7.1 Anderson models.

The matrix $\Omega^{-1}$ of (6.1.16) may be written in the following way:

\[
(7.1.1) \quad \Omega(\rho)^{-1} = (1 + \rho^2)I - 2\rho\Theta + \rho(1 - \rho)C,
\]

where

\[
(7.1.2) \quad \Theta = \frac{1}{2} \begin{bmatrix}
1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & \cdots & 0 & 1 & 1
\end{bmatrix}
\]

(i.e., $\Theta = [\theta_{ij}]$ with $\theta_{11} = \theta_{nn} = \theta_{i+1,i} = \theta_{i,i+1} = \frac{1}{2}$ and $\theta_{ij} = 0$ otherwise) and where

\[
(7.1.3) \quad C = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}
\]

(i.e., $C = [c_{ij}]$ with $c_{11} = c_{nn} = 1$ and $c_{ij} = 0$ otherwise).

Anderson (1948) noted that if $\Omega(\rho)^{-1}$ in (7.1.1) were replaced by

\[
(7.1.4) \quad \Phi(\rho)^{-1} = (1 + \rho^2)I - 2\rho\Theta,
\]

the resulting model would be much more tractable. [To justify the notation on the left it must be shown that matrix on the right in (7.1.4) is positive definite for $-1 < \rho < 1$; this will be shown in Lemma 7.1.1 below. The formula (7.1.4) will then define the matrix function $\Phi(\rho)$ which is used to approximate $\Omega(\rho)$]. Since the two inverse matrices differ only in the corners, by the quantity $\rho(1 - \rho)$, one might conjecture that the properties of the model (7.1.4) constitute a fairly good approximation to those of the exact model (7.1.1). To what extent this conjecture is true has been studied in the special cases of estimating a mean or a linear trend (see §6.2.1 above), but in the general case it remains a largely unexplored problem; in this section, we shall not attempt to make an assessment, but simply analyze the regression model (6.1.1) on the assumption that the residuals $\varepsilon_i$ are governed by a stochastic process for which $\sigma^{-2} \mathbf{E} \{ \varepsilon \varepsilon' \} = \Phi(\rho)$ has an inverse of the form (7.1.4).

More generally, consider a regression model

\[
(7.1.5) \quad y = X\beta + \varepsilon; \quad \mathbf{E}\{\varepsilon\} = 0, \quad \mathbf{E}\{\varepsilon \varepsilon'\} = V(\rho)
\]
such that \(V(\rho)\) satisfies
\[
V(\rho)^{-1} = \kappa(\rho)\{I + h(\rho)\Theta\},
\]
where (i) \(\kappa(\rho)\) is a real-valued function of \(\rho\), (ii) \(\Theta\) is a fixed symmetric matrix, and (iii) \(h(\rho)\) is a real-valued strictly monotone function of \(\rho\) in \(-1 < \rho < 1\), such that \(I + h(\rho)\Theta\) is positive definite. [This latter condition is required to justify the notation \(V(\rho)^{-1}\).] Let \(P\) be an \(n \times n\) orthogonal matrix \((P^TP = I)\) such that
\[
P^TV(\rho)P = \Lambda(\rho),
\]
where \(\Lambda(\rho)\) is a diagonal matrix whose entries are the eigenvalues of \(V(\rho)\). Premultiplying both sides of (7.1.7) by \(P\) we have
\[
V(\rho)P = P\Lambda(\rho),
\]
or, denoting \(P = [p^0, p^1, \ldots, p^{n-1}]\) and \(\Lambda(\rho) = \text{diag}\{\lambda_0(\rho), \lambda_1(\rho), \ldots, \lambda_{n-1}(\rho)\}\),
\[
V(\rho)p^j = p^j\lambda_j(\rho),
\]
i.e., the columns \(p^j\) of \(P\) are eigenvectors of \(V(\rho)\) corresponding to the eigenvalues \(\lambda_j(\rho)\). (The numbering from 0 to \(n - 1\) is for later convenience.)

Inverting both sides of (7.1.7), we have
\[
P^TV(\rho)^{-1}P = \Lambda(\rho)^{-1},
\]
i.e., \(V(\rho)^{-1}\) has the same eigenvectors as \(V(\rho)\), and the reciprocal eigenvalues. Thus from (7.1.6) we have
\[
\Lambda(\rho)^{-1} = P^TV(\rho)^{-1}P = \kappa(\rho)\{I + h(\rho)P^T\Theta P\}. 
\]
Since the left side of (7.1.11) is a diagonal matrix, the matrix \(P^T\Theta P\) on the right must be also, hence
\[
P^T\Theta P = \Upsilon,
\]
say, where \(\Upsilon = \text{diag}\{\upsilon_j\}\) and the \(\upsilon_j\) are the eigenvalues of \(\Theta\). [The reader should note the distinction between the lower-case Greek upsilon (\(\upsilon\)) used here and the italic ‘vee’ (\(\upsilon\)) as well as the Greek nu (\(\nu\)).] From (7.1.12), the eigenvectors of \(\Theta\) are the same as those of \(V(\rho)^{-1}\), which in turn are the same as those of \(V(\rho)\). Since \(\Theta\) is fixed and given, it follows that the assumption (7.1.6) implies that the eigenvectors of \(V(\rho)\) are given and independent of \(\rho\).

The case (7.1.4) corresponds to that of (7.1.6) in which \(\Theta\) is given by (7.1.2) and
\[
V(\rho)^{-1} = \sigma^{-2}\Phi(\rho)^{-1} = \frac{1 + \rho^2}{\sigma^2}\left\{I - \frac{2\rho}{1 + \rho^2}\Theta\right\},
\]
i.e.,
\[
\kappa(\rho) = \frac{1 + \rho^2}{\sigma^2}, \quad h(\rho) = -\frac{2\rho}{1 + \rho^2}.
\]
We verify that \(h'(\rho) = -2(1 + \rho^2)^{-2}(1 - \rho^2) < 0\) for \(-1 < \rho < 1\), so \(h(\rho)\) is monotone strictly decreasing. It remains to verify that \((1 + \rho^2)I - 2\rho\Theta\) is positive definite for \(-1 < \rho < 1\). Premultiplying both sides of (7.1.4) by \(P^\theta\), and postmultiplying by \(P\), we obtain with (7.1.12) the eigenvalues of \(\Phi(\rho)^{-1}\):
\[
\Psi(\rho)^{-1} = P^T\Phi(\rho)^{-1}P = (1 + \rho^2)I - 2\rho\Upsilon,
\]
i.e.,
\[
\psi_j(\rho)^{-1} = 1 + \rho^2 - 2\rho\upsilon_j \quad (j = 0, 1, \ldots, n - 1).
\]
Lemma 7.1.1. The eigenvalues \( v_j \) of the matrix \( \Theta \) of (7.1.2) all lie in the interval 
\(-1 \leq v_j \leq 1\); hence, the eigenvalues \( \psi(\rho)^{-1} \) of the matrix \( \Phi(\rho)^{-1} \) of (7.1.4) are 
positive for all \( \rho \) in the interval \(-1 < \rho < 1\).

Proof. An eigenvalue \( v \) of \( \Theta \) must satisfy

\[
|Iv - \Theta| = \begin{vmatrix}
    v - \frac{1}{2} & -\frac{1}{2} & 0 & \cdots & 0 & 0 & 0 \\
    -\frac{1}{2} & v - \frac{1}{2} & 0 & \cdots & 0 & 0 & 0 \\
    0 & -\frac{1}{2} & v & \cdots & 0 & 0 & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
    0 & 0 & 0 & \cdots & v - \frac{1}{2} & 0 & 0 \\
    0 & 0 & 0 & \cdots & -\frac{1}{2} & v & -\frac{1}{2} \\
    0 & 0 & 0 & \cdots & 0 & -\frac{1}{2} & v - \frac{1}{2}
\end{vmatrix} = 0.
\]

If \( |v| > 1 \) then the absolute value of each diagonal element of \( Iv - \Theta \) is greater than 
the sum of the absolute values of the remaining elements in the same column, i.e., 
\( Iv - \Theta \) has a dominant diagonal; it follows that \( Iv - \Theta \) is nonsingular for \( |v| > 1 \) 
(see McKenzie 1960), contradicting (7.1.17). Therefore the eigenvalues of \( \Theta \) must satisfy

\[
-1 \leq v_j \leq 1.
\]

Letting \( P \) be an orthogonal matrix diagonalizing \( \Theta \) to

\[
P^\prime \Theta P = \Upsilon,
\]

and applying this diagonalization to the right side of (7.1.4), we have for the eigenvalues of the (as yet undetermined) matrix on the left,

\[
f(\rho; v) = \rho^2 - 2\rho v + 1.
\]

Viewed as a quadratic polynomial in real \( \rho \), obviously \( f(\rho) > 0 \) for sufficiently large 
\( |\rho| \). We verify that the zeros of \( f \) are \( \rho = v \pm \sqrt{v^2 - 1} \), which are complex if \( |v| < 1 \); 
they are real and repeated for \( v = \pm 1 \); therefore \( f(\rho, v) > 0 \) for \( |v| \neq 1 \). But clearly 
\( f(\rho; v) = (\rho \mp 1)^2 > 0 \) for \( v = \pm 1 \) and \(-1 < \rho < 1 \). It follows that the matrix on 
the right side of (7.1.4) is positive definite for \(-1 < \rho < 1 \), hence the left side is 
well defined. \( \square \)

The following result, due to von Neumann (1941, p. 371), provides much more 
detailed information about \( \Theta \).

