Correlation of Residuals in Successive Fittings with Least-Squares

BU-242-M | Robert Jacobsen | June, 1967

Abstract

This paper derives the covariance relations of the residuals in successive least-squares fits, with application to tests of heteroscedasticity.

Biometrics Unit, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York.

Correlation of Residuals in Successive Fittings with Least-Squares

BU-242-M Robert Jacobsen June, 1967

We give some simplified proofs and extensions of results in A. Hedayat's paper No. BU-135.

Let V be the observation space of dim. N, $\mathcal{Y}$ the observed point.

Let $E_\theta \mathcal{Y} = X\,\theta$, $\theta \in \Theta = R^p$, $X: \Theta \to V$ linear.

Let Ω denote the mean space, Im X, and Cov $\mathcal{Y}$ = D, where D is diagonal with respect to the orthonormal standard basis $e_1, \cdots, e_N$.

Denote V_i = the span of $\{e_1, \cdots, e_i\}$, and $\Omega_i = P_{V_i}\Omega$, where P_W denotes orthogonal projection onto $W \subset V$.

We are concerned with computing the covariance relations among the least-squares estimates of $E\mathcal{Y}$ and the residuals based on different numbers of observations.

(1) Now $\operatorname{cov}\left[(e_k, P_{V_i - \Omega_i} P_{V_i} \mathcal{Y}), (e_\ell, P_{V_j - \Omega_j} P_{V_j} \mathcal{Y}) \right]$

$$k = 1, \cdots, i; \quad \ell = 1, \cdots, j; \quad 1 \le i \le j \le N,$$

is the covariance between the k^{th} coordinate of the residual vector, based on a fit to the 1^{st} i observations, and the ℓ^{th} coordinate of the residual based on the 1^{st} j observations.

Biometrics Unit, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York.

$(e_k, P_{V_i-\Omega_i} P_{V_i} \mathcal{Y}) = (P^{1}_{V_i-\Omega_i} P^{1}_{V_i} e_k, \mathcal{Y}) = (P_{V_i-\Omega_i} e_k, \mathcal{Y})$, as a projection is self-adjoint.

$P_{V_i-\Omega_i} e_k = (I - P_{\Omega_i}) e_k$, as $e_k \in V_i$. So (1) becomes $(P_{V_i-\Omega_i} e_k, D\, P_{V_j-\Omega_j} e_\ell)$

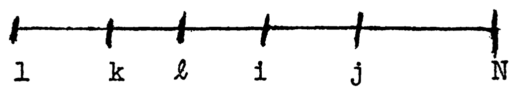
$$= (e_k, D\, e_\ell) - (e_k, D\, P_{\Omega_j} e_\ell) - (D\, e_\ell, P_{\Omega_i} e_k) + (P_{\Omega_i} e_k, D\, P_{\Omega_j} e_\ell), \qquad (2)$$

by the definition of cov $\mathcal{Y}$.

Evaluation of (2).

Assume $D = \sigma^2 I$.

Case 1.



$1 \le k \le i < \ell \le j \le N$

Write $e_k = P_{\Omega_i} e_k + P_{V_i-\Omega_i} e_k$

Now $V_i - \Omega_i \perp \Omega_i$ and $V_j - V_i$.

But $\Omega_j \subset \Omega_i \oplus V_j - V_i$.

So $V_i - \Omega_i \perp \Omega_j$.

(4) Hence, $(e_k, P_{\Omega_j} e_\ell) = (P_{\Omega_i} e_k, P_{\Omega_j} e_\ell) + (P_{V_i-\Omega_i} e_k, P_{\Omega_j} e_\ell)$

$$= (P_{\Omega_i} e_k, P_{\Omega_j} e_\ell).$$

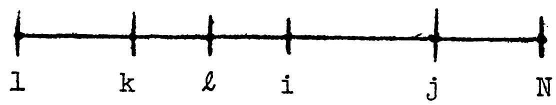
(5) Further, as $i < \ell$, $(e_\ell, P_{\Omega_i} e_k) = 0$.

(6) And, as $k < \ell$, $(e_k, e_\ell) = 0$.

So (2) becomes 0.

Therefore, any component of the residual based on the first i observations is uncorrelated with any component > i of the residual based on the first j observations, $j > i$, in the homoscedastic case.

