 6 & 13 \end{bmatrix}, \quad \begin{bmatrix} 3 & 2 & -2 \\ 2 & -2 & -2 \\ -2 & -2 & 10 \end{bmatrix}.$$

Exercise B.10 Show that the matrix B given below is positive definite, and find

a matrix Q such that $B = QQ'$. (Hint: The first row of Q can be taken as $(1,-1,0)$.)

$$B = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix}.$$

Exercise B.11 Let

$$A = \begin{bmatrix} 2 & 0 & 4 \\ 1 & 5 & 7 \\ 1 & -5 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 4 & 1 \\ 2 & 5 & 1 \\ -3 & 0 & 1 \end{bmatrix}.$$

Use Theorem B.35 to find the perpendicular projection operator onto the column space of each matrix.

Exercise B.12 Show that for a perpendicular projection matrix M,

$$\sum_i \sum_j m_{ij}^2 = r(M).$$

Exercise B.13 Prove that if $M = M'M$, then $M = M'$ and $M = M^2$.

Exercise B.14 Let M_1 and M_2 be perpendicular projection matrices, and let M_0 be a perpendicular projection operator onto $C(M_1) \cap C(M_2)$. Show that the following are equivalent:

(a) $M_1 M_2 = M_2 M_1$.

(b) $M_1 M_2 = M_0$.

(c) $\left\{C(M_1) \cap [C(M_1) \cap C(M_2)]^\perp\right\} \perp \left\{C(M_2) \cap [C(M_1) \cap C(M_2)]^\perp\right\}$.

Hints: (i) Show that $M_1 M_2$ is a projection operator. (ii) Show that $M_1 M_2$ is symmetric. (iii) Note that $C(M_1) \cap [C(M_1) \cap C(M_2)]^\perp = C(M_1 - M_0)$.

Exercise B.15 Let M_1 and M_2 be perpendicular projection matrices. Show that

(a) the eigenvalues of $M_1 M_2$ have length no greater than 1 in absolute value (they may be complex);

(b) $\text{tr}(M_1 M_2) \leq r(M_1 M_2)$.

Hints: For part (a) show that with $x'Mx \equiv \|Mx\|^2$, $\|Mx\| \leq \|x\|$ for any perpendicular projection operator M. Use this to show that if $M_1 M_2 x = \lambda x$, then $\|M_1 M_2 x\| \geq |\lambda| \, \|M_1 M_2 x\|$.

Exercise B.16 For vectors x and y, let $M_x = x(x'x)^{-1}x'$ and $M_y = y(y'y)^{-1}y'$. Show that $M_x M_y = M_y M_x$ if and only if $C(x) = C(y)$ or $x \perp y$.

Exercise B.17 Consider the matrix

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

(a) Show that A is a projection matrix.
(b) Is A a perpendicular projection matrix? Why or why not?
(c) Describe the space that A projects onto and the space that A projects along. Sketch these spaces.
(d) Find another projection operator onto the space that A projects onto.

Exercise B.18 Let A be an arbitrary projection matrix. Show that $C(I-A) = C(A')^{\perp}$.

Hints: Recall that $C(A')^{\perp}$ is the null space of A. Show that $(I-A)$ is a projection matrix.

Exercise B.19 Show that if A^{-} is a generalized inverse of A, then so is

$$G = A^{-}AA^{-} + (I - A^{-}A)B_1 + B_2(I - AA^{-})$$

for any choices of B_1 and B_2 with conformable dimensions.

Exercise B.20 Let A be positive definite with eigenvalues $\lambda_1, \ldots, \lambda_n$. Show that A^{-1} has eigenvalues $1/\lambda_1, \ldots, 1/\lambda_n$ and the same eigenvectors as A.

Exercise B.21 For A nonsingular, let

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

and let $A_{1\cdot 2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$. Show that if all inverses exist,

$$A^{-1} = \begin{bmatrix} A_{1\cdot 2}^{-1} & -A_{1\cdot 2}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}A_{1\cdot 2}^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}A_{1\cdot 2}^{-1}A_{12}A_{22}^{-1} \end{bmatrix}$$

and that

$$A_{22}^{-1} + A_{22}^{-1}A_{21}A_{1\cdot 2}^{-1}A_{12}A_{22}^{-1} = \left[A_{22} - A_{21}A_{11}^{-1}A_{12}\right]^{-1}.$$

Appendix C
Some Univariate Distributions

The tests and confidence intervals presented in this book rely almost exclusively on the χ^2, t, and F distributions. This appendix defines each of the distributions.

Definition C.1. Let $Z_1, \ldots, Z_n$ be independent with $Z_i \sim N(\mu_i, 1)$. Then

$$W = \sum_{i=1}^{n} Z_i^2$$

has a *noncentral chi-squared distribution* with n degrees of freedom and *noncentrality parameter* $\gamma = \sum_{i=1}^{n} \mu_i^2/2$. Write $W \sim \chi^2(n, \gamma)$.

See Rao (1973, Section 3b.2) for a proof that the distribution of W depends only on n and γ.

It is evident from the definition that if $X \sim \chi^2(r, \gamma)$ and $Y \sim \chi^2(s, \delta)$ with X and Y independent, then $(X+Y) \sim \chi^2(r+s, \gamma+\delta)$. A central χ^2 distribution is a distribution with a noncentrality parameter of zero, i.e., $\chi^2(r, 0)$. We will use $\chi^2(r)$ to denote a $\chi^2(r, 0)$ distribution. The 100αth percentile of a $\chi^2(r)$ distribution is the point $\chi^2(\alpha, r)$ that satisfies the equation

$$\Pr\left[\chi^2(r) \leq \chi^2(\alpha, r)\right] = \alpha.$$

Note that if $0 \leq a < 1$, the $100a$ percentile of a central $\chi^2(b)$ is denoted $\chi^2(a, b)$. However, if a is a positive integer, $\chi^2(a, b)$ denotes a noncentral chi-squared distribution.

Definition C.2. Let $X \sim N(\mu, 1)$ and $Y \sim \chi^2(n)$ with X and Y independent. Then

$$W = \frac{X}{\sqrt{Y/n}}$$

has a *noncentral t distribution* with n degrees of freedom and noncentrality parameter μ. Write $W \sim t(n,\mu)$. If $\mu = 0$, we say that the distribution is a central t distribution and write $W \sim t(n)$. The 100αth percentile of a $t(n)$ distribution is denoted $t(\alpha,n)$.

Definition C.3. Let $X \sim \chi^2(r,\gamma)$ and $Y \sim \chi^2(s,0)$ with X and Y independent. Then

$$W = \frac{X/r}{Y/s}$$

has a *noncentral F distribution* with r numerator and s denominator degrees of freedom and noncentrality parameter γ. Write $W \sim F(r,s,\gamma)$. If $\gamma = 0$, write $W \sim F(r,s)$ for the central F distribution. The 100αth percentile $F(r,s)$ is denoted $F(\alpha,r,s)$.

As indicated, if the noncentrality parameter of any of these distributions is zero, the distribution is referred to as a *central distribution* (e.g., central F distribution). The central distributions are those commonly used in statistical methods courses. If any of these distributions is not specifically identified as a noncentral distribution, it should be assumed to be a central distribution.

It is easily seen from Definition C.1 that any noncentral chi-squared distribution *tends* to be larger than the central chi-squared distribution with the same number of degrees of freedom. Similarly, from Definition C.3, a noncentral F tends to be larger than the corresponding central F distribution. (These ideas are made rigorous in Exercise C.1.) The fact that the noncentral F distribution tends to be larger than the corresponding central F distribution is the basis for many of the tests used in linear models. Typically, test statistics are used that have a central F distribution if the reduced (null) model is true and a noncentral F distribution if the full model is true but the null model is not. Since the noncentral F distribution tends to be larger, large values of the test statistic are more consistent with the full model than with the null. Thus, the form of an appropriate rejection region when the full model is true is to reject the null hypothesis for large values of the test statistic.

The power of these F tests is simply a function of the noncentrality parameter. Given a value for the noncentrality parameter, there is no theoretical difficulty in finding the power of an F test. The power simply involves computing the probability of the rejection region when the probability distribution is a noncentral F. Davies (1980) gives an algorithm for making these and more general computations.

We now prove a theorem about central F distributions that will be useful in Chapter 5.

Theorem C.4. If $s > t$, then $sF(1-\alpha,s,v) \geq tF(1-\alpha,t,v)$.

PROOF. Let $X \sim \chi^2(s)$, $Y \sim \chi^2(t)$, and $Z \sim \chi^2(v)$. Let Z be independent of X and Y. Note that $(X/s)/(Z/v)$ has an $F(s,v)$ distribution; so $sF(1-\alpha,s,v)$ is the $100(1-\alpha)$ percentile of the distribution of $X/(Z/v)$. Similarly, $tF(1-\alpha,t,v)$ is the $100(1-\alpha)$ percentile of the distribution of $Y/(Z/v)$.

We will first argue that to prove the theorem it is enough to show that

$$\Pr[X \le d] \le \Pr[Y \le d] \tag{1}$$

for all real numbers d. We will then show that (1) is true.

If (1) is true, if c is any real number, and if $Z = z$, by independence we have

$$\Pr[X \le cz/v] = \Pr[X \le cz/v | Z = z] \le \Pr[Y \le cz/v | Z = z] = \Pr[Y \le cz/v].$$

Taking expectations with respect to Z,

$$\begin{aligned}\Pr\left[X\big/(Z/v) \le c\right] &= \mathrm{E}\left(\Pr[X \le cz/v | Z = z]\right) \\ &\le \mathrm{E}\left(\Pr[Y \le cz/v | Z = z]\right) \\ &= \Pr\left[Y\big/(Z/v) \le c\right].\end{aligned}$$

Since the cumulative distribution function (cdf) for $X\big/(Z/v)$ is always no greater than the cdf for $Y\big/(Z/v)$, the point at which a probability of $1-\alpha$ is attained for $X\big/(Z/v)$ must be no less than the similar point for $Y\big/(Z/v)$. Therefore,

$$sF(1-\alpha, s, v) \ge tF(1-\alpha, t, v).$$

To see that (1) holds, let Q be independent of Y and $Q \sim \chi^2(s-t)$. Then, because Q is nonnegative,

$$\Pr[X \le d] = \Pr[Y + Q \le d] \le \Pr[Y \le d]. \qquad \square$$

Exercise

Definition C.5. Consider two random variables W_1 and W_2. W_2 is said to be *stochastically larger* than W_1 if for every real number w

$$\Pr[W_1 > w] \le \Pr[W_2 > w].$$

If for some random variables W_1 and W_2, W_2 is stochastically larger than W_1, then we also say that the distribution of W_2 is stochastically larger than the distribution of W_1.

Exercise C.1 Show that a noncentral chi-squared distribution is stochastically larger than the central chi-squared distribution with the same degrees of freedom. Show that a noncentral F distribution is stochastically larger than the corresponding central F distribution.