Lemma 7.1.2 (von Neumann). The eigenvalues of the matrix \( \Theta \) of (7.1.2) are 
given by

\[
v_j = \cos \frac{j\pi}{n} \quad \text{for} \quad j = 0, 1, \ldots, n - 1,
\]

and the corresponding eigenvectors by

\[
\xi^j = \begin{bmatrix}
    \cos \frac{j\pi}{2n} \\
    \cos \frac{3j\pi}{2n} \\
    \vdots \\
    \cos \frac{(2n - 1)j\pi}{2n}
\end{bmatrix} \quad \text{for} \quad j = 0, 1, \ldots, n - 1.
\]
Proof. If $\xi$ is an eigenvector of $\Theta$ and $v$ the corresponding eigenvalue, then from $\Theta \xi = v \xi$, $\xi \neq 0$ and (7.1.2), we have

\[
\begin{align*}
\xi_1 + \xi_2 &= 2v\xi_1 \\
\xi_1 + \xi_3 &= 2v\xi_2 \\
\xi_2 + \xi_4 &= 2v\xi_3 \\
&\vdots \\
\xi_{n-3} + \xi_{n-1} &= 2v\xi_{n-2} \\
\xi_{n-2} + \xi_n &= 2v\xi_{n-1} \\
\xi_{n-1} + \xi_n &= 2v\xi_n
\end{align*}
\]

(7.1.22)

where not all $\xi_t = 0$. Denoting

\[
\xi_0 = \xi_1 \quad \text{and} \quad \xi_{n+1} = \xi_n.
\]

(7.1.23)

(7.1.22) may be written in the form

\[
\xi_{t-1} + \xi_{t+1} = 2v\xi_t \quad (t = 1, 2, \ldots, n).
\]

(7.1.24)

From (7.1.18) it follows that any eigenvalue $v$ of $\Theta$ can be written

\[
v = \cos \alpha
\]

for some $\alpha$; thus (7.1.24) may be written

\[
\xi_{t-1} + \xi_{t+1} = 2 \cos \alpha \cdot \xi_t \quad (t = 1, 2, \ldots, n).
\]

(7.1.26)

From the trigonometric identity

\[
\cos \phi + \cos \psi = 2 \cos \frac{1}{2} (\phi + \psi) \cos \frac{1}{2} (\phi - \psi)
\]

(7.1.27)

we now verify that (7.1.26) is satisfied by

\[
\xi_t = \cos (t - \frac{1}{2}) \alpha;
\]

(7.1.28)

for, substituting (7.1.28) in both sides of (7.1.26) we obtain

\[
\begin{align*}
\xi_{t-1} + \xi_{t+1} &= \cos (t - 1 - \frac{1}{2}) \alpha + \cos (t + 1 - \frac{1}{2}) \alpha \\
&= 2 \cos \frac{1}{2} (2t - 1) \alpha \cdot \cos \frac{1}{2} (-2\alpha) \\
&= 2 \cos \alpha \cos (t - \frac{1}{2}) \alpha \\
&= 2 \cos \alpha \cdot \xi_t.
\end{align*}
\]

(7.1.29)

It remains to determine $\alpha$. The first equation of (7.1.23) implies no restriction, since $\xi_0 = \cos (-\frac{1}{2} \alpha) = \cos (\frac{1}{2} \alpha) = \xi_1$ automatically. However, the second equation of (7.1.23) requires

\[
\cos (n - \frac{1}{2}) \alpha = \xi_n = \xi_{n+1} = \cos (n + \frac{1}{2}) \alpha,
\]

(7.1.30)
implying that the angle \((n + \frac{1}{2})\alpha\) differs from both \((n - \frac{1}{2})\alpha\) and \(-(n - \frac{1}{2})\alpha\) by an integral multiple of \(2\pi\) radians. The second of these conditions states that

\[
(7.1.31) \quad (n + \frac{1}{2})\alpha = 2\pi j - (n - \frac{1}{2})\alpha
\]

where \(j\) is an integer; this implies that

\[
(7.1.32) \quad \alpha = \frac{j\pi}{n}.
\]

Now for all \(j = 0, 1, \ldots, n - 1\) we have \(0 \leq \alpha < \pi\), whence substitution of (7.1.32) in (7.1.28) gives \(\xi_t = \cos(j\pi/2n) > 0\); therefore for no \(j = 0, 1, \ldots, n - 1\) are all \(\xi_t = 0, t = 1, 2, \ldots, n\). Thus if we define

\[
(7.1.33) \quad v_j = \cos\frac{j\pi}{n} \quad \text{and} \quad \xi_{tj} = \cos\frac{(2t - 1)j\pi}{2n} \quad (t = 1, 2, \ldots, n; \ j = 0, 1, \ldots, n - 1)
\]

then these satisfy the required conditions (7.1.22), hence the \(v_j\)'s of (7.1.33) are the eigenvalues of \(\Theta\) and (7.1.33) also defines the corresponding eigenvectors \(\xi_{tj}\). □

From (7.1.32) it follows that

\[
(7.1.34) \quad 1 = v_0 > v_1 > \ldots > v_{n-1} > -1;
\]

thus, the eigenvalues of \(\Theta\) are all distinct. It follows that its eigenvectors (7.1.33) are mutually orthogonal (cf. Perlis 1952, p. 184). This follows directly from the fact that

\[v_i \xi^i \xi^j = \xi^i \Theta \xi^j = \xi^i \xi^j v_j\]

whence \((v_i - v_j)\xi^i \xi^j = 0\), so \(v_i \neq v_j\) implies \(\xi^i \xi^j = 0\).

In order to obtain an orthonormal set of eigenvectors of \(\Theta\) it is necessary to ascertain the lengths of the vectors \(\xi^j\). This will be done in the next lemma, which incidentally also provides a direct proof of the mutual orthogonality of the eigenvectors \(\xi^j\).

**Lemma 7.1.3.** The eigenvectors (7.1.33) have length \(|\xi^j| = \sqrt{n}\) for \(j = 0\) and \(\sqrt{n/2}\) for \(j = 1, 2, \ldots, n - 1\). Accordingly, an orthonormal set of eigenvectors \(p^j\) of \(\Theta\) \((j = 0, 1, \ldots, n - 1)\) is given by

\[
(7.1.35) \quad p^j_t = \frac{1}{\sqrt{n}} \cos\frac{(2t - 1)j\pi}{2n} \quad (j = 1, 2, \ldots, n - 1)
\]

for \(t = 1, 2, \ldots, n\).

**Proof.** First we verify that

\[
(7.1.36) \quad \sum_{t=1}^{n} \cos\frac{(2t - 1)i\pi}{2n} \cos\frac{(2t - 1)j\pi}{2n} = \begin{cases} 0 & \text{if } i \neq j \\ n & \text{if } i = j = 0 \\ n/2 & \text{if } i = j \neq 0 \end{cases}
\]
for $0 \leq i, j \leq n - 1$. This is the discrete counterpart of the orthogonality relations for continuous $t$, $0 \leq t \leq n$:

$$
\int_{\frac{1}{2}}^{\frac{n+1}{2}} \cos \frac{(2t-1)i\pi}{2n} \cos \frac{(2t-1)j\pi}{2n} \, dt = \begin{cases} 
\frac{n}{2} & \text{if } i = j = 0 \\
0 & \text{if } i \neq j, \text{ or } i = j \neq 0 \\
n/2 & \text{if } i = j \neq 0;
\end{cases}
$$

and is established in essentially the same way, by means of the trigonometric identity

$$
\cos \phi \cos \psi = \frac{1}{2}[\cos(\phi + \psi) + \cos(\phi - \psi)]
$$

(which is the inverse of (7.1.27)):

$$(7.1.38) \quad \cos \frac{(2t-1)i\pi}{2n} \cos \frac{(2t-1)j\pi}{2n} = \frac{1}{2} \left\{ \cos \frac{(2t-1)(i+j)\pi}{2n} + \cos \frac{(2t-1)(i-j)\pi}{2n} \right\}.$$ 

When $i = j = 0$, the right side of (7.1.38) is equal to 1, and the middle sum of (7.1.36) and integral of (7.1.37) are obtained immediately. The remaining conditions follow once it is established that

$$
\sum_{t=1}^{n} \cos \frac{(2t-1)m\pi}{2n} = \int_{\frac{1}{2}}^{\frac{n+1}{2}} \cos \frac{(2t-1)m\pi}{2n} \, dt = 0 \quad (0 < |m| < 2n).
$$

For the continuous case of (7.1.39), set $u = \frac{(2t-1)m\pi}{2n}$, whence $t = \frac{nu}{m\pi} + \frac{1}{2}$.

Then

$$
\int \cos \frac{(2t-1)m\pi}{2n} \, dt = \int \cos u \frac{n}{m\pi} \, du = \frac{n}{m\pi} \sin u = \frac{n}{m\pi} \sin \frac{(2t-1)m\pi}{2n}
$$

whence

$$
\int_{\frac{1}{2}}^{\frac{n+1}{2}} \cos \frac{(2t-1)m\pi}{2n} \, dt = \frac{n}{m\pi} [\sin m\pi - \sin 0] = 0
$$

since $\sin m\pi = \sin 0 = 0$ for all integers $m$.

To obtain the discrete formula of (7.1.39) we may proceed as follows. Let $F(t)$ be a real-valued function defined for integral values of $t$, and let the first-difference operator be defined by

$$
\Delta F(t) = F(t+1) - F(t) \equiv f(t).
$$

Analogously to the indefinite integral, the indefinite sum $\Sigma$ is defined by

$$
\Sigma f(t) = \Sigma \Delta F(t) = F(t).
$$

(cf. Jordan 1960, pp. 100–117). The corresponding definite sum is

$$
\sum_{t=1}^{n} f(t) = \sum_{t=1}^{n} F(t+1) - \sum_{t=1}^{n} F(t) = F(n+1) - F(1)
$$
which may be written\(^1\)

\[(7.1.45)\]
\[
\sum_{t=1}^{n} f(t) = \left[\Sigma f(t)\right]_{1}^{n+1} = F(t)_{1}^{n+1}.
\]

Applying this to \(f(t) = \cos(at + b)\), let us first compute

\[(7.1.46)\]
\[
\Delta \sin(at + b') = \sin(at + b' + a) - \sin(at + b') = 2 \sin \frac{1}{2} a \cdot \cos(at + b' + \frac{1}{2} a),
\]

where use has been made of the trigonometric identity

\[(7.1.47)\]
\[
\sin \phi - \sin \psi = 2 \sin \frac{1}{2} (\phi - \psi) \cos \frac{1}{2} (\phi + \psi).
\]

Put \(b = b' + \frac{1}{2} a\). Then applying the operator \(\Sigma\) to (7.1.46) we obtain

\[(7.1.48)\]
\[
\sin(at + b - \frac{1}{2} a) = 2 \sin \frac{1}{2} a \Sigma \cos(at + b),
\]

whence, as long as \(a\) is not an even multiple of \(\pi\),

\[(7.1.49)\]
\[
\Sigma \cos(at + b) = \frac{\sin[a(t - \frac{1}{2}) + b]}{2 \sin \frac{1}{2} a}.
\]

When

\[(7.1.50)\]
\[
\frac{at + b}{2n} = \frac{(2t - 1)m\pi}{2n} = \frac{m\pi}{n} - \frac{m\pi}{2n} \quad (0 < |m| < 2n),
\]

this yields the indefinite sum

\[(7.1.51)\]
\[
\Sigma \cos \frac{(2t - 1)m\pi}{2n} = \frac{\sin \left(\frac{(t-1)m\pi}{n}\right)}{2 \sin \frac{m\pi}{2n}}.
\]