Appendix D
Multivariate Distributions

Let $(x_1,\ldots,x_n)'$ be a random vector. The joint cumulative distribution function (cdf) of $(x_1,\ldots,x_n)'$ is

$$F(u_1,\ldots,u_n) \equiv \Pr[x_1 \leq u_1,\ldots,x_n \leq u_n].$$

If $F(u_1,\ldots,u_n)$ is the cdf of a discrete random variable, we can define a (joint) probability mass function

$$f(u_1,\ldots,u_n) \equiv \Pr[x_1 = u_1,\ldots,x_n = u_n].$$

If $F(u_1,\ldots,u_n)$ admits the nth order mixed partial derivative, then we can define a (joint) density function

$$f(u_1,\ldots,u_n) \equiv \frac{\partial^n}{\partial u_1 \cdots \partial u_n} F(u_1,\ldots,u_n).$$

The cdf can be recovered from the density as

$$F(u_1,\ldots,u_n) = \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_n} f(w_1,\ldots,w_n) dw_1 \cdots dw_n.$$

For a function $g(\cdot)$ of $(x_1,\ldots,x_n)'$ into $\mathbf{R}$, the expected value is defined as

$$\mathrm{E}[g(x_1,\ldots,x_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u_1,\ldots,u_n) f(u_1,\ldots,u_n) du_1 \cdots du_n.$$

We now consider relationships between two random vectors, say $x = (x_1,\ldots,x_n)'$ and $y = (y_1,\ldots,y_m)'$. Assume that the joint vector $(x',y')' = (x_1,\ldots,x_n,y_1,\ldots,y_m)'$ has a density function

$$f_{x,y}(u,v) \equiv f_{x,y}(u_1,\ldots,u_n,v_1,\ldots,v_m).$$

Similar definitions and results hold if $(x',y')'$ has a probability mass function.

The distribution of one random vector, say x, ignoring the other vector, y, is called the *marginal distribution* of x. The marginal cdf of x can be obtained by substituting the value $+\infty$ into the joint cdf for all of the y variables:

$$F_x(u) = F_{x,y}(u_1, \ldots, u_n, +\infty, \ldots, +\infty).$$

The marginal density can be obtained either by partial differentiation of $F_x(u)$ or by integrating the joint density over the y variables:

$$f_x(u) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{x,y}(u_1, \ldots, u_n, v_1, \ldots, v_m) dv_1 \cdots dv_m.$$

The conditional density of a vector, say x, given the value of the other vector, say $y = v$, is obtained by dividing the density of $(x', y')'$ by the density of y evaluated at v, i.e.,

$$f_{x|y}(u|v) \equiv f_{x,y}(u, v) \big/ f_y(v).$$

The conditional density is a well-defined density, so expectations with respect to it are well defined. Let g be a function from $\mathbf{R}^n$ into $\mathbf{R}$,

$$\mathrm{E}[g(x)|y = v] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x|y}(u|v) du,$$

where $du = du_1 du_2 \cdots du_n$. The standard properties of expectations hold for conditional expectations. For example, with a and b real,

$$\mathrm{E}[ag_1(x) + bg_2(x)|y = v] = a\mathrm{E}[g_1(x)|y = v] + b\mathrm{E}[g_2(x)|y = v].$$

The conditional expectation of $\mathrm{E}[g(x)|y = v]$ is a function of the value v. Since y is random, we can consider $\mathrm{E}[g(x)|y = v]$ as a random variable. In this context we write $\mathrm{E}[g(x)|y]$. An important property of conditional expectations is

$$\mathrm{E}[g(x)] = \mathrm{E}[\mathrm{E}[g(x)|y]].$$

To see this, note that $f_{x|y}(u|v) f_y(v) = f_{x,y}(u, v)$ and

$$\begin{aligned}
\mathrm{E}[\mathrm{E}[g(x)|y]] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{E}[g(x)|y = v] f_y(v) dv \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x|y}(u|v) du \right] f_y(v) dv \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x|y}(u|v) f_y(v) du\, dv \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) f_{x,y}(u, v) du\, dv \\
&= \mathrm{E}[g(x)].
\end{aligned}$$

In fact, both the notion of conditional expectation and this result can be generalized. Consider a function $g(x,y)$ from $\mathbf{R}^{n+m}$ into $\mathbf{R}$. If $y = v$, we can define $\mathrm{E}[g(x,y)|y = v]$ in a natural manner. If we consider y as random, we write $\mathrm{E}[g(x,y)|y]$. It can be easily shown that

$$\mathrm{E}[g(x,y)] = \mathrm{E}[\mathrm{E}[g(x,y)|y]]\,.$$

A function of x or y alone can also be considered as a function from $\mathbf{R}^{n+m}$ into $\mathbf{R}$.

A second important property of conditional expectations is that if $h(y)$ is a function from $\mathbf{R}^m$ into $\mathbf{R}$, we have

$$\mathrm{E}[h(y)g(x,y)|y] = h(y)\mathrm{E}[g(x,y)|y]\,. \tag{1}$$

This follows because if $y = v$,

$$\begin{aligned}\mathrm{E}[h(y)g(x,y)|y = v] &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(v)g(u,v)f_{x|y}(u|v)du \\ &= h(v)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(u,v)f_{x|y}(u|v)du \\ &= h(v)\mathrm{E}[g(x,y)|y = v]\,.\end{aligned}$$

This is true for all v, so (1) holds. In particular, if $g(x,y) \equiv 1$, we get

$$\mathrm{E}[h(y)|y] = h(y).$$

Finally, we can extend the idea of conditional expectation to a function $g(x,y)$ from $\mathbf{R}^{n+m}$ into $\mathbf{R}^s$. Write $g(x,y) = [g_1(x,y),\ldots,g_s(x,y)]'$. Then define

$$\mathrm{E}[g(x,y)|y] = (\mathrm{E}[g_1(x,y)|y]\,,\ldots,\mathrm{E}[g_s(x,y)|y])'\,.$$

If their densities exist, two random vectors are *independent* if and only if their joint density is equal to the product of their marginal densities, i.e., x and y are independent if and only if

$$f_{x,y}(u,v) = f_x(u)f_y(v).$$

Note that if x and y are independent,

$$f_{x|y}(u|v) = f_x(u).$$

If the random vectors x and y are independent, then any vector-valued functions of them, say g(x) and h(y), are also independent. This follows easily from a more general definition of the independence of two random vectors: The random vectors x and y are independent if for any two (reasonable) sets A and B,

$$\Pr[x \in A, y \in B] = \Pr[y \in A]\Pr[y \in B].$$

To prove that functions of random variables are independent, recall that the set inverse of a function $g(u)$ on a set A_0 is $g^{-1}(A_0) = \{u|g(u) \in A_0\}$. That $g(x)$ and $h(y)$ are independent follows from the fact that for any (reasonable) sets A_0 and B_0,

$$\begin{aligned}\Pr[g(x) \in A_0, h(y) \in B_0] &= \Pr[x \in g^{-1}(A_0), y \in h^{-1}(B_0)] \\ &= \Pr[x \in g^{-1}(A_0)] \Pr[y \in h^{-1}(B_0)] \\ &= \Pr[g(x) \in A_0] \Pr[h(y) \in B_0].\end{aligned}$$

The *characteristic function* of a random vector $x = (x_1, \ldots, x_n)'$ is a function from $\mathbf{R}^n$ to $\mathbf{C}$, the complex numbers. It is defined by

$$\varphi_x(t_1, \ldots, t_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[i \sum_{j=1}^{n} t_j u_j \right] f_x(u_1, \ldots, u_n) du_1 \cdots du_n.$$

We are interested in characteristic functions because if $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_n)'$ are random vectors and if

$$\varphi_x(t_1, \ldots, t_n) = \varphi_y(t_1, \ldots, t_n)$$

for all $(t_1, \ldots, t_n)$, then x and y have the same distribution.

For convenience, we have assumed the existence of densities. With minor modifications, the definitions and results of this appendix hold for any probability defined on $\mathbf{R}^n$.

Exercise

Exercise D.1 Let x and y be independent. Show that

(a) $\mathrm{E}[g(x)|y] = \mathrm{E}[g(x)]$;

(b) $\mathrm{E}[g(x)h(y)] = \mathrm{E}[g(x)]\,\mathrm{E}[h(y)]$.

Appendix E
Inference for One Parameter

Since the third edition of this book, I have thought hard about the philosophy of testing as a basis for non-Bayesian statistical inference, cf. Christensen (2005, 2008). This appendix has been modified accordingly. The approach taken is one I call Fisherian, as opposed to the Neyman–Pearson approach. The theory presented here has no formal role for alternative hypotheses.

A statistical testing problem is essentially a form of proof by contradiction. We have a *null model* for the data and we determine whether the observed data seem to contradict that null model or whether they are consistent with it. If the data contradict the null model, something must be wrong with the null model. Having data consistent with the null model certainly does not suggest that the null model is correct but may suggest that the model is tentatively adequate. The catch is that we rarely get an absolute contradiction to the null model, so we use probability to determine the extent to which the data seem inconsistent with the null model.

In the current discussion, *it is convenient to break the null model into two parts: a general model for the data and a particular statement about a single parameter of interest, called the null hypothesis (H_0).*

Many statistical tests and confidence intervals for a single parameter are applications of the same theory. (Tests and confidence intervals for variances are an exception.) To use this theory we need to know four things: [1] The unobservable parameter of interest (*Par*). [2] The estimate of the parameter (*Est*). [3] The standard error of the estimate (SE(*Est*)), wherein SE(*Est*) is typically an estimate of the standard deviation of *Est*, but if we happened to know the actual standard deviation, we would be happy to use it. And [4] an appropriate *reference distribution*. Specifically, we need the distribution of

$$\frac{Est - Par}{\mathrm{SE}(Est)}.$$

If the SE(*Est*) is estimated, the reference distribution is usually the t distribution with some known number of degrees of freedom df, say, $t(df)$. If the SE(*Est*) is known, then the distribution is usually the standard normal distribution, i.e., a $t(\infty)$.

In some problems (e.g., problems involving the binomial distribution) large sample results are used to get an approximate distribution and then the technique proceeds as if the approximate distribution were correct. When appealing to large sample results, the known distribution of part [4] is the standard normal (although I suspect that a $t(df)$ distribution with a reasonable, finite number of degrees of freedom would give more realistic results).

These four required items are derived from the model for the data (although sometimes the standard error incorporates the null hypothesis). For convenience, we may refer to these four items as "the model."

The $1-\alpha$ percentile of a distribution is the point that cuts off the top α of the distribution. For a t distribution, denote this $t(1-\alpha, df)$ as seen in Figure E.1. Formally, we can write

$$\Pr\left[\frac{Est - Par}{\mathrm{SE}(Est)} \geq t(1-\alpha, df)\right] = \alpha.$$

By symmetry about zero, we also have

$$\Pr\left[\frac{Est - Par}{\mathrm{SE}(Est)} \leq -t(1-\alpha, df)\right] = \alpha.$$

To keep the discussion as simple as possible, numerical examples have been restricted to one-sample normal theory. However, the results also apply to inferences on each individual mean and the difference between the means in two-sample problems, contrasts in analysis of variance, coefficients in regression, and, in general, to one-dimension estimable parametric functions in arbitrary linear models.

E.1 Testing

We want to test the null hypothesis

$$H_0 : Par = m,$$

where m is some known number. In *significance (Fisherian) testing*, we cannot do that. *What we can do* is test the null model, which is the combination of the model and the null hypothesis. The test is based on the assumption that both the model and H_0 are true. As mentioned earlier, it is rare that data contradict the null model absolutely, so we check to see if the data seem inconsistent with the null model.

What kind of data are inconsistent with the null model? Consider the *test statistic*

$$\frac{Est - m}{\mathrm{SE}(Est)}.$$

With m known, the test statistic is an observable random variable. If the null model is true, the test statistic has a known $t(df)$ distribution as illustrated in Figure E.1. The

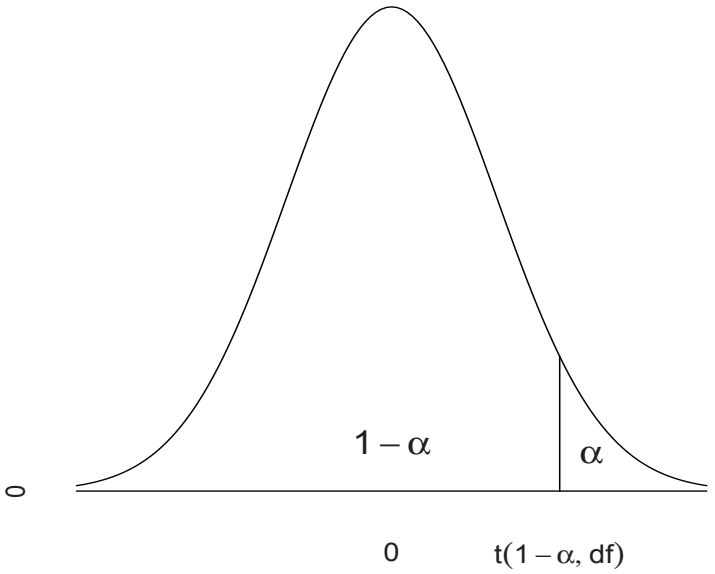

Fig. E.1 Percentiles of $t(df)$ distributions.

$t(df)$ distribution is likely to give values near 0 and is increasingly less likely to give values far from 0. Therefore, weird data, i.e., those that are most inconsistent with the null model, are large positive and large negative values of $[Est - m]/\mathrm{SE}(Est)$. The density (shape) of the $t(df)$ distribution allows us to order the possible values of the test statistic in terms of how weird they are relative to the null model.

To decide on a formal test, we need to decide which values of the test statistic will cause us to reject the null model and which will not. In other words, "How weird must data be before we question the null model?" We solve this problem by picking a small probability α that determines a *rejection region*, sometimes called a *critical region*. The rejection region consists of the weirdest test statistic values under the null model, but is restricted to have a probability of only α under the null model. Since a $t(df)$ distribution is symmetric about 0 and the density decreases as we go away from 0, the α critical region consists of points less than $-t(1-\alpha/2, df)$ and points larger than $t(1-\alpha/2, df)$. In other words, the α level test for the model with $H_0 : Par = m$ is to reject the null model if

$$\frac{Est - m}{\mathrm{SE}(Est)} \geq t\left(1 - \frac{\alpha}{2}, df\right)$$

or if

$$\frac{Est - m}{\mathrm{SE}(Est)} \leq -t\left(1 - \frac{\alpha}{2}, df\right).$$

This is equivalent to rejecting the null model if

$$\frac{|Est - m|}{\mathrm{SE}(Est)} \geq t\left(1 - \frac{\alpha}{2}, df\right).$$

What causes us to reject the null model? Either having a true model that is so different from the null that the data look "weird," or having the null model true and getting unlucky with the data.

Observing weird data, i.e., data that are inconsistent with the null model, gives us cause to question the validity of the null model. Specifying a small α level merely ensures that everything in the rejection region really constitutes weird data. More properly, specifying a small α level is our means of determining what constitutes weird data. Although α can be viewed as a probability, it is better viewed as a measure of how weird the data must be relative to the null model before we will reject. We want α small so that we only reject the null model for data that are truly weird, but we do not want α so small that we fail to reject the null model even when very strange data occur.

Rejecting the null model means that *either* the null hypothesis *or* the model is deemed incorrect. Only if we are confident that the model is correct can we conclude that the null hypothesis is wrong. If we want to make conclusions about the null hypothesis, it is important to do everything possible to assure ourselves that the model is reasonable.

If we do not reject the null model, we merely have data that are consistent with the null model. That in no way implies that the null model is true. Many other models will also be consistent with the data. Typically, $Par = m + 0.00001$ fits the data about as well as the null model. Not rejecting the test does not imply that the null model is true any more than rejecting the null model implies that the underlying model is true.

EXAMPLE E.1. Suppose that 16 independent observations are taken from a normal population. Test $H_0 : \mu = 20$ with α level 0.01. The observed values of $\bar{y}_\cdot$ and s^2 were 19.78 and 0.25, respectively.

[1] $Par = \mu$,
[2] $Est = \bar{y}_\cdot$,
[3] $\mathrm{SE}(Est) = \sqrt{s^2/16}$. In this case, the $\mathrm{SE}(Est)$ is estimated.
[4] $[Est - Par]/\mathrm{SE}(Est) = [\bar{y}_\cdot - \mu]/\sqrt{s^2/16}$ has a $t(15)$ distribution.

With $m = 20$, the $\alpha = 0.01$ test is to reject the H_0 model if

$$|\bar{y}_\cdot - 20|/[s/4] \geq 2.947 = t(0.995, 15).$$

Having $\bar{y}_\cdot = 19.78$ and $s^2 = 0.25$, we reject if

$$\frac{|19.78-20|}{\sqrt{.25/16}} \geq 2.947.$$

Since $|19.78-20|/\sqrt{.25/16} = |-1.76|$ is less than 2.947, we do not reject the null model at the $\alpha = 0.01$ level.

Nobody actually does this! Or at least, nobody should do it. Although this procedure provides a philosophical basis for our statistical inferences, there are two other procedures, both based on this, that give uniformly more information. This procedure requires us to specify the model, the null hypothesis parameter value m, and the α level. For a fixed model and a fixed null parameter m, P values are more informative because they allow us to report test results for all α levels. Alternatively, for a fixed model and a fixed α level, confidence intervals report the values of all parameters that are consistent with the model and the data. (Parameter values that are inconsistent with the model and the data are those that would be rejected, assuming the model is true.) We now discuss these other procedures.

E.2 *P* values

The P value of a test is the probability under the null model of seeing data as weird or weirder than we actually saw. Weirdness is determined by the distribution of the test statistic. If the observed value of the test statistic from Section 1 is t_{obs}, then the P value is the probability of seeing data as far or farther from 0 than t_{obs}. In general, we do not know if t_{obs} will be positive or negative, but its distance from 0 is $|t_{obs}|$. The P value is the probability that a $t(df)$ distribution is less than or equal to $-|t_{obs}|$ or greater than or equal to $|t_{obs}|$.

In Example E.1, the value of the test statistic is -1.76. Since $t(0.95, 15) = 1.75$, the P value of the test is approximately (just smaller than) 0.10. An $\alpha = 0.10$ test would use the $t(0.95, 15)$ value.

It is not difficult to see that the P value is the α level at which the test would just barely be rejected. So if $P \leq \alpha$, the null model is rejected, and if $P > \alpha$, the data are deemed consistent with the null model. Knowing the P value lets us do all α level tests of the null model. In fact, historically and philosophically, P values come before α level tests. Rather than noticing that the α level test has this relationship with P values, it is more general to define the α level test as rejecting precisely when $P \leq \alpha$. We can then observe that, for our setup, the α level test has the form given in Section 1.

While an α level constitutes a particular choice about how weird the data must be before we decide to reject the null model, the P value measures the evidence against the null hypothesis. The smaller the P value, the more evidence against the null model.

E.3 Confidence Intervals

A $(1-\alpha)100\%$ *confidence interval (CI)* for *Par* is defined to be the set of all parameter values m that would not be rejected by an α level test. In Section 1 we gave the rule for when an α level test of $H_0 : Par = m$ rejects. Conversely, the null model will not be rejected if

$$-t\left(1-\frac{\alpha}{2},df\right) < \frac{Est-m}{\mathrm{SE}(Est)} < t\left(1-\frac{\alpha}{2},df\right). \tag{1}$$

Some algebra, given later, establishes that we do not reject the null model if and only if

$$Est - t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est) < m < Est + t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est). \tag{2}$$

This interval consists of all the parameter values m that are consistent with the data and the model as determined by an α level test. The endpoints of the CI can be written

$$Est \pm t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est).$$

On occasion (as with binomial data), when doing an α level test or a P value, we may let the standard error depend on the null hypothesis. To obtain a confidence interval using this approach, we need a standard error that does not depend on m.

EXAMPLE E.2. We have 10 independent observations from a normal population with variance 6. $\bar{y}.$ is observed to be 17. We find a 95% CI for μ, the mean of the population.

[1] $Par = \mu$,
[2] $Est = \bar{y}.$,
[3] $\mathrm{SE}(Est) = \sqrt{6/10}$. In this case, $\mathrm{SE}(Est)$ is known and not estimated.
[4] $[Est - Par]/\mathrm{SE}(Est) = [\bar{y}. - \mu]/\sqrt{6/10} \sim N(0,1) = t(\infty)$.

The confidence coefficient is $95\% = (1-\alpha)100\%$, so $1-\alpha = 0.95$ and $\alpha = 0.05$. The percentage point from the normal distribution that we require is $t\left(1-\frac{\alpha}{2},\infty\right) = t(0.975,\infty) = 1.96$. The limits of the 95% CI are, in general,

$$\bar{y}. \pm 1.96\sqrt{6/10}$$

or, since $\bar{y}. = 17$,

$$17 \pm 1.96\sqrt{6/10}.$$

The μ values in the interval $(15.48, 18.52)$ are consistent with the data and the normal random sampling model as determined by an $\alpha = 0.05$ test.

To see that statements (1) and (2) are algebraically equivalent, the argument runs as follows:

$$-t\left(1-\frac{\alpha}{2},df\right) < \frac{Est-m}{\mathrm{SE}(Est)} < t\left(1-\frac{\alpha}{2},df\right)$$

if and only if $-t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est) < Est-m < t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est)$;
if and only if $t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est) > -Est+m > -t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est)$;
if and only if $Est+t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est) > m > Est-t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est)$;
if and only if $Est-t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est) < m < Est+t\left(1-\frac{\alpha}{2},df\right)\mathrm{SE}(Est)$.

E.4 Final Comments on Significance Testing

The most arbitrary element in Fisherian testing is the choice of a test statistic. Although alternative hypotheses do not play a formal role in significance testing, interesting possible alternative hypotheses do inform the choice of test statistic.

For example, in linear models we often test a full model $Y = X\beta + e$ against a reduced model $Y = X_0\gamma + e$, with $e \sim N(0,\sigma^2 I)$ and $C(X_0) \subset C(X)$. Although we choose a test statistic based on comparing these models, the significance test is only a test of whether the data are consistent with the reduced model. Rejecting the F test does not suggest that the full model is correct, it only suggests that the reduced model is wrong. Nonetheless, it is of interest to see how the test behaves if the full model is correct. But models other than the full model can also cause the test to reject, see Appendix F, especially Section F.2. For example, it is of interest to examine the *power* of a test. The power of an α level test at some alternative model is the probability of rejecting the null model when the alternative model is true. But in significance testing, there is no thought of accepting any alternative model. Any number of things can cause the rejection of the null model. Similar comments hold for testing generalized linear models.

When testing a null model based on a single parameter hypothesis $H_0 : Par = m$, interesting possible alternatives include $Par \neq m$. Our test statistic is designed to be sensitive to these alternatives, but problems with the null model other than $Par \neq m$ can cause us to reject the null model.

In general, a test statistic can be any function of the data for which the distribution under the null model is known (or can be approximated). But finding a usable test statistic can be difficult. Having to choose between alternative test statistics for the same null model is something of a luxury. For example, to test the null model of equal means in a balanced one-way ANOVA, we can use either the F test of Chapter 4 or the Studentized range test of Section 5.4

Appendix F
Significantly Insignificant Tests

Philosophically, the test of a null model occurs almost in a vacuum. Either the data contradict the null model or they are consistent with it. The discussion of model testing in Section 3.2 largely assumes that the full model is true. While it is interesting to explore the behavior of the F test statistic when the full model is true, and indeed it is reasonable and appropriate to choose a test statistic that will work well when the full model is true, the act of rejecting the null model in no way implies that the full model is true. It is perfectly reasonable that the null (reduced) model can be rejected when the full model is false.