The required definite sum is then, in accordance with (7.1.45),

\[(7.1.52)\]
\[
\sum_{t=1}^{n} \cos \frac{(2t - 1)m\pi}{2n} = \frac{\sin m\pi}{2 \sin \frac{m\pi}{2n}} = 0.
\]

The orthogonal matrix \(P = [p_{ij}]\) of (7.1.43) which diagonalizes \(\Theta\) to \(P'\Theta P = \Upsilon = \text{diag}\{\upsilon_0, \upsilon_1, \ldots, \upsilon_{n-1}\}\), where \(\upsilon_j = \cos(j\pi/n)\), is now given by (7.1.35). □

As an example, for \(n = 3\):

\[
P = \begin{bmatrix}
\frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{2} \cos \frac{\pi}{6} & \frac{\sqrt{2}}{2} \cos \frac{\pi}{3} \\
\frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{2} \cos \frac{\pi}{2} & \frac{\sqrt{2}}{2} \cos \pi \\
\frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{2} \cos \frac{5\pi}{6} & \frac{\sqrt{2}}{2} \cos \frac{5\pi}{3}
\end{bmatrix} = \begin{bmatrix}
\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\
\frac{1}{\sqrt{3}} & 0 & -\frac{\sqrt{2}}{\sqrt{3}} \\
\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}}
\end{bmatrix}.
\]

\(^1\)By convention, the upper limit is not included in the definite sum; cf. Jordan (1960, p. 117).
For the case $n = 17$, the graphs of the first five eigenvectors (7.1.35) are plotted continuously in Figure 7.1.1, with time measured downwards.

We may now return to the question discussed in Section 2.6; when are least-squares estimators minimum-variance affine unbiased? In terms of Theorem 2.6.9, we have the criterion: if $E(\varepsilon\varepsilon') = \sigma^2 \Phi(\rho)$, where $\Phi(\rho)$ is given by (7.1.4), then if the $k$ columns of $X$ are linear combinations of an $n \times k$ submatrix $P_1$ of the
matrix \( P = [p_{ij}] \) (given by (7.1.19) and (7.1.35)) of eigenvectors of \( \Phi(\rho) \), then the ordinary least-squares estimator of \( \beta \) is Gauss-Markov. In the particular case in which \( P = [P_1, P_2] \) is a partition of \( P \) into its first \( k \) and last \( n - k \) columns as ordered in (7.1.35), the condition \( X = P_1 K_1 \) states that the independent variables can be expressed as linear combinations of the finite Fourier series (7.1.35) in the lowest \( k \) frequencies.²

To get an idea of the extent to which these conditions are met in practical cases, two examples are illustrated in Figures 7.1.2 and 7.1.3. In Figure 7.1.2, data for the 17-year period 1923–39 are taken from a paper by Theil and Nagar (1961) who studied the regression of the logarithm of consumption of clothing in Holland on the logarithm of real per capita income \( (x_{t1}) \) and the logarithm of a deflated price index of clothing \( (x_{t2}) \). Letting \( X \) be the \( 17 \times 3 \) observation matrix, and \( P_1 = [p^0, p^1, p^2] \), we may calculate the empirical regression of \( X \) on \( P_1 \) from the formula \( \hat{\Gamma} = (P_1' P_1)^{-1} P_1' X \) and then plot the fitted series

\[
(7.1.53) \quad \hat{x}_{ti} = \sum_{j=0}^{2} p_{ij}\hat{\gamma}_{ij} = \sum_{j=0}^{2} \cos \left( \frac{(2t - 1)j\pi}{2n} \right) \hat{\gamma}_{ij} \quad (i = 1, 2; \; t = 1, 2, \ldots, 17)
\]

where \( \hat{x}_{t1}, \hat{x}_{t2} \) are respectively the values for the logarithm of income and price predicted by the eigenvectors \( [p^0, p^1, p^2] \), shown by the continuous solid curves marked 012 in Figure 7.1.2; the interpolated data for \( x_{t1}, x_{t2} \) are shown by broken dotted lines. Also shown are curves marked 013 corresponding to the choice of regression vectors \( [p^0, p^1, p^3] \).³ It is clear that both give a reasonably good fit for the price data, whereas it would be necessary to use the four vectors \( [p^0, p^1, p^2, p^3] \) to obtain a more satisfactory fit for the income data, which reflect the boom of the 1920s and the ensuing depression.

²On Fourier series see Jordan (1960, pp. 463–4).
³The regression coefficients \( \hat{\gamma}_{ij} \) of (7.1.53) are easily computed by taking sums of cross-products. The computations give

\[\hat{\gamma}_{10} = 2.01225, \hat{\gamma}_{11} = 0.005332, \hat{\gamma}_{12} = -0.0169827, \text{ and } \hat{\gamma}_{13} = -0.022351\]

for income, and

\[\hat{\gamma}_{20} = 2.01225, \hat{\gamma}_{21} = 0.005332, \hat{\gamma}_{22} = -0.0169827, \text{ and } \hat{\gamma}_{23} = -0.022351\]

for price.
Figure 7.1.3 provides a similar illustration for data on income and the relative price of spirits (both in logarithms) in the U.K. during the 69-year period 1870–1938; these data are taken from the paper of Durbin and Watson (1951, p. 160), who in turn obtained them from Prest (1949). The fit is good for the income data, but that for the price data is marred by the erratic swings during and after the first

\footnote{The data themselves are not to be found in Prest (1949), however. For these data the}
World War.

Regression coefficients $\gamma_{ij}$ of (7.1.53) are found to be

$$\gamma_{10} = 1.963494, \gamma_{11} = -0.1310585, \text{ and } \gamma_{12} = -0.0308038$$

for income, and

$$\gamma_{20} = 2.118375, \gamma_{21} = -0.2540703, \text{ and } \gamma_{22} = 0.1095974$$

for price.
Figure 7.1.3

7.2 Testing for autocorrelation: Anderson’s theorem and the Durbin-Watson test.

The most prominent method of testing for autocorrelation of residuals in re-
gression is that of Durbin and Watson (1950, 1951). The theory behind this test rests upon some fundamental results of Anderson (1948). Thus, we first examine Anderson’s theory and proposed test.

Anderson (1948, p. 94) proved the following fundamental theorem which underlies the method of Durbin and Watson.

**Theorem 7.2.1 (Anderson).** Let the density function of \( \varepsilon \) be

\[
K \exp\left\{-\frac{1}{2} \alpha [\varepsilon'(\Psi + \lambda \Theta)e]\right\},
\]

where \( K \) is a positive constant, \( \alpha > 0 \), \( \Psi \) is a symmetric positive-definite matrix, \( \Theta \) is a symmetric matrix, and \( \lambda \) a parameter such that \( \Psi + \lambda \Theta \) is positive definite. Secondly, let \( X \) be an \( n \times k \) matrix of rank \( k \) whose \( k \) columns are linear combinations of some subset of \( k \) of the eigenvectors of \( \Theta \). Let the vector of residuals from the least-squares regression of \( y \) on \( X \) be denoted by

\[
e = y - Xb = y - XX'y = (I - XX')y = Ey,
\]

where \( E = I - XX' = I - X(X'X)^{-1}X' \). Then a uniformly most powerful test of the hypothesis \( \lambda = 0 \), against the alternative \( \lambda < 0 \), at level \( \gamma \), is \( r > r^* \) where

\[
r = \frac{\varepsilon'^\prime \Theta e}{\varepsilon'^\prime \Psi e},
\]

and \( r^* \) is chosen so that the probability that \( r > r^* \) is \( \gamma \) when \( \lambda = 0 \).

In the special case under consideration, we take \( \Psi = I, \alpha = (1 + \rho^2)/\sigma^2 \), \( \lambda = -2\rho/(1 + \rho^2) \), and \( \Theta \) as in (7.1.2) above. The joint density function (7.2.1) therefore becomes

\[
K \exp\left\{-\frac{1}{2\sigma^2 \varepsilon^\prime[(1 + \rho^2)I - 2\rho \Theta]\varepsilon}\right\}
\]

and the corresponding test statistic (7.2.3) is (cf. Anderson, 1948, p. 108)

\[
r = \frac{\varepsilon'^\prime \Theta e}{\varepsilon'^\prime e} = \frac{1}{2} \frac{\varepsilon'^\prime + \frac{1}{2} \varepsilon'^\prime + \sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2}.
\]

The matrix of the quadratic form of (7.2.4) is the matrix \( \Phi(\rho)^{-1} \) of (7.1.4); writing it out in full:

\[
\Phi(\rho)^{-1} = \\
\begin{bmatrix}
1 + \rho^2 - \rho & -\rho & 0 & \cdots & 0 & 0 & 0 \\
-\rho & 1 + \rho^2 & -\rho & \cdots & 0 & 0 & 0 \\
0 & -\rho & 1 + \rho^2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 + \rho^2 & -\rho & 0 \\
0 & 0 & 0 & \cdots & -\rho & 1 + \rho^2 & -\rho \\
0 & 0 & 0 & \cdots & 0 & -\rho & 1 + \rho^2 - \rho
\end{bmatrix}
\]

As is seen by comparing (7.1.4) with (7.1.1), the two matrix functions are identical except for the quantity \( \rho^2 - \rho \) in the corners, and they coincide for \( \rho = 0 \) and \( \rho = 1 \). The approximation of \( \Omega(\rho)^{-1} \) by \( \Phi(\rho)^{-1} \) is crucial to the whole discussion of the Durbin-Watson statistic.

Anderson’s theorem may now be restated as follows.
Theorem 7.2.1'. Let the disturbances \( \varepsilon \) in the regression model be jointly normally distributed with mean 0 and variance matrix \( \sigma^2 \Phi(\rho) \) where \( \Phi(\rho)^{-1} \) is given by (7.1.4); and let it be further assumed that \( X = P_1K \), where \( P_1 \) is an \( n \times k \) matrix of \( k \) eigenvectors of \( \Theta \), and \( K \) is a \( k \times k \) nonsingular matrix. Then if \( r^* \) is chosen so that the probability that \( r > r^* \) is \( \gamma \) when \( \rho = 0 \), a uniformly most powerful test of the hypothesis \( \rho = 0 \), against the alternative \( 0 < \rho < 1 \), at level \( \gamma \), is provided by \( r > r^* \), where \( r \) is given by (7.2.5).