Throughout this book we have examined standard approaches to testing in which F tests are rejected only for large values. The rationale for this is based on the full model being true. We now examine the significance of small F statistics. Small F statistics can be caused by an unsuspected lack of fit or, when the mean structure of the reduced model is correct, they can be caused by not accounting for negatively correlated data or not accounting for heteroscedasticity. We also demonstrate that large F statistics can be generated by not accounting for positively correlated data or heteroscedasticity, even when the mean structure of the reduced model is correct.

Christensen (1995, 2005, 2008) argues that (non-Bayesian) testing should be viewed as an exercise in examining whether or not the data are consistent with a particular (predictive) model. While possible alternative hypotheses may drive the choice of a test statistic, any unusual values of the test statistic should be considered important. By this standard, perhaps the only general way to decide which values of the test statistic are unusual is to identify as unusual those values that have small probabilities or small densities under the model being tested.

The F test statistic is driven by the idea of testing the reduced model against the full model. However, given the test statistic, any unusual values of that statistic should be recognized as indicating data that are inconsistent with the model being tested. If the full model is true, values of F much larger than 1 are inconsistent with the reduced model. Values of F much larger than 1 are consistent with the full model but, as we shall see, they are consistent with other models as well. Similarly, values of F much smaller than 1 are also inconsistent with the reduced model and we will examine models that can generate small F statistics.

I have been hesitant to discuss what I think of as a Fisherian F test, since nobody actually performs them. (That includes me, because it is so much easier to use the reported P values provided by standard computer programs.) Although the test statistic comes from considering both the reduced (null) model and the full model, once the test statistic is chosen, the full model no longer plays a role. From Theorem 3.2.1(ii), if the reduced model is true,

$$F \equiv \frac{Y'(M-M_0)Y/r(M-M_0)}{Y'(I-M)Y/r(I-M)} \sim F(r(M-M_0), r(I-M), 0).$$

We use the density to define "weird" values of the F distribution. The smaller the density, the weirder the observation. Write $r_1 \equiv r(M-M_0)$ and $r_2 \equiv r(I-M)$, denote the density $g(f|r_1,r_2)$, and let F_{obs} denote the observed value of the F statistic. Since the P value of a test is the probability under the null model of seeing data as weird or weirder than we actually saw, and weirdness is defined by the density, the P value of the test is

$$P = \Pr[g(F|r_1,r_2) \le g(F_{obs}|r_1,r_2)],$$

wherein F_{obs} is treated as fixed and known. This is computed under the only distribution we have, the $F(r_1,r_2)$ distribution. An α level test is defined as rejecting the null model precisely when $P \le \alpha$.

If $r_1 > 2$, the $F(r_1,r_2)$ density has the familiar shape that starts at 0, rises to a maximum in the vicinity of 1, and drops back down to zero for large values. Unless F_{obs} happens to be the mode, there are two values $f_1 < f_2$ that have

$$g(F_{obs}|r_1,r_2) = g(f_1|r_1,r_2) = g(f_2|r_1,r_2).$$

(One of f_1 and f_2 will be F_{obs}.) In this case, the P value reduces to

$$P = \Pr[F \le f_1] + \Pr[F \ge f_2].$$

In other words, the Fisherian F test is a two-sided F test, rejecting both for very small and very large values of F_{obs}. For $r_1 = 1,2$, the Fisherian test agrees with the usual test because then the $F(r_1,r_2)$ density starts high and decreases as f gets larger.

I should also admit that there remain open questions about the appropriateness of using densities, rather than actual probabilities, to define the weirdness of observations. The remainder of this appendix is closely related to Christensen (2003).

F.1 Lack of Fit and Small F Statistics

The standard assumption in testing models is that there is a full model $Y = X\beta + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = \sigma^2 I$ that fits the data. We then test the adequacy of a reduced model $Y = X_0\gamma + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = \sigma^2 I$ in which $C(X_0) \subset C(X)$, cf. Section 3.2. Based on second moment arguments, the test statistic is a ratio of variance estimates.

We construct an unbiased estimate of σ^2, $Y'(I-M)Y/r(I-M)$, and another statistic $Y'(M-M_0)Y/r(M-M_0)$ that has $\mathrm{E}[Y'(M-M_0)Y/r(M-M_0)] = \sigma^2 + \beta'X'(M-M_0)X\beta/r(M-M_0)$. Under the assumed covariance structure, this second statistic is an unbiased estimate of σ^2 if and only if the reduced model is correct. The test statistic

$$F = \frac{Y'(M-M_0)Y/r(M-M_0)}{Y'(I-M)Y/r(I-M)}$$

is a (biased) estimate of

$$\frac{\sigma^2 + \beta'X'(M-M_0)X\beta/r(M-M_0)}{\sigma^2} = 1 + \frac{\beta'X'(M-M_0)X\beta}{\sigma^2 r(M-M_0)}.$$

Under the null model, F is an estimate of the number 1. When the full model is true, values of F much larger than 1 suggest that F is estimating something larger than 1, which suggests that $\beta'X'(M-M_0)X\beta/\sigma^2 r(M-M_0) > 0$, something that occurs if and only if the reduced model is false. The standard normality assumption leads to an exact central F distribution for the test statistic under the null model, so we are able to quantify how unusual it is to observe any F statistic greater than 1. Although the test is based on second moment considerations, under the normality assumption it is also the generalized likelihood ratio test, see Exercise 3.1, and a uniformly most powerful invariant test, see Lehmann (1986, Section 7.1).

In testing lack of fit, the same basic ideas apply except that we start with the (reduced) model $Y = X\beta + e$. The ideal situation would be to know that if $Y = X\beta + e$ has the wrong mean structure, then a model of the form

$$Y = X\beta + W\delta + e, \quad C(W) \perp C(X) \tag{1}$$

fits the data where assuming $C(W) \perp C(X)$ creates no loss of generality. Unfortunately, there is rarely anyone to tell us the true matrix W. Lack of fit testing is largely about constructing a full model, say, $Y = X_*\beta_* + e$ with $C(X) \subset C(X_*)$ based on reasonable assumptions about the nature of any lack of fit. The test for lack of fit is simply the test of $Y = X\beta + e$ against the constructed model $Y = X_*\beta_* + e$. Typically, the constructed full model involves somehow generalizing the structure already observed in $Y = X\beta + e$. Section 6.6 discusses the rationale for several choices of constructed full models. For example, the traditional lack of fit test for simple linear regression begins with the replication model $y_{ij} = \beta_0 + \beta_1 x_i + e_{ij}$, $i = 1, \ldots, a$, $j = 1, \ldots, N_i$. It then assumes $\mathrm{E}(y_{ij}) = f(x_i)$ for some function $f(\cdot)$, in other words, it assumes that the several observations associated with x_i have the same expected value. Making no additional assumptions leads to fitting the full model $y_{ij} = \mu_i + e_{ij}$ and the traditional lack of fit test. Another way to think of this traditional test views the reduced model relative to the one-way ANOVA as having only the linear contrast important. The traditional lack of fit test statistic becomes

$$F = \frac{SSTrts - SS(lin)}{a-2} \Big/ MSE, \tag{2}$$

where $SS(lin)$ is the sum of squares for the linear contrast. If there is no lack of fit in the reduced model, F should be near 1. If lack of fit exists because the more general mean structure of the one-way ANOVA fits the data better than the simple linear regression model, the F statistic tends to be larger than 1.

Unfortunately, if the lack of fit exists because of features that are not part of the original model, generalizing the structure observed in $Y = X\beta + e$ is often inappropriate. Suppose that the simple linear regression model is balanced, i.e., all $N_i = N$, that for each i the data are taken in time order $t_1 < t_2 < \cdots < t_N$, and that the lack of fit is due to the true model being

$$y_{ij} = \beta_0 + \beta_1 x_i + \delta t_j + e_{ij}, \quad \delta \neq 0. \tag{3}$$

Thus, depending on the sign of δ, the observations within each group are subject to an increasing or decreasing trend. Note that in this model, for fixed i, the $\mathrm{E}(y_{ij})$s are *not* the same for all j, thus invalidating the assumption of the traditional test. In fact, this causes the traditional lack of fit test to have a *small* F statistic. One way to see this is to view the problem in terms of a balanced two-way ANOVA. The true model (3) is a special case of the two-way ANOVA model $y_{ij} = \mu + \alpha_i + \eta_j + e_{ij}$ in which the only nonzero terms are the linear contrast in the α_is and the linear contrast in the η_js. Under model (3), the numerator of the statistic (2) gives an unbiased estimate of σ^2 because $SSTrts$ in (2) is $SS(\alpha)$ for the two-way model and the only nonzero α effect is being eliminated from the treatments. However, the mean squared error in the denominator of (2) is a weighted average of the error mean square from the two-way model and the mean square for the η_js in the two-way model. The sum of squares for the significant linear contrast in the η_js from model (3) is included in the error term of the lack of fit test (2), thus biasing the error term to estimate something larger than σ^2. In particular, the denominator has an expected value of $\sigma^2 + \delta^2 a \sum_{j=1}^{N} (t_j - \bar{t}.)^2 / a(N-1)$. Thus, if the appropriate model is (3), the statistic in (2) estimates $\sigma^2 / [\sigma^2 + \delta^2 a \sum_{j=1}^{N} (t_j - \bar{t}.)^2 / a(N-1)]$ which is a number that is less than 1. Values of F much smaller than 1, i.e., very near 0, are consistent with a lack of fit that exists within the groups of the one-way ANOVA. Note that in this balanced case, true models involving interaction terms, e.g., models like

$$y_{ij} = \beta_0 + \beta_1 x_i + \delta t_j + \gamma x_i t_j + e_{ij},$$

also tend to make the F statistic small if either $\delta \neq 0$ or $\gamma \neq 0$. Finally, if there exists lack of fit both between the groups of observations and within the groups, if can be very difficult to identify. For example, if $\beta_2 \neq 0$ and either $\delta \neq 0$ or $\gamma \neq 0$ in the true model

$$y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \delta t_j + \gamma x_i t_j + e_{ij},$$

there is both a traditional lack of fit between the groups (the significant $\beta_2 x_i^2$ term) and lack of fit within the groups ($\delta t_j + \gamma x_i t_j$). In this case, neither the numerator nor the denominator in (2) is an estimate of σ^2.

More generally, start with a model $Y = X\beta + e$. This is tested against a larger model $Y = X_*\beta_* + e$ with $C(X) \subset C(X_*)$, regardless of where the larger model comes

from. The F statistic is

$$F = \frac{Y'(M_* - M)Y/r(M_* - M)}{Y'(I - M_*)Y/r(I - M_*)}.$$

We assume that the true model is (1). The F statistic estimates 1 if the original model $Y = X\beta + e$ is correct. It estimates something greater than 1 if the larger model $Y = X_*\beta_* + e$ is correct, i.e., if $W\delta \in C(X)^{\perp}_{C(X_*)}$. F estimates something less than 1 if $W\delta \in C(X_*)^{\perp}$, i.e., if $W\delta$ is actually in the error space of the larger model, because then the numerator estimates σ^2 but the denominator estimates

$$\sigma^2 + \delta' W'(I - M_*)W\delta/r(I - M_*) = \sigma^2 + \delta' W' W\delta/r(I - M_*).$$

If $W\delta$ is in neither of $C(X)^{\perp}_{C(X_*)}$ nor $C(X_*)^{\perp}$, it is not clear how the test will behave because neither the numerator nor the denominator estimates σ^2. Christensen (1989, 1991) contains related discussion of these concepts.

The main point is that, when testing a full model $Y = X\beta + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = \sigma^2 I$ against a reduced model $Y = X_0\gamma + e$, $C(X_0) \subset C(X)$, if the F statistic is small, it suggests that $Y = X_0\gamma + e$ may suffer from lack of fit in which the lack of fit exists in the error space of $Y = X\beta + e$. We will see in the next section that other possible explanations for a small F statistic are the existence of "negative correlation" in the data or heteroscedasticity.

F.2 The Effect of Correlation and Heteroscedasticity on F Statistics

The test of a reduced model assumes that the full model $Y = X\beta + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = \sigma^2 I$ holds and tests the adequacy of a reduced model $Y = X_0\gamma + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = \sigma^2 I$, $C(X_0) \subset C(X)$. Rejecting the reduced model does not imply that the full model is correct. The mean structure of the reduced model may be perfectly valid, but the F statistic can become large or small because the assumed covariance structure is incorrect.

We begin with a concrete example, one-way ANOVA. Let $i = 1, \ldots, a$, $j = 1, \ldots, N$, and $n \equiv aN$. Consider a reduced model $y_{ij} = \mu + e_{ij}$ which in matrix terms we write $Y = J\mu + e$, and a full model $y_{ij} = \mu_i + e_{ij}$, which we write $Y = Z\gamma + e$. In matrix terms the usual one-way ANOVA F statistic is

$$F = \frac{Y'[M_Z - (1/n)J_n^n]Y/(a-1)}{Y'(I - M_Z)Y/a(N-1)}. \tag{1}$$

We now assume that the true model is $Y = J\mu + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = \sigma^2 V$ and examine the behavior of the F statistic (1).

For a homoscedastic balanced one-way ANOVA we want to characterize the concepts of overall positive correlation, positive correlation within groups, and positive

correlation for evaluating differences between groups. Consider first a simple example with $a = 2$, $N = 2$. The first two observations are a group and the last two are a group. Consider a covariance structure

$$V_1 = \begin{bmatrix} 1 & 0.9 & 0.1 & 0.09 \\ 0.9 & 1 & 0.09 & 0.1 \\ 0.1 & 0.09 & 1 & 0.9 \\ 0.09 & 0.1 & .9 & 1 \end{bmatrix}.$$

There is an overall positive correlation, high positive correlation between the two observations in each group, and weak positive correlation between the groups. A second example,

$$V_2 = \begin{bmatrix} 1 & 0.1 & 0.9 & 0.09 \\ 0.1 & 1 & 0.09 & 0.9 \\ 0.9 & 0.09 & 1 & 0.1 \\ 0.09 & 0.9 & 0.1 & 1 \end{bmatrix},$$

has an overall positive correlation but weak positive correlation between the two observations in each group, with high positive correlation between some observations in different groups.

We now make a series of definitions for homoscedastic balanced one-way ANOVA based on the projection operators in (1) and V. Overall positive correlation is characterized by $\mathrm{Var}(\bar{y}_{\cdot\cdot}) > \sigma^2/n$, which in matrix terms is written

$$n\frac{\mathrm{Var}(\bar{y}_{\cdot\cdot})}{\sigma^2} = \mathrm{tr}[(1/n)JJ'V] > \frac{1}{n}\mathrm{tr}(V)\mathrm{tr}[(1/n)JJ'] = \frac{1}{n}\mathrm{tr}(V). \tag{2}$$

Overall negative correlation is characterized by the reverse inequality. For homoscedastic models the term $\mathrm{tr}(V)/n$ is 1. For heteroscedastic models the term on the right is the average variance of the observations divided by σ^2.

Positive correlation within groups is characterized by $\sum_{i=1}^{a}\mathrm{Var}(\bar{y}_{i\cdot})/a > \sigma^2/N$, which in matrix terms is written

$$\sum_{i=1}^{a} N\frac{\mathrm{Var}(\bar{y}_{i\cdot})}{\sigma^2} = \mathrm{tr}[M_Z V] > \frac{1}{n}\mathrm{tr}(V)\mathrm{tr}[M_Z] = \frac{a}{n}\mathrm{tr}(V). \tag{3}$$

Negative correlation within groups is characterized by the reverse inequality.

Positive correlation for evaluating differences between groups is characterized by

$$\frac{\sum_{i=1}^{a}\mathrm{Var}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}{a} > \frac{a-1}{a}\frac{\sigma^2}{N}.$$

Note that equality obtains if $V = I$. In matrix terms, this is written

$$\frac{N}{\sigma^2}\sum_{i=1}^{a}\mathrm{Var}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = \mathrm{tr}([M_Z - (1/n)JJ']V) > \frac{1}{n}\mathrm{tr}(V)\mathrm{tr}[M_Z - (1/n)JJ'] = \frac{a-1}{n}\mathrm{tr}(V) \tag{4}$$

and negative correlation for evaluating differences between groups is characterized by the reverse inequality. If all the observations in different groups are uncorrelated, there will be positive correlation for evaluating differences between groups if and only if there is positive correlation within groups. This follows because having a block diagonal covariance matrix $\sigma^2 V$ implies that $\mathrm{tr}(M_Z V) = \mathrm{tr}[(1/N)Z'VZ] = a\mathrm{tr}[(1/n)J'VJ] = a\mathrm{tr}[(1/n)JJ'V]$.

For our example V_1,

$$2.09 = (1/4)[4(2.09)] = \mathrm{tr}[(1/n)J_n^n V_1] > \frac{1}{n}\mathrm{tr}(V_1) = 4/4 = 1,$$

so there is an overall positive correlation,

$$3.8 = 2(1/2)[3.8] = \mathrm{tr}[M_Z V_1] > \frac{a}{n}\mathrm{tr}(V_1) = (2/4)4 = 2,$$

so there is positive correlation within groups, and

$$1.71 = 3.8 - 2.09 = \mathrm{tr}([M_Z - (1/n)J_n^n]V_1) > \frac{a-1}{n}\mathrm{tr}(V_1) = (1/4)4 = 1,$$

so there is positive correlation for evaluating differences between groups.

For the second example V_2,

$$2.09 = (1/4)[4(2.09)] = \mathrm{tr}[(1/n)J_n^n V_2] > \frac{1}{n}\mathrm{tr}(V_2) = 4/4 = 1,$$

so there is an overall positive correlation,

$$2.2 = 2(1/2)[2.2] = \mathrm{tr}[M_Z V_2] > \frac{a}{n}\mathrm{tr}(V_2) = (2/4)4 = 2,$$

so there is positive correlation within groups, but

$$0.11 = 2.2 - 2.09 = \mathrm{tr}([M_Z - (1/n)J_n^n]V_2) < \frac{a-1}{n}\mathrm{tr}(V_2) = (1/4)4 = 1,$$

so positive correlation for evaluating differences between groups does not exist.

The existence of positive correlation within groups and positive correlation for evaluating differences between groups causes the one-way ANOVA F statistic in (1) to get large even when there are no differences in the group means. Assuming that the correct model is $Y = J\mu + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = \sigma^2 V$, by Theorem 1.3.1, the numerator of the F statistic estimates

$$\mathrm{E}\{Y'[M_Z - (1/n)J_n^n]Y/(a-1)\} = \mathrm{tr}\{[M_Z - (1/n)J_n^n]V\}/(a-1) > \frac{a-1}{n}\mathrm{tr}(V)/(a-1) = \mathrm{tr}(V)/n$$

and the denominator of the F statistic estimates

$$\begin{aligned}
\mathrm{E}\{Y'(I-M_Z)Y/a(N-1)\} &= \mathrm{tr}\{[I-M_Z]V\}/a(N-1) \\
&= (\mathrm{tr}\{V\} - \mathrm{tr}\{[M_Z]V\})/a(N-1) \\
&< \left(\mathrm{tr}\{V\} - \frac{a}{n}\mathrm{tr}(V)\right)/a(N-1) \\
&= \frac{n-a}{n}\mathrm{tr}(V)/a(N-1) = \mathrm{tr}(V)/n.
\end{aligned}$$

In (1), F is an estimate of

$$\begin{aligned}
\frac{\mathrm{E}\{Y'[M_Z-(1/n)J_n^n]Y/(a-1)\}}{\mathrm{E}\{Y'(I-M_Z)Y/a(N-1)\}} &= \frac{\mathrm{tr}\{[M_Z-(1/n)J_n^n]V\}/(a-1)}{\mathrm{tr}\{[I-M_Z]V\}/a(N-1)} \\
&> \frac{\mathrm{tr}(V)/n}{\mathrm{tr}(V)/n} = 1,
\end{aligned}$$

so having both positive correlation within groups and positive correlation for evaluating differences between groups tends to make F statistics large. Exactly analogous computations show that both negative correlation within groups and negative correlation for evaluating differences between groups tends to make F statistics less than 1.

Another example elucidates some additional points. Suppose the observations have the AR(1) correlation structure discussed in Subsection 13.3.1:

$$V_3 = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

Using the same grouping structure as before, when $0 < \rho < 1$, we have overall positive correlation because

$$1 + \frac{\rho}{2}(3 + 2\rho + \rho^2) = \mathrm{tr}[(1/n)JJ'V_3] > 1,$$

and we have positive correlation within groups because

$$2(1+\rho) = \mathrm{tr}[M_Z V_3] > 2.$$

If $-1 < \rho < 0$, the inequalities are reversed. Similarly, for $-1 < \rho < 0$ we have negative correlation for evaluating differences between groups because

$$1 + \frac{\rho}{2}(1 - 2\rho - \rho^2)^2 = \mathrm{tr}([M_Z - (1/n)JJ']V_3) < 1.$$

However, we only get positive correlation for evaluating differences between groups when $0 < \rho < \sqrt{2} - 1$. Thus, for negative ρ we tend to get small F statistics, for $0 < \rho < \sqrt{2} - 1$ we tend to get large F statistics, and for $\sqrt{2} - 1 < \rho < 1$ the result is not clear.

To illustrate, suppose $\rho = 1$ and the observations all have the same mean, then with probability 1, all the observations are equal and, in particular, $\bar{y}_{i\cdot} = \bar{y}_{\cdot\cdot}$ with probability 1. It follows that

$$0 = \frac{\sum_{i=1}^{a} \text{Var}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}{a} < \frac{a-1}{a}\frac{\sigma^2}{N}$$

and no positive correlation exists for evaluating differences between groups. More generally, for very strong positive correlations, both the numerator and the denominator of the F statistic estimate numbers close to 0 and both are smaller than they would be under $V = I$. On the other hand, it is not difficult to see that, for $\rho = -1$, the F statistic is 0.

In the balanced heteroscedastic one-way ANOVA, V is diagonal. This generates equality between the left sides and right sides of (2), (3), and (4), so under heteroscedasticity F still estimates the number 1. We now generalize the ideas of within group correlation and correlation for evaluating differences between groups, and see that heteroscedasticity can affect unbalanced one-way ANOVA.