Under these conditions just specified it follows from Theorem 2.6.9 that the ordinary least-squares estimator of \( \beta \) is the Gauss-Markov estimator. Thus we come to the following conclusion: if (1) the variance of the residuals is correctly given by \( \sigma^2 \Phi(\rho) \), where \( \Phi(\rho)^{-1} \) is given by (7.1.4), and if (2) the regression vectors \( X \) are linear combinations of a subset of \( k \) eigenvectors of \( \Theta \), then Anderson’s statistic (7.2.5) provides a uniformly most powerful test of the hypothesis \( \rho = 0 \) against \( 0 < \rho < 1 \); but at the same time the ordinary least-squares estimator is the Gauss-Markov estimator. Hence, if the purpose of testing for autocorrelated residuals was to guard against loss of efficiency, then under those conditions in which Anderson’s test is uniformly most powerful, there is no loss of efficiency in any case; and the only circumstances in which there is a loss of efficiency are precisely those in which Anderson’s test is not uniformly most powerful.

Durbin and Watson (1950, p. 424) called their statistic “a slight modification” of Anderson’s. The modification consists of replacing \( \Theta \) by the matrix \( A = 2(I - \Theta) \). Consequently their statistic is related to Anderson’s by the formula

\[
(7.2.7) \quad d = \frac{e'Ae}{e'e} = 2\frac{e'e - e'\Theta e}{e'e} = 2(1 - r).
\]

Putting \( d^* = 2(1 - r^*) \), Anderson’s test

\[ r > r^* \]

becomes (cf. Durbin & Watson 1951, p. 161)

\[ d < d^* . \]

The two statistics being in one-to-one correspondence, everything that was said about the Anderson statistic applies equally well, of course, to the Durbin-Watson statistic. And also to the von Neumann (1941) ratio, which constitutes the special case in which \( k = 1 \) and \( X \) is a column of ones. In the latter case \( X \) is an eigenvector of \( \Theta \), as well (of course) as of \( A \), corresponding to the zero eigenvalue of the latter matrix (which has rank \( n - 1 \)).

The Durbin-Watson statistic is usually written in the form

\[
(7.2.8) \quad d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.
\]

This may be related to the formula (7.2.7) as follows. Denoting \( u_t = e_t - e_{t-1} \), the \((n - 1) \times 1 \) column vector \( u = (u_2, u_3, \ldots, u_n)' \) may be written \( u = W'e \) where \( W \)
is the \( n \times (n - 1) \) matrix

\[
W = \begin{bmatrix}
-1 & 0 & 0 & \cdots & 0 & 0 \\
1 & -1 & 0 & \cdots & 0 & 0 \\
0 & 1 & -1 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -1 & 0 \\
0 & 0 & 0 & \cdots & 1 & -1 \\
0 & 0 & 0 & \cdots & 0 & 1 \\
\end{bmatrix}.
\]

(7.2.9)

Defining now the matrix

\[
A = WW' = \begin{bmatrix}
1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
-1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\
0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\
0 & 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\
0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \\
\end{bmatrix},
\]

(7.2.10)

we may write (7.2.8) in the form

\[
d = \frac{\varepsilon' \varepsilon}{\varepsilon' \varepsilon} = \frac{e'WW'e}{e'\varepsilon} = \frac{e'Ae}{y'ey} = \frac{y'EAe}{y'ey} = \frac{\varepsilon' EAE \varepsilon}{\varepsilon' \varepsilon},
\]

where use has been made of (7.2.2) and the fact that \( y = X \beta + \varepsilon \).

Since the matrix \( E \) of (7.2.2) appearing in (7.2.11) is idempotent and symmetric, it is clear that the symmetric matrices \( EAE \) and \( E \) appearing in the numerator and denominator of (7.2.11) commute. It is well known (cf. Bellman 1960, p. 56) that two symmetric matrices of the same order can be simultaneously diagonalized by the same orthogonal transformation if and only if they commute. For the special case at hand the result follows simply, as shown in Durbin & Watson (1950, pp. 412–13).

**Lemma 7.2.1.** The Durbin-Watson statistic may be expressed as

\[
d = \frac{\sum_{i=1}^{n-k} \nu_i \xi_i^2}{\sum_{i=1}^{n-k} \xi_i^2}
\]

where the \( \nu_i \) are the \( n - k \) positive eigenvalues of \( EAE \).

**Proof.** Let \( Q \) be an \( n \times n \) orthogonal matrix diagonalizing the idempotent matrix \( E \) to

\[
Q'EQ = \begin{bmatrix}
I_{n-k} & 0 \\
0 & 0 \\
\end{bmatrix}
\]

where \( I_{n-k} \) is the identity matrix of order \( n - k \) corresponding to the \( n - k \) unit eigenvalues of \( E \). Then

\[
Q' EAEQ = Q' EQ' AQ E Q
\]

\[
= \begin{bmatrix}
I_{n-k} & 0 \\
0 & 0 \\
\end{bmatrix} \begin{bmatrix}
B_{11} & B_{12} \\
B_{21} & B_{22} \\
\end{bmatrix} \begin{bmatrix}
I_{n-k} & 0 \\
0 & 0 \\
\end{bmatrix}
\]

\[
= \begin{bmatrix}
B_{11} & 0 \\
0 & 0 \\
\end{bmatrix},
\]
where $[B_{ij}]$ is the appropriate partition of $B \equiv Q' A Q$. Now let $R_{11}$ be an orthogonal matrix of order $n - k$ diagonalizing $B_{11}$ to

$$R'_{11} B_{11} R_{11} = N \equiv \begin{bmatrix}
\nu_1 & 0 & \cdots & 0 \\
0 & \nu_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \nu_{n-k}
\end{bmatrix}.$$

Then the matrix

$$R = \begin{bmatrix} R_{11} & 0 \\ 0 & I_k \end{bmatrix}$$

is orthogonal, as is the matrix $H \equiv QR$. We verify that

$$H' E H = \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad H' E A E H = \begin{bmatrix} N & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus, setting $\varepsilon = H\zeta$ gives (7.2.12). \qed

A special case of interest is that of the von Neumann ratio, in which $X = \iota$ (a column of ones); since the columns (and rows) of $A$ sum to zero, $\iota^t A = A \iota = 0$, so $E A E = A$ and (7.2.12) becomes

$$d = \frac{\sum_{i=1}^{n-1} \lambda_i z_i^2}{\sum_{i=1}^{n-1} z_i^2} \quad (7.2.13)$$

where the $\lambda$s are the $n - 1$ nonvanishing eigenvalues of $A$. [This is related to the von Neumann ratio $\eta = \delta^2 / \sigma^2$ (cf. von Neumann 1941, p. 369) by $(n - 1)\eta = nd$.]

From the relation $A = 2(I - \Theta)$ we have $P^t A P = 2(I - \Upsilon) \equiv \Lambda$ from (7.1.12), hence from (7.1.33) it follows that

$$\lambda_j = 2 \left(1 - \cos \frac{j\pi}{n}\right) \quad (j = 0, 1, \ldots, n - 1), \quad (7.2.14)$$

whence from (7.1.34),

$$\lambda_0 < \lambda_1 < \cdots < \lambda_{n-1} < \lambda_n = 4. \quad (7.2.15)$$

($\lambda_n$ is defined as in (7.2.14) but it is not an eigenvalue of $A$. ) We observe from (7.1.33) that

$$v_j + v_{n-j} = \cos \frac{j\pi}{n} + \cos \left(\pi - \frac{j\pi}{n}\right) = 0$$

whence

$$\lambda_j + \lambda_{n-j} = 4. \quad (7.2.16)$$

The following lemma was introduced by Durbin & Watson (1950, p. 416; 1951, pp. 177–8). It actually goes back to Courant (1922, p. 281, Satz 3), and the essential feature of it was noticed even earlier by Fischer (1905). Courant described it (1922, p. 33) as an application of the minimum-maximum property of the eigenvalues. See also Hannan (1960, p. 119) and Bellman (1960, pp. 110–122).
Lemma 7.2.2. Let $A$ be any real $n \times n$ symmetric matrix, and $X$ any $n \times k$ matrix of rank $k$. Then if $s$ of the columns of $X$ are linear transforms of some $s$ eigenvectors of $A$, and if the eigenvalues of $A$ associated with the remaining $n - s$ eigenvectors of $A$ are denoted

\[(7.2.17)\]

\[\lambda_1' \leq \lambda_2' \leq \cdots \leq \lambda_{n-s}',\]

then

\[(7.2.18)\]

\[\lambda_i' \leq \nu_i \leq \lambda_{i+k-s}' \quad (i = 1, 2, \ldots, n - k).\]

Since the Durbin-Watson lemma holds for any real symmetric matrix $A$, in general the inequalities (7.2.17) are weak; in the case of the matrix $A$ of (7.2.10), however—the only case under consideration here—the inequalities are strict in view of (7.2.15).

Now assume that $X$ contains a column of ones (corresponding to the constant term). This column is an eigenvector of $A$ corresponding to the eigenvalue $\lambda_0 = 0$. Then even if more than one column of $X$ satisfies the conditions of the Durbin-Watson lemma, we may still take $s = 1$ in Lemma 7.2.2 and it is always the case that

\[(7.2.19)\]

\[\lambda_i \leq \nu_i \leq \lambda_{i+k-1} \quad (i = 1, 2, \ldots, n - k).\]

Therefore,

\[(7.2.20)\]

\[d_L = \frac{\sum_{i=1}^{n-k} \lambda_i' \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2} \leq \frac{\sum_{i=1}^{n-k} \nu_i \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2} \leq \frac{\sum_{i=1}^{n-k} \lambda_{i+k-1}' \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2} = d_U\]

for all $\zeta$. This may be written $d_L \leq d \leq d_U$, if the $d$ symbols are interpreted as functions of $\zeta$.

In the special case in which the $k$ columns of $X$ are eigenvectors of $A$ corresponding to its $k$ smallest eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{k-1}$, then $s = k$ and application of the Durbin-Watson lemma gives $\nu_i = \lambda_i' = \lambda_{i+k-1}'$ from (7.2.18), hence $d = d_U$. The other limiting case is that in which the regression vectors are eigenvalues of $A$ corresponding to the eigenvalues $\lambda_0, \lambda_{n-k+1}, \ldots, \lambda_{n-1}$, i.e., the smallest eigenvalue $\lambda_0 = 0$ and the $k - 1$ largest; in this case (7.2.18) gives $\nu_i = \lambda_i' = \lambda_i$, and $d = d_L$. (A direct derivation of this result is given in Lemma 7.2.3 below.)