In general, we test a full model $Y = X\beta + e$, $\text{E}(e) = 0$, $\text{Cov}(e) = \sigma^2 I$ against a reduced model $Y = X_0\gamma + e$, in which $C(X_0) \subset C(X)$. We examine the F statistic when the true model is $Y = X_0\gamma + e$, $\text{E}(e) = 0$, $\text{Cov}(e) = \sigma^2 V$. Using arguments similar to those for balanced one-way ANOVA, having

$$\text{tr}[MV] > \frac{1}{n}\text{tr}(V)\text{tr}[M] = \frac{r(X)}{n}\text{tr}(V)$$

and

$$\text{tr}([M - M_0]V) > \frac{1}{n}\text{tr}(V)\text{tr}[M - M_0] = \frac{r(X) - r(X_0)}{n}\text{tr}(V)$$

causes large F statistics even when the mean structure of the reduced model is true, and reversing the inequalities causes small F statistics. These are merely sufficient conditions so that the tests intuitively behave certain ways. The actual behavior of the tests under normal distributions can be determined numerically, cf. Christensen and Bedrick (1997).

These covariance conditions can be caused by patterns of positive and negative correlations as discussed earlier, but they can also be caused by heteroscedasticity. For example, consider the behavior of the unbalanced one-way ANOVA F test when the observations are uncorrelated but heteroscedastic. For concreteness, assume that $\text{Var}(y_{ij}) = \sigma_i^2$. Because the observations are uncorrelated, we need only check the condition

$$\text{tr}[MV] \equiv \text{tr}[M_Z V] > \frac{1}{n}\text{tr}(V)\text{tr}[M_Z] = \frac{a}{n}\text{tr}(V),$$

which amounts to

$$\sum_{i=1}^{a} \sigma_i^2 / a > \sum_{i=1}^{a} \frac{N_i}{n}\sigma_i^2.$$

Thus, when the groups' means are equal, F statistics will get large if many observations are taken in groups with small variances and few observations are taken on

groups with large variances. F statistics will get small if the reverse relationship holds.

The general condition

$$\text{tr}[MV] > \frac{1}{n}\text{tr}(V)\text{tr}[M] = \frac{r(X)}{n}\text{tr}(V)$$

is equivalent to

$$\frac{\sum_{i=1}^{n}\text{Var}(x_i'\hat{\beta})}{r(X)} > \frac{\sum_{i=1}^{n}\text{Var}(y_i)}{n}.$$

So, under homoscedasticity, positive correlation in the full model amounts to having an average variance for the predicted values (averaging over the rank of the covariance matrix of the predicted values) that is larger than the common variance of the observations. Negative correlation in the full model involves reversing the inequality. Similarly, having positive correlation for distinguishing the full model from the reduced model means

$$\frac{\sum_{i=1}^{n}\text{Var}(x_i'\hat{\beta} - x_{0i}'\hat{\gamma})}{r(X) - r(X_0)} = \frac{\text{tr}[(M - M_0)V]}{r(M - M_0)} > \frac{\text{tr}(V)}{n} = \frac{\sum_{i=1}^{n}\text{Var}(y_i)}{n}.$$

Appendix G
Randomization Theory Models

The division of labor in statistics has traditionally designated randomization theory as an area of nonparametric statistics. Randomization theory is also of special interest in the theory of experimental design because randomization has been used to justify the analysis of designed experiments.

It can be argued that the linear models given in Chapter 8 are merely good approximations to more appropriate models based on randomization theory. One aspect of this argument is that the F tests based on the theory of normal errors are a good approximation to randomization (permutation) tests. Investigating this is beyond the scope of a linear models book, cf. Hinkelmann and Kempthorne (1994) and Puri and Sen (1971). Another aspect of the approximation argument is that the BLUEs under randomization theory are precisely the least squares estimates. By Theorem 10.4.5, to establish this we need to show that $C(VX) \subset C(X)$ for the model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = V,$$

where V is the covariance matrix under randomization theory. This argument will be examined here for two experimental design models: the model for a completely randomized design and the model for a randomized complete block design. First, we introduce the subject with a discussion of simple random sampling.

G.1 Simple Random Sampling

Randomization theory for a simple random sample assumes that observations y_i are picked at random (without replacement) from a larger finite population. Suppose

the elements of the population are $s_1, s_2, \ldots, s_N$. We can define elementary sampling random variables for $i = 1, \ldots, n$ and $j = 1, \ldots, N$,

$$\delta_j^i = \begin{cases} 1, & \text{if } y_i = s_j \\ 0, & \text{otherwise.} \end{cases}$$

Under simple random sampling without replacement

$$\mathrm{E}[\delta_j^i] = \Pr[\delta_j^i = 1] = \frac{1}{N}.$$

$$\mathrm{E}[\delta_j^i \delta_{j'}^{i'}] = \Pr[\delta_j^i \delta_{j'}^{i'} = 1] = \begin{cases} 1/N, & \text{if } (i,j) = (i',j') \\ 1/N(N-1), & \text{if } i \neq i' \text{ and } j \neq j' \\ 0, & \text{otherwise.} \end{cases}$$

If we write $\mu = \sum_{j=1}^N s_j/N$ and $\sigma^2 = \sum_{j=1}^N (s_j - \mu)^2/N$, then

$$y_i = \sum_{j=1}^N \delta_j^i s_j = \mu + \sum_{j=1}^N \delta_j^i (s_j - \mu).$$

Letting $e_i = \sum_{j=1}^N \delta_j^i (s_j - \mu)$ gives the linear model

$$y_i = \mu + e_i.$$

The population mean μ is a fixed unknown constant. The e_is have the properties

$$\mathrm{E}[e_i] = \mathrm{E}\left[\sum_{j=1}^N \delta_j^i (s_j - \mu)\right] = \sum_{j=1}^N \mathrm{E}\left[\delta_j^i\right](s_j - \mu) = \sum_{j=1}^N (s_j - \mu)/N = 0,$$

$$\mathrm{Var}(e_i) = \mathrm{E}[e_i^2] = \sum_{j=1}^N \sum_{j'=1}^N (s_j - \mu)(s_{j'} - \mu)\mathrm{E}[\delta_j^i \delta_{j'}^i] = \sum_{j=1}^N (s_j - \mu)^2/N = \sigma^2.$$

For $i \neq i'$,

$$\begin{aligned} \mathrm{Cov}(e_i, e_{i'}) &= \mathrm{E}[e_i e_{i'}] = \sum_{j=1}^N \sum_{j'=1}^N (s_j - \mu)(s_{j'} - \mu)\mathrm{E}[\delta_j^i \delta_{j'}^{i'}] \\ &= [N(N-1)]^{-1} \sum_{j \neq j'} (s_j - \mu)(s_{j'} - \mu) \\ &= [N(N-1)]^{-1} \left(\left[\sum_{j=1}^N (s_j - \mu) \right]^2 - \sum_{j=1}^N (s_j - \mu)^2 \right) \\ &= -\sigma^2/(N-1). \end{aligned}$$

In matrix terms, the linear model can be written

$$Y = J\mu + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 V,$$

where

$$V = \begin{bmatrix} 1 & -(N-1)^{-1} & -(N-1)^{-1} & \cdots & -(N-1)^{-1} \\ -(N-1)^{-1} & 1 & -(N-1)^{-1} & \cdots & -(N-1)^{-1} \\ -(N-1)^{-1} & -(N-1)^{-1} & 1 & \cdots & -(N-1)^{-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -(N-1)^{-1} & -(N-1)^{-1} & -(N-1)^{-1} & \cdots & 1 \end{bmatrix}.$$

Clearly $VJ = [(N-n)/(N-1)]J$, so the BLUE of μ is $\bar{y}_{\cdot\cdot}$.

G.2 Completely Randomized Designs

Suppose that there are t treatments, each to be randomly assigned to N units out of a collection of $n = tN$ experimental units. A one-way ANOVA model for this design is

$$y_{ij} = \mu_i + e_{ij}, \tag{1}$$

$i = 1, \ldots, t$, $j = 1, \ldots, N$. Suppose further that the ith treatment has an effect τ_i and that the experimental units without treatment effects would have readings $s_1, \ldots, s_n$. The elementary sampling random variables are

$$\delta_k^{ij} = \begin{cases} 1, & \text{if replication } j \text{ of treatment } i \text{ is assigned to unit } k \\ 0, & \text{otherwise.} \end{cases}$$

With this restricted random sampling,

$$\mathrm{E}[\delta_k^{ij}] = \Pr[\delta_k^{ij} = 1] = \frac{1}{n}$$

$$\mathrm{E}[\delta_k^{ij}\delta_{k'}^{i'j'}] = \Pr[\delta_k^{ij}\delta_{k'}^{i'j'} = 1] = \begin{cases} 1/n, & \text{if } (i,j,k) = (i',j',k') \\ 1/n(n-1), & \text{if } k \neq k' \text{ and } (i,j) \neq (i',j)' \\ 0, & \text{otherwise.} \end{cases}$$

We can write

$$y_{ij} = \tau_i + \sum_{k=1}^{n} \delta_k^{ij} s_k.$$

Taking $\mu = \sum_{k=1}^{n} s_k/n$ and $\mu_i = \mu + \tau_i$ gives

$$y_{ij} = \mu_i + \sum_{k=1}^{n} \delta_k^{ij}(s_k - \mu).$$

To obtain the linear model (1), let $e_{ij} = \sum_{k=1}^{n} \delta_k^{ij}(s_k - \mu)$. Write $\sigma^2 = \sum_{k=1}^{n}(s_k - \mu)^2/n$. Then

$$\mathrm{E}[e_{ij}] = \mathrm{E}\left[\sum_{k=1}^{n} \delta_k^{ij}(s_k - \mu)\right] = \sum_{k=1}^{n} \mathrm{E}\left[\delta_k^{ij}\right](s_k - \mu) = \sum_{k=1}^{n} (s_k - \mu)/n = 0,$$

$$\mathrm{Var}(e_{ij}) = \mathrm{E}[e_{ij}^2] = \sum_{k=1}^{n}\sum_{k'=1}^{n} (s_k - \mu)(s_{k'} - \mu)\mathrm{E}[\delta_k^{ij}\delta_{k'}^{ij}] = \sum_{k=1}^{n} (s_k - \mu)^2/n = \sigma^2.$$

For $(i,j) \neq (i',j')$,

$$\begin{aligned}
\mathrm{Cov}(e_{ij}, e_{i'j'}) &= \mathrm{E}[e_{ij}e_{i'j'}] = \sum_{k=1}^{n}\sum_{k'=1}^{n} (s_k - \mu)(s_{k'} - \mu)\mathrm{E}[\delta_k^{ij}\delta_{k'}^{i'j'}] \\
&= [n(n-1)]^{-1} \sum_{k \neq k'} (s_k - \mu)(s_{k'} - \mu) \\
&= [n(n-1)]^{-1} \left(\left[\sum_{k=1}^{n} (s_k - \mu)\right]^2 - \sum_{k=1}^{n} (s_k - \mu)^2 \right) \\
&= -\sigma^2/(n-1).
\end{aligned}$$

In matrix terms, writing $Y = (y_{11}, y_{12}, \ldots, y_{tN})'$, we get

$$Y = X \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_t \end{bmatrix} + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 V,$$

where

$$\begin{aligned}
V &= \begin{bmatrix}
1 & -1/(n-1) & -1/(n-1) & \cdots & -1/(n-1) \\
-1/(n-1) & 1 & -1/(n-1) & \cdots & -1/(n-1) \\
-1/(n-1) & -1/(n-1) & 1 & \cdots & -1/(n-1) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-1/(n-1) & -1/(n-1) & -1/(n-1) & \cdots & 1
\end{bmatrix} \\
&= \frac{n}{n-1} I - \frac{1}{n-1} J_n^n.
\end{aligned}$$

It follows that

$$VX = \frac{n}{n-1} X - \frac{1}{n-1} J_n^n X.$$

Since $J \in C(X)$, $C(VX) \subset C(X)$, and least squares estimates are BLUEs. Standard errors for estimable functions can be found as in Section 11.1 using the fact that this model involves only one cluster.

Exercise G.1 Establish whether least squares estimates are BLUEs in a completely randomized design with unequal numbers of observations on the treatments.

G.3 Randomized Complete Block Designs

Suppose there are a treatments and b blocks. The experimental units must be grouped into b blocks, each of a units. Let the experimental unit effects be s_{kj}, $k = 1,\ldots,a$, $j = 1,\ldots,b$. Treatments are assigned at random to the a units in each block. The elementary sampling random variables are

$$\delta_{kj}^i = \begin{cases} 1, & \text{if treatment } i \text{ is assigned to unit } k \text{ in block } j \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathrm{E}[\delta_{kj}^i] = \Pr[\delta_{kj}^i = 1] = \frac{1}{a}.$$

$$\mathrm{E}[\delta_{kj}^i \delta_{k'j'}^{i'}] = \Pr[\delta_{kj}^i \delta_{k'j'}^{i'} = 1] = \begin{cases} 1/a, & \text{if } (i,j,k) = (i',j',k') \\ 1/a^2, & \text{if } j \neq j' \\ 1/a(a-1), & \text{if } j = j',\ k \neq k',\ i \neq i' \\ 0, & \text{otherwise.} \end{cases}$$

If α_i is the additive effect of the ith treatment and $\beta_j \equiv \bar{s}_{\cdot j}$, then

$$y_{ij} = \alpha_i + \beta_j + \sum_{k=1}^{a} \delta_{kj}^i (s_{kj} - \beta_j).$$

Letting $e_{ij} = \sum_{k=1}^{a} \delta_{kj}^i (s_{kj} - \beta_j)$ gives the linear model

$$y_{ij} = \alpha_i + \beta_j + e_{ij}. \tag{1}$$

The column space of the design matrix for this model is precisely that of the model considered in Section 8.3. Let $\sigma_j^2 = \sum_{k=1}^{a} (s_{kj} - \beta_j)^2/a$. Then

$$\mathrm{E}[e_{ij}] = \sum_{k=1}^{a} (s_{kj} - \beta_j)/a = 0,$$

$$\begin{aligned} \mathrm{Var}(e_{ij}) &= \sum_{k=1}^{a} \sum_{k'=1}^{a} (s_{kj} - \beta_j)(s_{k'j} - \beta_j) \mathrm{E}[\delta_{kj}^i \delta_{k'j}^i] \\ &= \sum_{k=1}^{a} (s_{kj} - \beta_j)^2/a = \sigma_j^2. \end{aligned}$$

For $j \neq j'$,

$$\begin{aligned} \mathrm{Cov}(e_{ij}, e_{i'j'}) &= \sum_{k=1}^{a} \sum_{k'=1}^{a} (s_{kj} - \beta_j)(s_{k'j'} - \beta_{j'}) \mathrm{E}[\delta_{kj}^i \delta_{k'j'}^{i'}] \\ &= a^{-2} \sum_{k=1}^{a} (s_{kj} - \beta_j) \sum_{k'=1}^{a} (s_{k'j'} - \beta_{j'}) \end{aligned}$$

$$= 0.$$

For $j = j'$, $i \neq i'$,

$$\begin{aligned}
\mathrm{Cov}(e_{ij}, e_{i'j'}) &= \sum_{k=1}^{a}\sum_{k'=1}^{a} (s_{kj} - \beta_j)(s_{k'j} - \beta_j)\mathrm{E}[\delta_{kj}^{i}\delta_{k'j}^{i'}] \\
&= \sum_{k \neq k'} (s_{kj} - \beta_j)(s_{k'j} - \beta_j)/a(a-1) \\
&= [a(a-1)]^{-1}\left(\left[\sum_{k=1}^{a}(s_{kj} - \beta_j)\right]^2 - \sum_{k=1}^{a}(s_{kj} - \beta_j)^2\right) \\
&= -\sigma_j^2/(a-1).
\end{aligned}$$

Before proceeding, we show that although the terms β_j are not known, the differences among these are known constants under randomization theory. For any unit k in block j, some treatment is assigned, so $\sum_{i=1}^{a} \delta_{kj}^{i} = 1$.

$$\begin{aligned}
\bar{y}_{\cdot j} &= \frac{1}{a}\left[\sum_{i=1}^{a}\left(\alpha_i + \beta_j + \sum_{k=1}^{a}\delta_{kj}^{i}(s_{kj} - \beta_j)\right)\right] \\
&= \frac{1}{a}\left[\sum_{i=1}^{a}\alpha_i + a\beta_j + \sum_{k=1}^{a}(s_{kj} - \beta_j)\sum_{i=1}^{a}\delta_{kj}^{i}\right] \\
&= \bar{\alpha}_{\cdot} + \beta_j + \sum_{k=1}^{a}(s_{kj} - \beta_j) \\
&= \bar{\alpha}_{\cdot} + \beta_j\,.
\end{aligned}$$

Therefore, $\bar{y}_{\cdot j} - \bar{y}_{\cdot j'} = \beta_j - \beta_{j'} = \bar{s}_{\cdot j} - \bar{s}_{\cdot j'}$. Since these differences are fixed and known, there is no basis for a test of $H_0 : \beta_1 = \cdots = \beta_b$. In fact, the linear model is not just model (1) but model (1) subject to these estimable constraints on the βs.

To get best linear unbiased estimates we need to assume that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_b^2 = \sigma^2$. We can now write the linear model in matrix form and establish that least squares estimates of treatment means and contrasts in the α_is are BLUEs. In the discussion that follows, we use notation from Section 7.1. Model (1) can be rewritten

$$Y = X\eta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = V, \tag{2}$$

where $\eta = [\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b]'$. If we let X_2 be the columns of X corresponding to $\beta_1, \ldots, \beta_b$, then (cf. Section 11.1)

$$V = \sigma^2\,[a/(a-1)]\left[I - (1/a)X_2X_2'\right] = \sigma^2\,[a/(a-1)]\left[I - M_\mu - M_\beta\right].$$

If model (2) were the appropriate model, checking that $C(VX) \subset C(X)$ would be trivial based on the fact that $C(X_2) \subset C(X)$. However, we must account for the estimable constraints on the model discussed above. In particular, consider

$$M_\beta X\eta = [t_{ij}],$$

where

$$t_{ij} = \beta_j - \bar{\beta}_\cdot = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} = \bar{s}_{\cdot j} - \bar{s}_{\cdot\cdot}\,.$$

This is a fixed known quantity. Proceeding as in Section 3.3, the model is subject to the estimable constraint

$$M_\beta X\eta = M_\beta Y.$$

Normally a constraint has the form $\Lambda'\beta = d$, where d is known. Here $d = M_\beta Y$, which appears to be random but, as discussed, $M_\beta Y$ is not random; it is fixed and upon observing Y it is known.

The equivalent reduced model involves $X_0 = (I - M_{MP})X = (I - M_\beta)X$ and a known vector $Xb = M_\beta Y$. Thus, the constrained model is equivalent to

$$(Y - M_\beta Y) = (I - M_\beta)X\gamma + e. \tag{3}$$

We want to show that least squares estimates of contrasts in the αs based on Y are BLUEs with respect to this model. First we show that least squares estimates from model (3) based on $(Y - M_\beta Y) = (I - M_\beta)Y$ are BLUEs. We need to show that

$$C(V(I - M_\beta)X) = C[(I - M_\mu - M_\beta)(I - M_\beta)X] \subset C[(I - M_\beta)X].$$

Because $(I - M_\mu - M_\beta)(I - M_\beta) = (I - M_\mu - M_\beta)$, we have

$$C(V(I - M_\beta)X) = C[(I - M_\mu - M_\beta)X],$$

and because $C(I - M_\mu - M_\beta) \subset C(I - M_\beta)$ we have

$$C[(I - M_\mu - M_\beta)X] \subset C[(I - M_\beta)X].$$

To finish the proof that least squares estimates based on Y are BLUEs, note that the estimation space for model (3) is $C[(I - M_\beta)X] = C(M_\mu + M_\alpha)$. BLUEs are based on

$$(M_\mu + M_\alpha)(I - M_\beta)Y = (M_\mu + M_\alpha)Y.$$

Thus, any linear parametric function in model (2) that generates a constraint on $C(M_\mu + M_\alpha)$ has a BLUE based on $(M_\mu + M_\alpha)Y$ (cf. Exercise 3.9.5). In particular, this is true for contrasts in the αs. Standard errors for estimable functions are found in a manner analogous to Section 11.1. This is true even though model (3) is not the form considered in Section 11.1 and is a result of the orthogonality relationships that are present.

The assumption that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_b^2$ is a substantial one. Least squares estimates without this assumption are unbiased, but may be far from optimal. It is important to choose blocks so that their variances are approximately equal.

Exercise G.2 Find the standard error for a contrast in the α_is of model (1).

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, Second Edition. John Wiley and Sons, New York.

Andrews, D. F. (1974). A robust method for multiple regression. *Technometrics*, **16**, 523-531.

Arnold, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*. John Wiley and Sons, New York.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). *Journal of the Royal Statistical Society, Series B*, **44**, 1-36.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press, Oxford.

Atwood, C. L. and Ryan, T. A., Jr. (1977). A class of tests for lack of fit to a regression model. Unpublished manuscript.

Bailey, D. W. (1953). *The Inheritance of Maternal Influences on the Growth of the Rat*. Ph.D. Thesis, University of California.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, **91**, 1450-1460.

Belsley, D. A. (1991). *Collinearity Diagnostics: Collinearity and Weak Data in Regression*. John Wiley and Sons, New York.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons, New York.

Benedetti, J. K. and Brown, M. B. (1978). Strategies for the selection of log-linear models. *Biometrics*, **34**, 680-686.

Berger, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Revised Second Edition. Springer-Verlag, New York.

Berry, D. A. (1996). *Statistics: A Bayesian Perspective*. Duxbery, Belmont, CA.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates*. John Wiley and Sons, New York.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211-246.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley and Sons, New York.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. John Wiley and Sons, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, New York.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Second Edition. Springer-Verlag, New York.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, Second Edition. John Wiley and Sons, New York.

Casella, G. (2008). *Statistical Design.* Springer-Verlag, New York.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, Second Edition. Duxbury Press, Pacific Grove, CA.

Christensen, R. (1984). A note on ordinary least squares methods for two-stage sampling. *Journal of the American Statistical Association*, **79**, 720-721.

Christensen, R. (1987a). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer-Verlag, New York.

Christensen, R. (1987b). The analysis of two-stage sampling data by ordinary least squares. *Journal of the American Statistical Association*, **82**, 492-498.

Christensen, R. (1989). Lack of fit tests based on near or exact replicates. *The Annals of Statistics*, **17**, 673-683.

Christensen, R. (1991). Small sample characterizations of near replicate lack of fit tests. *Journal of the American Statistical Association*, **86**, 752-756.

Christensen, R. (1993). Quadratic covariance estimation and equivalence of predictions. *Mathematical Geology*, **25**, 541-558.

Christensen, R. (1995). Comment on Inman (1994). *The American Statistician*, **49**, 400.

Christensen, R. (1996a). *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall, London.

Christensen, R. (1996b). Exact tests for variance components. *Biometrics*, **52**, 309-315.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, Second Edition. Springer-Verlag, New York.

Christensen, R. (2001). *Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression, and Response Surface Maximization*, Second Edition. Springer-Verlag, New York.