It is on the basis of (7.2.20) that Durbin and Watson obtained significance points $d^*_L$ and $d^*_U$, corresponding to the distributions of $d_L$ and $d_U$ under the null hypothesis, to obtain a region of rejection defined by

\[d < d^*_L,\]

a region of acceptance defined by

\[d \geq d^*_U,\]

leaving the famous “zone of indeterminacy”

\[d^*_L \leq d < d^*_U,\]
where the symbol $d$ should now be understood to denote the value of the corresponding function of $e$ defined by (7.2.8) (i.e., a function of sample observations).

Taking stock of the situation, we may distinguish two cases. One is the case in which the second condition of Anderson’s theorem applies, i.e., the columns of $X$ are all linear transforms of a certain set of $k$ eigenvectors of $A$ (including a column of ones); this includes the limiting cases $d_L$ and $d_U$. Under these conditions the Anderson theory applies, and in principle the significance points $d^*$ could be tabulated in each case from the distributions of $d$, which would be of the form (7.2.12) where the $v_i$ consist of the $n - k$ eigenvalues $\lambda_i$ of $A$ other than the $k$ eigenvalues corresponding to the given $k$ eigenvectors. We would have $d_L^* \leq d^* \leq d_U^*$ and the test which rejects the null hypothesis $\rho = 0$ when $d < d^*$ and accepts it when $d \geq d^*$ would be uniformly most powerful against $\rho > 0$ when $\varepsilon$ is normal with inverse covariance matrix proportional to (7.2.6). Up to the approximation of (7.1.1) by (7.2.6), whose effect on the power of the test is not known, we may therefore reasonably consider the Durbin-Watson test to be highly satisfactory in this case. But this is also the case in which the ordinary least-squares estimator is Gauss-Markov.

The second and more usual situation is that in which the regression vectors (the columns of $X$) are not expressible as linear combinations of the same set of $k$ of the $n$ eigenvectors of $A$; this is the case in which least-squares estimation can be expected to be inefficient when serial correlation is present, and a test of serial correlation most desirable. But this is also the circumstance in which little is known about the properties of the Durbin-Watson test; these authors claim (1950, p. 425) that in this case “we still have a valid test, though possibly of reduced power.” The loss of power must surely be very substantial, and the reason for suspecting this outcome was pointed out by Durbin and Watson themselves (1950, p. 425): “any test based on least squares residuals cannot even be a likelihood ratio test,” since the least-squares estimates will not be maximum-likelihood estimates when $0 < \rho < 1$ and $E(\varepsilon \varepsilon') = \sigma^2 \Phi(\rho)$ unless the regression vectors fulfill the second condition of Anderson’s theorem. (Further, it is actually not obvious how one should define the power of the Durbin-Watson test since the true significance level of the test is inherently unknown.) A similar situation characterizes all these approaches to the problem of testing for serial correlation; commenting on some results of R. L. Anderson and T. W. Anderson, Durbin and Watson remarked (1950, p. 409): “Perversely enough, this is the very case in which the test is least needed, since the least-squares regression coefficients are best linear unbiased estimates even in the non-null case, and in addition estimates of their variance can be obtained which are at least asymptotically unbiased.” But this holds also in the present case, as shown by Wold (1950, p. 281). The perversity alluded to is no mere accident or curiosum, but is in the nature of the case, and characterizes the Durbin-Watson statistic as well.

A further comment should be made about the Durbin-Watson test. It was developed in the early 1950s, which was before the age of computers. The only reason for computing the bounds rather than approximating the significance points of the statistic itself (which of course depends on the particular $X$ matrix that occurs in an application) is that in the 1950s it was computationally infeasible to do so. This is no longer the case. We will come back to this question in section 7.3.

We shall now derive some diagonalization theorems underlying the derivation of the von Neumann ratio (1941, pp. 368–9) and the Durbin-Watson statistic (1950,
pp. 412–3); see also Hannan (1960, pp. 118–9). These theorems are concerned with certain aspects of the simultaneous diagonalization of \( E = I - XX^\dagger = I - X(X'X)^{-1}X' \) and \( A = WW' \), where these matrices have the fixed meanings attached to them in (7.2.2), (7.2.9), and (7.2.10). Since \( E \) has rank \( n - k \) and \( A \) has rank \( n - 1 \), we shall be mainly concerned with the conditions that ensure that \( EAE \) will have rank \( n - k \), as well as with those conditions under which \( E, EAE, \) and \( A \) are simultaneously diagonalized.

Recall from Chapter 2, section 2.8, that for any \( n \times k \) matrix \( X \) of rank \( p \), an \( n \times l \) matrix \( Y \) of rank \( \nu = n - p \) is called complementary to \( X \) in case the \( n \times (k + l) \) augmented matrix \([X \ Y]\) has the full rank \( n \). Moreover, \( Y \) is called polar to \( X \) if, in addition, \( X'Y = 0 \). In what follows, the discussions will be confined to the case \( p = k \) and \( \nu = l \).

**Lemma 7.2.3.** Let \( X \) be \( n \times k \) of rank \( k \), and let \( Y \) be some \( n \times l \) matrix of rank \( l = n - k \), such that \( X'Y = 0 \) (i.e., \( Y \) is polar to \( X \)). Then the columns of \( X \) are linear combinations of the \( k \) eigenvectors of \( YY' \) associated with its \( k \) vanishing eigenvalues.

**Proof.** Let \( P \) be an orthogonal matrix diagonalizing \( YY' \) to

\[
P'YY'P = \begin{bmatrix} 0 & 0 \\ 0 & \Delta \end{bmatrix},
\]

where \( \Delta \) is a diagonal matrix of order \( l \) with positive diagonal elements. Let \( P \) be partitioned into its first \( k \) and last \( l \) columns, as \( P = [P_1 \mid P_2] \). Then from (7.2.21) we have \( P_1'YY'P_1 = 0 \) whence \( P_1'Y = 0 \); thus \( P_1 \) is in the row null space of \( Y \), which is the column space of \( X \); therefore since both span the same space we have \( X = P_1K \), where \( K \) is some \( k \times k \) nonsingular matrix. \( \square \)

A special case of interest is that in which \( X = \iota \) (a column of ones) and \( Y = W \). Since \( W'i = 0 \) we have \( W'e = W'y = W'\varepsilon \) (where \( b = y - Xb \) and \( b = X'y \)).

Defining (cf. von Neumann et. al., 1941) the mean-square successive difference by

\[
\delta^2 = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{n - 1} = \frac{e'WW'e}{n - 1} = \frac{e'Az}{n - 1}
\]

and the sample variance (the maximum-likelihood estimator) by

\[
s^2 = \frac{\sum_{i=1}^{n}e_i^2}{n} = \frac{e'\varepsilon}{n}
\]

(where \( E = I - \iota\iota' \)), upon putting \( \varepsilon = P\zeta \) we obtain the von Neumann ratio (von Neumann, 1941)

\[
\eta = \frac{\delta^2}{s^2} = \frac{n}{n - 1} \cdot \frac{e'Az}{e'\varepsilon} = \frac{n}{n - 1} \cdot \frac{\sum_{i=1}^{n-1}A\zeta_i^2}{\sum_{i=1}^{n-1}S_i^2}
\]

where the \( \lambda_i's \) are the positive eigenvalues of \( A = WW' \).

The following lemma asserts that under certain conditions \( Y \) may always be chosen to have a particular form in order for the diagonalization property of Lemma 7.2.3 to hold.
Lemma 7.2.4. The matrix $EW = (I - XX^\dagger)W$ is polar to $X$ if and only if the augmented $n \times (k + n - 1)$ matrix $[X, W]$ has rank $n$.

Proof. Certainly $X'(I - XX^\dagger)W = 0$, so it remains only to prove that $(I - XX^\dagger)W$ has rank $l = n - k$ if and only if $[X, W]$ has rank $n$.

First, let $[X, W]$ have rank $n$; then, since $I - XX^\dagger$ is idempotent of rank $n - k = l$, the matrix

$$
(I - XX^\dagger)[X, W] = [0, (I - XX^\dagger)W]
$$

has rank $l$, whence $(I - XX^\dagger)W$ has rank $l$.

Conversely, let $(I - XX^\dagger)W$ have rank $l$. Suppose rank $[X, W] < n$; since rank $W = n - 1$, the columns of $X$ must be in the column space of $W$, whence $X = WK$ where $K$ is some $(n - 1) \times k$ matrix of rank $k$. Therefore

$$
(I - XX^\dagger)WK = (I - XX^\dagger)X = 0,
$$

so $(I - XX^\dagger)W$ is orthogonal to $K$. Thus the matrix

$$
[W'(I - XX^\dagger), K],
$$

which has $n - 1$ rows, has rank $l + k = n$, which is impossible. Therefore $[X, W]$ has rank $n$. □

It follows from Lemma 7.2.4 that, since rank$([I, W]) = n$, as long as $X$ has the vector $i$ as one of its columns, the matrix $Y = (I - XX^\dagger)W$ will be polar to $X$. Lemma 7.2.9 will therefore apply to the matrix

$$
YY' = (I - XX^\dagger)WW'(I - XX^\dagger) = EAE
$$

where $E = I - XX^\dagger$. Then we shall have $y'YY'y = e'YY'e = e'EAE'e = e'Ae$.

Theorem 7.2.2. Let $P$ be an orthogonal matrix diagonalizing $A = WW'$ to the diagonal matrix $P'AP = \Lambda$, and let $X$ be an $n \times k$ matrix of rank $k$ such that the augmented matrix $[X, W]$ has rank $n$. Then $P$ also diagonalizes $(I - XX^\dagger)A(I - XX^\dagger)$ if and only if the columns of $X$ are linear combinations of $k$ eigenvectors of $A$. The $n - k$ positive eigenvalues of $EAE$ are equal to those of $A$ associated with the $n - k$ complementary eigenvectors of $A$.

Proof. Denote $E = I - XX^\dagger$. Let $P$ also diagonalize $EAE = EWW'E$; since $[X, W]$ has rank $n$, $Y = EW$ is polar to $X$ from Lemma 7.2.10, so from Lemma 7.2.9 it follows that the columns of $X$ are linear combinations of the $k$ eigenvectors of $YY' = EWW'E$ associated with its $k$ vanishing eigenvalues. Now since $P$ diagonalizes both $A$ and $EAE$ by assumption, its columns are eigenvectors of both, so the assertion follows.

Conversely, let $P = [P_1, P_2]$ be a partition of $P$ into its first $k$ and last $l$ columns, and let $X = P_1K$ where $K$ is some $k \times k$ nonsingular matrix. Then $E = I - XX^\dagger = I - P_1P_1'$, and given $P'AP = \Lambda$ we have

$$
EAE = (I - P_1P_1')P\Lambda P'(I - P_1P_1')
$$

$$
= (I - P_1P_1')[P_1, P_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} P_1' \\ P_2' \end{bmatrix} (I - P_1P_1')
$$

$$
= [0, P_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} 0 \\ P_2' \end{bmatrix} = P_2\Lambda_2 P_2'.
$$
Thus,

\[
P'EAE = \begin{bmatrix} P'_1 \\ P'_2 \end{bmatrix} P_2 \Lambda_2 P'_2 \begin{bmatrix} P_1 & P_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_2 \end{bmatrix}.
\]

\[\square\]

7.3. Distribution and Beta approximation of the Durbin-Watson statistic.

The exact distribution of the von Neumann ratio was computed and tabulated by Hart (1942), who presented plots of the density function for \( n = 3, \ldots, 7 \). It is clear from (7.2.8) and (7.2.23) that 3 is the smallest sample size for which either the von Neumann ratio or the Durbin-Watson statistic is defined, and in the latter case one requires \( n - k \geq 2 \); the graphs displayed by Hart may therefore be taken as a good indication of the density of the Durbin-Watson statistics for degrees of freedom \( n - k = 2, \ldots, 6 \). Her graph for \( n = 3 \) shows the density of the von Neumann ratio to be U-shaped; for \( n = 4 \) it is unimodal with an infinite-sloping cusp at the mode; the densities for \( n = 5 \) and 6 still have singularities, and even though singularities exist at higher sample sizes (as pointed out by von Neumann in Hart (1942)) they are not detectable by the eye, so that the curve has a good bell shape for \( n \geq 7 \). This suggests that one should not think of applying the Durbin-Watson test for degrees of freedom less than 6, but that for \( n - k \geq 6 \) a beta approximation to the statistic could be expected to be reasonably good. Since the beta distribution is characterized by its mean and variance, this means that in order to compute the approximation one must be able to compute the mean and variance of the Durbin-Watson statistic. This is made possible by the fact that it is distributed independently of its own denominator—a theorem proved independently by Pitman (1937) and von Neumann (1941). Since Pitman’s theorem is somewhat more general, the statement and proof of Lemma 7.3.1 below will follow his.

We note that the random variables \( \zeta_i^2 \) appearing in (7.2.12), (7.2.13), and (7.2.25) have, after division by \( \sigma^2 \), independent central chi-square distributions under the null hypothesis \( \rho = 0 \). We note also that the Durbin-Watson statistic is homogeneous of degree 0 in the \( \zeta_i^2 \)’s. The random variables

\[ W_i = \frac{\zeta_i^2}{2\sigma^2} \quad (i = 1, 2, \ldots, l = n - k) \]

have independent gamma distributions; consequently we may write the Durbin-Watson statistic as

\[ D(W) = \frac{\sum_{i=1}^{l} \nu_i W_i}{\sum_{i=1}^{l} W_i} \]

**Lemma 7.3.1 (Pitman-von Neumann).** Let \( W_i (i = 1, 2, \ldots, l) \) be independently distributed random variables with gamma densities

\[ f(w_i) = \frac{w_i^{a_i-1} e^{-w_i}}{\Gamma(a_i)} \quad (a_i > 0). \]

Let \( D(W) = D(W_1, W_2, \ldots, W_l) \) be homogeneous of degree 0, and define

\[ S(W) = \sum_{i=1}^{l} W_i, \]
Then $D(W)$ and $S(W)$ are independent.

Proof. The moment-generating function of the joint distribution of $S(W)$ and $D(W)$ is (for $|t_1| < 1$)

$$m(t_1, t_2) = \mathbf{E}\{e^{t_1 S(W) + t_2 D(W)}\}$$

$$= \int_0^\infty \cdots \int_0^\infty e^{t_1 S(w) + t_2 D(w)} \prod_{i=1}^l f(w_i)d w_i$$

$$= \frac{1}{\prod_{i=1}^l \Gamma(a_i)} \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^l \left[a_i^{\alpha_i-1} e^{-S(w)} e^{t_1 S(w) + t_2 D(w)}\right] d w_1 \cdots d w_l$$

$$= \frac{1}{\prod_{i=1}^l \Gamma(a_i)} \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^l \left[a_i^{\alpha_i-1} e^{(t_1-1) S(w) + t_2 D(w)}\right] d w_1 \cdots d w_l.$$

Now performing the transformation of variables $z_i = (1-t_1)w_i$, and using the homogeneity of $D(w)$, this becomes

$$m(t_1, t_2) = \frac{(t_1 - 1)^{S(a)}}{\prod_{i=1}^l \Gamma(a_i)} \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^l z_i^{\alpha_i-1} e^{-S(w)+t_2 D(z)} d z_1 \cdots d z_l$$

$$= m_1(t_1)m_2(t_2),$$

which factors into two terms, one (to the left of the multiple integral) depending only on $t_1$, and the remaining term depending only on $t_2$. Consequently, $S(W)$ and $D(W)$ are independent. 

Since by Lemma 7.3.1, under the null hypothesis $\rho = 0$ $d$ is distributed independently of its own denominator $\sum_{i=1}^{n-k} \zeta_i^2$, we have

$$\mathbf{E}\left\{\sum_{i=1}^{n-k} \nu_i \zeta_i^2\right\}^r = \mathbf{E}\left\{d \cdot \sum_{i=1}^{n-k} \zeta_i^2\right\}^r = \mathbf{E}\{d\}^r \mathbf{E}\left\{\sum_{i=1}^{n-k} \zeta_i^2\right\}^r$$

for $r = 1, 2, \ldots$

hence

$$(7.3.1) \quad \mathbf{E}\{d\}^r = \frac{\mathbf{E}\left\{\sum_{i=1}^{n-k} \nu_i \zeta_i^2\right\}^r}{\mathbf{E}\left\{\sum_{i=1}^{n-k} \zeta_i^2\right\}^r}$$

for $r = 1, 2, \ldots$.

In particular, for $r = 1$, we have for the numerator of (7.3.1)

$$\mathbf{E}\{\sum_{i=1}^{n-k} \nu_i \zeta_i^2\} = \sum_{i=1}^{n-k} \nu_i,$$

since each $\zeta_i^2$, being a chi-square variable with one degree of freedom, has mean 1. Likewise for the denominator of (7.3.1),

$$\mathbf{E}\left\{\sum_{i=1}^{n-k} \zeta_i^2\right\} = n - k,$$
by the same reasoning as above, or because \( \sum_{i=1}^{n-k} \zeta_i^2 \) has a chi-square distribution with \( n - k \) degrees of freedom whose mean is then \( n - k \). Thus,

\[
(7.3.2) \quad \mathbb{E}\{d\} = \frac{\sum_{i=1}^{n-k} \nu_i}{n - k} = \bar{\nu}.
\]

The same procedure cannot be applied to the variance of (7.2.12), however, only to its second moment about zero. To see this, take the case \( n - k = 1 \). Then

\[
d = \nu_1 \bar{\zeta}_1^2 = \nu_1
\]

whence \( \text{Var}\{d\} = 0 \) while the ratio of variances is \( 2\nu_1^2/2 = \nu_1^2 \).

The deviation of \( d \) from its mean is, from (7.2.12) and (7.3.2),

\[
d - \mathbb{E}\{d\} = \frac{\sum_{i=1}^{n-k} (\nu_i - \bar{\nu}) \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2}.
\]

Now \( d - \mathbb{E}\{d\} \) is still distributed independently of \( \sum_{i=1}^{n-k} \zeta_i^2 \), so

\[
(7.3.3) \quad \text{Var}\{d\} = \mathbb{E}\{(d - \mathbb{E}\{d\})^2\} = \frac{\mathbb{E}\{(\sum_{i=1}^{n-k} (\nu_i - \bar{\nu}) \zeta_i^2)^2\}}{\mathbb{E}\{(\sum_{i=1}^{n-k} \zeta_i^2)^2\}}.
\]

It remains to compute the two second moments in the numerator and denominator of (7.3.3).

Computing the second moment in the denominator of (7.3.3) is straightforward, since \( \sum_{i=1}^{n-k} \zeta_i^2 \) is distributed as chi-square with \( n - k \) degrees of freedom, whose mean and variance are \( n - k \) and \( 2(n - k) \) respectively; the second moment about zero is then

\[
(7.3.4) \quad \mathbb{E}\left\{\sum_{i=1}^{n-k} \zeta_i^2\right\}^2 = \text{Var}\left\{\sum_{i=1}^{n-k} \zeta_i^2\right\} + \left[\mathbb{E}\left\{\sum_{i=1}^{n-k} \zeta_i^2\right\}\right]^2
\]

\[
= 2(n - k) + (n - k)^2
\]

\[
= (n - k)(n - k + 2)
\]

(compare equation (5.1.14) in §5.1 of Chapter 5, as well as the Corollary to Lemma 5.2.3).