Christensen, R. (2003). Significantly insignificant F tests. *The American Statistician*, **57**, 27-32.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, **59**, 121-126.

Christensen, R. (2008). Review of *Principals of Statistical Inference* by D. R. Cox. *Journal of the American Statistical Association*, **103**, 1719-1723.

Christensen, R. and Bedrick, E. J. (1997). Testing the independence assumption in linear models. *Journal of the American Statistical Association*, **92**, 1006-1016.

Christensen, R. and Bedrick, E. J. (1999). A survey of some new alternatives to Wald's variance component test. *Tatra Mountains Mathematical Publications*, **17**, 91-102.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Chapman and Hall/CRC Press, Boca Raton, FL.

Christensen, R., Johnson, W., and Pearson, L. M. (1992). Prediction diagnostics for spatial linear models. *Biometrika*, **79**, 583-591.

Christensen, R., Johnson, W., and Pearson, L. M. (1993). Covariance function diagnostics for spatial linear models. *Mathematical Geology*, **25**, 145-160.

Christensen, R. and Lin, Y. (2010). Perfect Estimation in Linear Models with Singular Covariance Matrices. Unpublished manuscript.

Christensen, R., Pearson, L. M., and Johnson, W. (1992). Case deletion diagnostics for mixed models. *Technometrics*, **34**, 38-45.

Christensen, R. and Utts, J. (1992). Testing for nonadditivity in log-linear and logit models. *Journal of Statistical Planning and Inference*, **33**, 333-343.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science*, **19**, 81-94.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, Second Edition. John Wiley and Sons, New York.
Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.
Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.
Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley and Sons, New York.
Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. John Wiley and Sons, New York.
Cornell, J. A. (1988). Analyzing mixture experiments containing process variables. A split plot approach. *Journal of Quality Technology*, **20**, 2-23.
Cox, D. R. (1958). *Planning of Experiments*. John Wiley and Sons, New York.
Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition. John Wiley and Sons, New York.
Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, **1**, 311-341.
Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. John Wiley and Sons, New York.
Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, Second Edition. John Wiley and Sons, New York.
Davies, R. B. (1980). The distribution of linear combinations of χ^2 random variables. *Applied Statistics*, **29**, 323-333.
de Finetti, B. (1974, 1975). *Theory of Probability*, Vols. 1 and 2. John Wiley and Sons, New York.
DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.
deLaubenfels, R. (2006). The victory of least squares and orthogonality in statistics. *The American Statistician*, **60**, 315-321.
Doob, J. L. (1953). *Stochastic Processes*. John Wiley and Sons, New York.
Draper, N. and Smith, H. (1998). *Applied Regression Analysis*, Third Edition. John Wiley and Sons, New York.
Duan, N. (1981). Consistency of residual distribution functions. Working Draft No. 801-1-HHS (106B-80010), Rand Corporation, Santa Monica, CA.
Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, **38**, 159-179.
Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. John Wiley and Sons, New York. Reprinted in 2007 by IMS Lecture Notes – Monograph Series.
Eaton, M. L. (1985). The Gauss-Markov theorem in multivariate analysis. In *Multivariate Analysis – VI*, edited by P. R. Krishnaiah. Elsevier Science Publishers B. V.
Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression, with discussion. *The Annals of Statistics*, **32**, 407499.
Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.
Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, Second Edition. MIT Press, Cambridge, MA.
Fisher, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, **85**, 597-612.
Fisher, R. A. (1935). *The Design of Experiments*, Ninth Edition, 1971. Hafner Press, New York.
Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, **16**, 499-511.
Geisser, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference*, edited by V.P. Godambe and D.A. Sprott. Holt, Reinhart, and Winston, Toronto.
Geisser, S. (1993). *Predictive Inference. An Introduction*. Chapman and Hall, New York.
Gnanadesikan, R. (1977). *Methods for Statistical Analysis of Multivariate Observations*. John Wiley and Sons, New York.

Goldstein, M. and Smith, A. F. M. (1974). Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society, Series B*, **26**, 284-291.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

Groß, J. (2004). The general Gauss–Markov model with possibly singular dispersion matrix. *Statistical Papers*, **25(3)**, 311-336.

Guttman, I. (1970). *Statistical Tolerance Regions*. Hafner Press, New York.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

Hartigan, J. (1969). Linear Bayesian methods. *Journal of the Royal Statistical Society, Series B*, **31**, 446-454.

Harville, D. A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, **80**, 132-138.

Haslett, J. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *Journal of the Royal Statistical Society, Series B*, **61**, 603-609.

Haslett, J. and Hayes, K. (1998). Residuals for the linear model with general covariance structure. *Journal of the Royal Statistical Society, Series B*, **60**, 201-215.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226-252.

Hinkelmann, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments*. John Wiley and Sons, New York.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, **56**, 495-504.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York.

Hoerl, A. E. and Kennard, R. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55-67.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman and Hall, London.

Huynh, H. and Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association*, **65**, 1582-1589.

Jeffreys, H. (1961). *Theory of Probability*, Third Edition. Oxford University Press, London.

Johnson, R. A. and Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ.

Khuri, A. I., Mathew, T., and Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*. John Wiley and Sons, New York.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*, Fifth Edition. McGraw-Hill Irwin, New York.

Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley and Sons, New York.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, Second Edition. John Wiley and Sons, New York.

Lin, T-H. and Harville, D. A. (1991). Some alternatives to Wald's confidence interval and test. *Journal of the American Statistical Association*, **86**, 179-187.

Lindley, D. V. (1971). *Bayesian Statistics: A Review*. SIAM, Philadelphia.

McCullagh, P. (2000). Invariance and factorial models, with discussion. *Journal of the Royal Statistical Society, Series B*, **62**, 209-238.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition. Chapman and Hall, London.

Mandansky, A. (1988). *Prescriptions for Working Statisticians*. Springer-Verlag, New York.

Mandel, J. (1961). Nonadditivity in two-way analysis of variance. *Journal of the American Statistical Association*, **56**, 878-888.

Mandel, J. (1971). A new analysis of variance model for nonadditive data. *Technometrics*, **13**, 1-18.
Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, **12**, 591-612.
Martin, R. J. (1992). Leverage, influence and residuals in regression models when observations are correlated. *Communications in Statistics – Theory and Methods*, **21**, 1183-1212.
Mathew, T. and Sinha, B. K. (1992). Exact and optimum tests in unbalanced split-plot designs under mixed and random models. *Journal of the American Statistical Association*, **87**, 192-200
Miller, F. R., Neill, J. W., and Sherfey, B. W. (1998). Maximin clusters for near replicate regression lack of fit tests. *The Annals of Statistics*, **26**, 1411-1433.
Miller, F. R., Neill, J. W., and Sherfey, B. W. (1999). Implementation of maximin power clustering criterion to select near replicates for regression lack-of-fit tests. *Journal of the American Statistical Association*, **94**, 610-620.
Miller, R. G., Jr. (1981). *Simultaneous Statistical Inference*, Second Edition. Springer-Verlag, New York.
Milliken, G. A. and Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association*, **65**, 797-807.
Moguerza, J. M. and Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, **21**, 322-336.
Monlezun, C. J. and Blouin, D. C. (1988). A general nested split-plot analysis of covariance. *Journal of the American Statistical Association*, **83**, 818-823.
Morrison, D. F. (1976). *Multivariate Statistical Methods*. McGraw-Hill, New York.
Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.
Neill, J. W. and Johnson, D. E. (1984). Testing for lack of fit in regression – a review. *Communications in Statistics, Part A – Theory and Methods*, **13**, 485-511.
Öfversten, J. (1993). Exact tests for variance components in unbalanced linear models. *Biometrics*, **49**, 45-57.
Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681-686.
Peixoto, J. L. (1993). Four equivalent definitions of reparameterizations and restrictions in linear models. *Communications in Statistics, A*, **22**, 283-299.
Picard, R. R. and Berk, K. N. (1990). Data splitting. *The American Statistician*, **44**, 140-147.
Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, **79**, 575-583.
Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley and Sons, New York.
Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston.
Rao, C. R. (1971). Estimation of variance and covariance components—MINQUE theory. *Journal of Multivariate Analysis*, **1**, 257-275.
Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Second Edition. John Wiley and Sons, New York.
Ravishanker, N. and Dey, D. (2002). *A First Course in Linear Model Theory*. Chapman and Hall/CRC Press, Boca Raton, FL.
Rencher, A. C. (2008). *Linear Models in Statistics*, Second Edition. John Wiley and Sons, New York.
Ripley, B. D. (1981). *Spatial Statistics*. John Wiley and Sons, New York.
St. Laurent, R. T. (1990). The equivalence of the Milliken-Graybill procedure and the score test. *The American Statistician*, **44**, 36-37.
Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York.
Schafer, D. W. (1987). Measurement error diagnostics and the sex discrimination problem. *Journal of Business and Economic Statistics*, **5**, 529-537.

Schatzoff, M., Tsao, R., and Fienberg, S. (1968). Efficient calculations of all possible regressions. *Technometrics*, **10**, 768-779.
Scheffé, H. (1959). *The Analysis of Variance*. John Wiley and Sons, New York.
Searle, S. R. (1971). *Linear Models*. John Wiley and Sons, New York.
Searle, S. R., Casella, G., and McCulloch, C. (1992). *Variance Components*. John Wiley and Sons, New York.
Searle, S. R. and Pukelsheim, F. (1987). Estimation of the mean vector in linear models, Technical Report BU-912-M, Biometrics Unit, Cornell University, Ithaca, NY.
Seber, G. A. F. (1966). *The Linear Hypothesis: A General Theory*. Griffin, London.
Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley and Sons, New York.
Seely, J. F. and El-Bassiouni, Y. (1983). Applying Wald's variance component test. *The Annals of Statistics*, **11**, 197-201.
Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, **67**, 215-216.
Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591-611.
Shewhart, W. A. (1931). *Economic Control of Quality*. Van Nostrand, New York.
Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. Graduate School of the Department of Agriculture, Washington. Reprint (1986), Dover, New York.
Shi, L. and Chen, G. (2009). Influence measures for general linear models with correlated errors. *The American Statistician*, **63**, 40-42.
Shillington, E. R. (1979). Testing lack of fit in regression without replication. *Canadian Journal of Statistics*, **7**, 137-146.
Shumway, R. H. and Stoffer, D. S. (2000). *Times Series Analysis and Its Applications*. Springer-Verlag, New York.
Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. John Wiley and Sons, New York.
Smith, A. F. M. (1986). Comment on an article by B. Efron. *The American Statistician*, **40**, 10.
Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317-343.
Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*, Seventh Edition. Iowa State University Press, Ames.
Sulzberger, P. H. (1953). The effects of temperature on the strength of wood, plywood and glued joints. Department of Supply, Report ACA-46, Aeronautical Research Consultative Committee, Australia.
Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.
Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics*, **5**, 232–242.
Utts, J. (1982). The rainbow test for lack of fit in regression. *Communications in Statistics—Theory and Methods*, **11**, 2801-2815.
Weisberg, S. (1985). *Applied Linear Regression*, Second Edition. John Wiley and Sons, New York
Wermuth, N. (1976). Model search among multiplicative models. *Biometrics*, **32**, 253-264.
Wichura, M. J. (2006). *The Coordinate-free Approach to Linear Models*. Cambridge University Press, New York.
Williams, E. J. (1959). *Regression Analysis*. John Wiley and Sons, New York.
Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York.
Zhu, M. (2008). Kernels and ensembles: Perspectives on statistical learning. *The American Statistician*, **62**, 97-109.

Author Index

Index