To compute the second moment in the numerator of (7.3.3) we need to compute the cumulants of the chi-square distribution, since these have the property that the cumulant of a sum is equal to the sum of the cumulants. The cumulant-generating function\(^5\) \( K_X(t) \) of a random variable \( X \) is defined in terms of the moment-generating function \( m_X(t) = \mathbb{E}\{e^{tX}\} \) of \( X \) by

\[
K_X(t) = \log m_X(t) = \log \mathbb{E}\{e^{tX}\};
\]

hence if $X$ and $Y$ are two independent random variables,

$$K_{X+Y}(t) = \log \mathbb{E}\{e^{t(X+Y)}\}$$

$$= \log \mathbb{E}\{e^{tX}e^{tY}\}$$

$$= \log(\mathbb{E}\{e^{tX}\}\mathbb{E}\{e^{tY}\})$$

$$= \log(\mathbb{E}\{e^{tX}\}) + \log(\mathbb{E}\{e^{tY}\})$$

$$= K_X(t) + K_Y(t).$$

(7.3.5)

The $r$th cumulant of a random variable $X$ is defined by

$$\kappa_r(X) = \left. \frac{d^rK_X(t)}{dt^r} \right|_{t=0}.$$

From this definition and (7.3.5) it is clear that for two independent random variables $X$ and $Y$,

$$\kappa_r(X + Y) = \kappa_r(X) + \kappa_r(Y).$$

The first two cumulants of any random variable $X$ are equal to its mean and variance respectively. In the case of the first cumulant,

$$\kappa_1(X) = \left. \frac{m'_X(t)}{m_X(t)} \right|_{t=0} = m'_X(0) = \mu'_1(X),$$

where $\mu'_r(X)$ is Pearson’s notation for the $r$th moment of $X$ about zero. [Note that $\mu'_r(X)$ must not be interpreted as the derivative of $\mu_r(X)$ with respect to $X$; on the other hand, in these formulas, $m'_X(t)$ and $m''_X(t)$ do denote the first and second derivatives of $m_X(t)$ with respect to $t.$] Thus, the first cumulant is equal to the mean. In the case of the second cumulant, we have

$$\kappa_2(X) = \left. \frac{m_X(t)m'_X(t) - [m'_X(t)]^2}{m_X(t)^2} \right|_{t=0}$$

$$= m'_X(0) - [m'_X(0)]^2$$

$$= \mu'_2(X) - [\mu'_1(X)]^2 = \mu_2(X)$$

where $\mu_r(X)$ is Pearson’s notation for the $r$th moment of $X$ about the mean, $\mu'_1(X)$. Thus the second cumulant of any distribution is equal to its variance.

Now, we know that $\zeta^2$ is distributed as chi-square with one degree of freedom, hence its variance is 2; thus, the variance of $(\nu_i - \bar{\nu})\zeta^2_i$ is $2(\nu_i - \bar{\nu})^2$; consequently, since this variance is equal to the second cumulant, $\kappa_2$, it follows by the additive property of cumulants that $\sum_{i=1}^{n-k}(\nu_i - \bar{\nu})\zeta^2_i$ has its second cumulant, and therefore its variance, equal to $2\sum_{i=1}^{n-k}(\nu_i - \bar{\nu})^2$. But $\sum_{i=1}^{n-k}(\nu_i - \bar{\nu})\zeta^2_i$ has zero mean, hence this is also its second moment about zero. Thus we conclude:

(7.3.6)

$$\mathbb{E}\left\{\sum_{i=1}^{n-k}(\nu_i - \bar{\nu})\zeta^2_i\right\}^2 = 2\sum_{i=1}^{n-k}(\nu_i - \bar{\nu})^2.$$

From (7.3.4) and (7.3.6) we then have

(7.3.7)

$$\text{Var}\{d\} = \frac{2\sum_{i=1}^{n-k}(\nu_i - \bar{\nu})^2}{(n-k)(n-k+2)}.$$
Now the beta distribution over the interval $[0, 1]$ has a density (cf. Johnson & Kotz 1970, Vol. 3, Ch. 24, p. 37)

$$p_X(x) = \frac{1}{B(p, q)} x^{p-1}(1-x)^{q-1} \quad (0 \leq x \leq 1, \ p > 0, \ q > 0),$$

whose $r$th moment about zero is

$$B(p + r, q) \quad \text{where} \quad B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p + q)} \quad \text{and} \quad \Gamma(m + 1) = m\Gamma(m) \text{ for } m \geq 0. \quad (7.3.8)$$

The transformation

$$x = \frac{z - a}{b - a}$$

brings this over to the beta distribution on the interval $[a, b]$ with density

$$p_z(z) = \frac{1}{B(p, q)} \frac{(z - a)^{p-1}(b-z)^{q-1}}{(b-a)^{p+q-2}} \quad (a \leq z \leq b, \ p > 0, \ q > 0). \quad (7.3.9)$$

Our problem is to show how the density of the Durbin-Watson statistic (7.2.11) may be approximated under the null hypothesis by the density of the random variable $Z$ with density (7.3.9).

The beta density (7.3.9) has four parameters: $a, b, p, \text{and} q$. Parameters $a$ and $b$ are the lower and upper bound respectively of the support of the beta density. These necessarily correspond to the smallest and largest of the eigenvalues $\nu_i$ of $EAX$ as is clear from the formula (7.2.12), since for any fixed value (say 1) of the denominator, if we assume that the eigenvalues are ordered as $0 < \nu_1 < \nu_2 < \ldots < \nu_{n-k}, \sum_{i=1}^{n-k} \nu_i \zeta_i^2$ is minimized subject to $\sum_{i=1}^{n-k} \zeta_i^2 = 1$ when $\zeta_i^2 = 1$ and $\zeta_i^2 = 0$ for $i \neq 1$; likewise, $\sum_{i=1}^{n-k} \nu_i \zeta_i^2$ is maximized subject to $\sum_{i=1}^{n-k} \zeta_i^2 = 1$ when $\zeta_i^2 = 1$ and $\zeta_i^2 = 0$ for $i \neq n - k$. In the first case $d = \nu_1$ and in the second case $d = \nu_{n-k}.$

There remain only the parameters $p$ and $q$. The problem is solved if one can put these parameters into one-to-one correspondence with the mean and variance of the beta distribution.

The $r$th moment about zero of the standard beta distribution over the interval $[0, 1]$ is

$$\mu_r'(X) = \frac{\Gamma(p + r)\Gamma(q)}{\Gamma(p + r + q)} \cdot \frac{\Gamma(p + q)}{\Gamma(p)\Gamma(q)} = \frac{(p + r - 1)! (p + q - 1)!}{(p + q + r - 1)! (p - 1)!}.$$}

Thus the first two moments are

$$\mu_1'(X) = \frac{p}{p + q} \quad \text{and} \quad \mu_2'(X) = \frac{(p + 1)p}{(p + q + 1)(p + q)}.$$}

The variance is therefore

$$\mathbb{E}((X - \mathbb{E}\{X\})^2) = \mathbb{E}\{X^2\} - (\mathbb{E}\{X\})^2$$

$$= \frac{(p + 1)p}{(p + q + 1)(p + q)} - \frac{p^2}{(p + q)^2} = \frac{pq}{(p + q + 1)(p + q)^2}.$$
From the transformation

\[ Z = a + (b - a)X \]

we readily compute the mean and variance of the random variable \( Z \) (with density (7.3.9)) to be

\[
\begin{align*}
\mathbf{E}\{Z\} &= a + (b - a)\frac{p}{p + q} \\
\text{Var}\{Z\} &= \frac{(b - a)^2 pq}{(p + q)^2(p + q + 1)}.
\end{align*}
\]  

(7.3.10)

Given \( a = \nu_1 \) and \( b = \nu_{n-k} \), these equations must be solved for \( p \) and \( q \) in terms of the mean and variance. Using Pearson’s notation \( \mu'_1(Z) = \mathbf{E}\{Z\} \) and \( \mu_2(Z) = \text{Var}\{Z\} \) for the mean and variance of \( Z \) (but omitting the argument \( Z \) for brevity), equations (7.3.10) may be rewritten as

\[
\begin{align*}
\frac{\mu'_1 - a}{b - a} &= \frac{p}{p + q} \quad \text{(hence } \frac{b - \mu'_1}{b - a} = \frac{q}{p + q}) ; \\
\frac{\mu_2}{(b - a)^2} &= \frac{p}{p + q} \cdot \frac{q}{p + q} \cdot \frac{1}{p + q + 1}.
\end{align*}
\]

(7.3.11)

Substituting the two expressions in the top line of (7.3.11) into the second line we obtain

\[
p + q = \frac{(\mu'_1 - a)(b - \mu'_1)}{\mu_2} - 1.
\]

(7.3.12)

Now we substitute (7.3.12) back into the top two expressions of (7.3.11) to obtain

\[
\begin{align*}
p &= \left( \frac{(\mu'_1 - a)(b - \mu'_1)}{\mu_2} - 1 \right) \frac{\mu'_1 - a}{b - a} ; \\
q &= \left( \frac{(\mu'_1 - a)(b - \mu'_1)}{\mu_2} - 1 \right) \frac{b - \mu'_1}{b - a}.
\end{align*}
\]

(7.3.13)

Setting \( a = \nu_1 \) and \( b = \nu_{n-k} \) in these two formulas, and substituting formula (7.3.2) for the mean, \( \nu \), and (7.3.7) for the moments \( \mu'_1 \) and \( \mu_2 \) respectively, we obtain the desired beta approximation of the Durbin-Watson statistic (7.2.12).

### 7.4 Bias in estimation of sampling variances.

If \( \Omega = I \) in the regression model

\[
y = X\beta + \varepsilon; \quad \mathbf{E}\{\varepsilon\} = 0; \quad \mathbf{E}\{\varepsilon\varepsilon'\} = \sigma^2\Omega,
\]

(7.4.1)

the variance matrix of the least-squares estimator

\[
b = (X'X)^{-1}X'y
\]

(7.4.2)

is

\[
\text{Var}\{b\} = \mathbf{E}\{(b - \beta)(b - \beta)'\} = \sigma^2(X'X)^{-1},
\]

(7.4.3)
and the best quadratic unbiased estimator of $\sigma^2$ is

\begin{equation}
(7.4.4) \quad s^2 = \frac{e'e}{n - k} = \frac{\sum_{t=1}^{n} e_t^2}{n - k}
\end{equation}

where

\begin{equation}
(7.4.5) \quad e = y - Xb
\end{equation}

(cf. Hsu, 1938; Theil, 1961, pp. 532–6). The best quadratic unbiased estimator of $\text{Var}\{b\}$ in (7.4.3) is then

\begin{equation}
(7.4.6) \quad s^2(X'X)^{-1} = \frac{e'e}{n - k} (X'X)^{-1}.
\end{equation}

Now consider the following specification problem. Suppose $\Omega \neq I$ in (7.4.1) and suppose the investigator uses (7.4.6) to estimate $\text{Var}\{b\}$, which is now

\begin{equation}
(7.4.7) \quad \text{Var}\{b\} = E\{(b - \beta)(b - \beta)'\} = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}
\end{equation}

rather than the expression given in (7.4.3). What can be said about the direction and magnitude of the bias in (7.4.6) as an estimator of (7.4.7)?

Define the bias by

\begin{equation}
(7.4.8) \quad B = E\{s^2(X'X)^{-1}\} - \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}.
\end{equation}

If the matrix $B$ is positive semi-definite, we will say that the sampling variances of the regression coefficients $b$ are biased upwards, or overestimated, and if $B$ is negative semi-definite, that they are biased downwards, or underestimated. The following basic result is due to Watson (1955, pp. 328–9).

**Lemma 7.4.1 (Watson).** The estimator $s^2(X'X)^{-1}$ of (7.4.6) is biased upwards or downwards as an estimator of $\text{Var}\{b\}$ in (7.4.7) according as the matrix

\begin{equation}
(7.4.9) \quad \tilde{\nu}X'X - X'\Omega X \quad \left( \text{where } \tilde{\nu} = \sum_{t=1}^{n} \frac{\nu_t}{n - k} \right)
\end{equation}

is positive or negative semi-definite, where $\nu_1, \nu_2, \ldots, \nu_{n-k}$ are the $n-k$ eigenvalues of the matrix $[I - X(X'X)^{-1}X']\Omega$ apart from $k$ zeros.

**Proof.** Define $E = I - X(X'X)^{-1}X'$; then

\begin{equation}
\epsilon = y - Xb = y - X(X'X)^{-1}X'y = Ey = E(X\beta + \varepsilon) = \varepsilon,
\end{equation}

and from (7.4.4) and (7.3.14),

\begin{equation}
(7.4.10) \quad E\{s^2\} = \frac{E\{e'e\}}{n - k} = \frac{E\{e'e\}}{n - k} = \frac{\text{tr}(E\varepsilon\varepsilon')}{n - k} = \frac{\sigma^2\text{tr}(E\Omega)}{n - k} = \frac{\sigma^2\text{tr}(E\Omega E)}{n - k}
\end{equation}

where use is made of the linearity of the trace operation and the fact that $\text{tr}(AB) = \text{tr}(BA)$. $E$ and therefore $E\Omega$ has rank $n - k$, hence has $n - k$ nonzero eigenvalues, whose sum is $\text{tr}(E\Omega)$. Denoting these by $\nu_1, \nu_2, \ldots, \nu_{n-k}$, it follows that

\begin{equation}
(7.4.11) \quad E\{s^2\} = \sigma^2 \frac{\sum_{t=1}^{n-k} \nu_t}{n - k};
\end{equation}
and substituting (7.4.11) in (7.4.8), the lemma follows upon premultiplying and postmultiplying (7.4.8) by $X'X$. \qed

Note that in the proof of Lemma 7.4.1 we used the fact that if $M$ is any $k \times k$ nonsingular matrix, the symmetric $k \times k$ matrix $B$ is positive semi-definite if and only if $M'BMM$ is. This follows from the fact that $x'Bx > 0$ if and only if $y'M'BMy > 0$ where $x = My$, and $x = 0$ if and only if $y = 0$.

The lemma, as stated, is not very useful in applications, since it requires knowledge of the eigenvalues of $E\Omega$. The following proposition applies the criterion to the case in which Anderson’s criterion of Theorem 2.6.9 is fulfilled.

**Theorem 7.4.1.** Let $P$ be an orthogonal matrix diagonalizing the matrix $\Omega$ of (6.1.13) to
\[(7.4.12) \quad PP' = \Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}, \quad P'P = I\]
and let $P_K$ be an $n \times k$ matrix formed from the $k$ columns $p_j$ of $P$ for which $j \in K$, where $K$ is a subset of exactly $k$ integers from the set $N = \{1, 2, \ldots, n\}$. Let the $n \times k$ matrix $X$ satisfy
\[(7.4.13) \quad X = P_KK\]
where $K$ is a $k \times k$ nonsingular matrix. Then a necessary and sufficient condition for overestimation (i.e., for $B$ of (7.4.8) to be positive semi-definite) is that
\[(7.4.14a) \quad \lambda_j \leq \frac{\sum_{i \in N \setminus K} \lambda_i}{n - k} \quad \text{for all } j \in K\]
with strict inequality holding for some $j \in K$; likewise, a necessary and sufficient condition for underestimation (i.e., for $B$ of (7.4.8) to be negative semi-definite) is that
\[(7.4.14b) \quad \lambda_j \geq \frac{\sum_{i \in N \setminus K} \lambda_i}{n - k} \quad \text{for all } j \in K\]
with strict inequality holding for some $j \in K$.

**Proof.** From the last term of (7.4.10), and making use of (7.4.11), (7.4.12), (7.4.13), we obtain
\[
\sum_{i=1}^{n-k} \lambda_i = \text{tr}(E\Omega) = \text{tr}(\Omega) - \text{tr}((X(X'X)^{-1}X')\Omega)) = \text{tr}(\Lambda P') - \text{tr}(P_KP_K'\Omega) = \text{tr}(\Lambda P') - \text{tr}(P_K'\Omega P_K) = \text{tr}(\Lambda) - \text{tr}(\Lambda_K) = \sum_{i \in N \setminus K} \lambda_i,
\]
where $\Lambda_K$ is the diagonal matrix defined by $\Lambda_K = P_K'\Omega P_K$. Substituting (7.4.15) in (7.4.11) and (7.4.13) in (7.4.9), and pre- and post-multiplying the result by $K^{-1}$ and $K^{-1}$ respectively, one obtains
\[(7.4.16) \quad K^{-1}X'BXK^{-1} = \frac{\sum_{i \in N \setminus K} \lambda_i}{n - k} - I_k - \Lambda_K,
\]
whence (7.4.14) follows. \qed

We obtain immediately the
Corollary. Let the eigenvalues $\lambda_j$ of $\Omega$ be distinct and arranged in descending order:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n$$

and let (7.4.13) hold with $K = \{1, 2, \ldots, k\}$. Then $B$ is negative definite, hence the estimator (7.4.6) is biased downwards.

This corollary is still not in a form that is useful in applications, since the $\lambda_j$’s depend on the elements of $\Omega$. To obtain useful results, we must specialize further the assumptions concerning $\Omega$. In particular, we now take up the case in which $\Omega = \Phi(\rho)$, where $\Phi(\rho)$ is given by (7.1.4).

From (7.1.4) and (7.1.15) we have

$$(7.4.17a) \quad P^T \Phi(\rho)^{-1} P = (1 + \rho^2)I - 2\rho \Upsilon,$$

whence from (7.1.44) the eigenvalues of $\Phi(\rho)^{-1}$ are

$$(7.4.17b) \quad \psi_j(\rho)^{-1} = 1 + \rho^2 - 2\rho \cos \frac{j\pi}{n}, \quad (j = 0, 1, \ldots, n - 1)$$

where, in view of (7.1.34) and (7.2.15),

$$(7.4.18a) \quad (1 - \rho^2) = \psi_0(\rho)^{-1} < \psi_1(\rho)^{-1} < \ldots$$

$$< \psi_{n-1}(\rho)^{-1} < \psi_n(\rho)^{-1} = (1 + \rho^2)^2 \quad \text{for } \rho > 0$$

and

$$(7.4.18b) \quad (1 - \rho^2) = \psi_0(\rho)^{-1} > \psi_1(\rho)^{-1} > \ldots$$

$$> \psi_{n-1}(\rho)^{-1} > \psi_n(\rho)^{-1} = (1 + \rho^2)^2 \quad \text{for } \rho < 0$$

$[\psi_n(\rho)^{-1}$ is not an eigenvalue of $\Phi(\rho)^{-1}$; it is included in (7.4.18) only for convenience.] Thus, the eigenvalues of $\Phi(\rho)$ satisfy

$$(7.4.19a) \quad \frac{1}{(1 - \rho^2)} = \psi_0(\rho) > \psi_1(\rho) > \ldots > \psi_{n-1}(\rho) > \psi_n(\rho) = \frac{1}{(1 + \rho)^2} \quad \text{for } \rho > 0$$

and

$$(7.4.19b) \quad \frac{1}{(1 - \rho^2)} = \psi_0(\rho) < \psi_1(\rho) < \ldots < \psi_{n-1}(\rho) < \psi_n(\rho) = \frac{1}{(1 + \rho)^2} \quad \text{for } \rho < 0.$$

This is summarized in

Theorem 7.4.2. Let $\text{Var}(\varepsilon) = \sigma^2 \Phi(\rho)$ where $\Phi(\rho)$ is given by (7.1.4), and let $P$ be the orthogonal matrix diagonalizing $\Phi(\rho)$ to $P^T \Phi(\rho) P = \Psi(\rho) = \text{diag}\{\psi_j(\rho)\}$

where

$$\psi_j(\rho) = \frac{1}{1 + \rho^2 - 2\rho \cos \frac{j\pi}{n}} \quad (j = 0, 1, \ldots, n - 1).$$

Let $P = [P_1, P_2]$ be a partition of $P$ into its first $k$ and last $n - k$ columns, where $P$ is given by (7.1.47), and let $X = P_1 K$ for some $k \times k$ matrix $K$. Then the matrix $B$ of (7.4.8) is positive definite if and only if $\rho < 0$, and negative definite if and only if $\rho > 0$, i.e., the sampling variances of the least-squares estimators $b$ are overestimated by (7.4.6) if and only if $\rho < 0$, and underestimated if and only if $\rho > 0$. 
7.5 Exercises.

1. An investigator is employing the model of Theorem 2.6.8:

\[ y = X\beta + \varepsilon = [\iota, Z]\beta + \varepsilon \quad \text{where} \quad \mathbf{E}\{\varepsilon\} = 0 \quad \text{and} \quad \mathbf{E}\{\varepsilon\varepsilon'\} = \sigma^2 \Omega, \]

where \( \iota \) denotes a column of \( n \) ones, and \( Z \) an \( n \times (k-1) \) matrix such that the rank of \( X = [\iota, Z] \) is equal to \( k \), and

\[ \Omega = (1 - \rho)I + \rho \mu \mu', \quad \text{where} \quad -\frac{1}{n-1} < \rho < 1. \]

This investigator estimates \( \beta \) by ordinary least squares, \( b = (X'X)^{-1}X'y \), and \( \sigma^2 \) by

\[ s^2 = \frac{\sum_{t=1}^{n} \varepsilon_i^2}{n - k}, \]

where \( \varepsilon = y - Xb \).

(a) Find the expected value of \( s^2 \).

(b) Defining the bias in estimation of sampling variances by

\[ B(\rho) = \mathbf{E}\{s^2\}(X'X)^{-1} - \text{Var}\{b\}, \]

determine the conditions under which the sampling variances of the individual regression coefficients \( b_j \) (the diagonal elements of \( s^2(X'X)^{-1} \)) are biased upward or downward as estimates of the true variances \( \text{Var}\{b_j\} \), for each \( j = 1, 2, \ldots, k \).

7.6 References.


1958. xii, 433 pp.