<u>Case 2.</u>

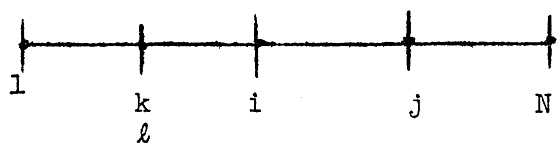


$1 \le k < \ell \le i \le j \le N$

(4) and (6) still hold.

So $(2) = -(e_\ell, P_{\Omega_i} e_k)\, \sigma^2$, which doesn't depend on j.

<u>Case 3.</u>



$1 \le k = \ell \le i \le j \le N$

(4) still holds.

So $(2) = \sigma^2 \left\{(e_\ell, e_\ell) - (e_\ell, P_{\Omega_i} e_\ell)\right\}$, which doesn't depend on j.

(7) Thus, $\text{var}\, (e_\ell, P_{V_i - \Omega_i} P_{V_i} y) = (1 - \|P_{\Omega_i} e_\ell\|^2)\, \sigma^2$

A formula for the correlation between two residuals can be given.

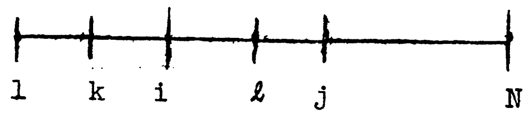
(8) $$\rho_{k,i,\ell,j} = \frac{-(e_\ell, P_{\Omega_i} e_k)}{\left[1 - (e_\ell, P_{\Omega_i} e_\ell)\right]^{\frac{1}{2}} \left[1 - (e_k, P_{\Omega_j} e_k)\right]^{\frac{1}{2}}}$$

for $1 \le k < \ell \le i \le j \le N$.

Evaluation of (2).

Assume $D = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_N^2 \end{pmatrix}$.

Case 1.



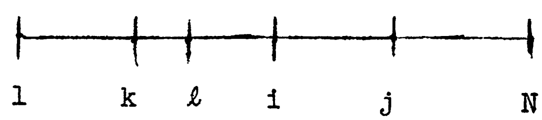
$1 \leq k \leq i < \ell \leq j \leq N$.

The 1st and 3rd terms of (2) vanish. (2) becomes

$$(P_{\Omega_i} e_k, D P_{\Omega_j} e_\ell) - (e_k, D P_{\Omega_j} e_\ell) = (D P_{\Omega_i} e_k, P_{\Omega_j} e_\ell) - \sigma_k^2 (e_k, P_{\Omega_j} e_\ell) \qquad (9)$$

which is not, in general, zero.

Case 2.

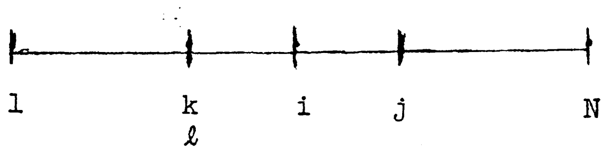


$1 \leq k < \ell \leq i \leq j \leq N$.

The 1st term of (2) vanishes. (2) becomes

$- (e_k, D P_{\Omega_j} e_\ell) - (D e_\ell, P_{\Omega_i} e_k) + (P_{\Omega_i} e_k, D P_{\Omega_j} e_\ell)$, which depends on j.

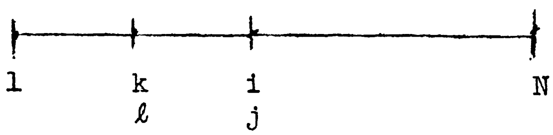
Case 3.



$1 \le k = \ell \le i < j \le N.$

(2) remains unchanged.

Case 4.



$1 \le k = \ell \le i = j \le N.$

The 2nd and 3rd terms of (2) become identical. (2) becomes

$$(10) \quad \sigma_\ell^2 - 2(e_\ell, D\, P_{\Omega_i}\, e_\ell) + (P_{\Omega_i}\, e_\ell, D\, P_{\Omega_i}\, e_\ell) = \mathrm{var}(e_\ell, P_{V_i - \Omega_i}\, P_{V_i}\, \mathcal{Y})$$

Now to investigate

$$(11) \quad \mathrm{cov}\left[(e_k, P_{\Omega_i}\, P_{V_i}\, \mathcal{Y}), (e_\ell, P_{V_j - \Omega_j}\, P_{V_j}\, \mathcal{Y})\right]$$

the covariance between the k^{th} coordinate of the estimated mean vector based on the first i observations and the ℓ^{th} coordinate of the residual based on the first j observations.

$$k = 1, \cdots, i; \quad \ell = 1, \cdots, j; \quad 1 \le i \le N; \; 1 \le j \le N.$$

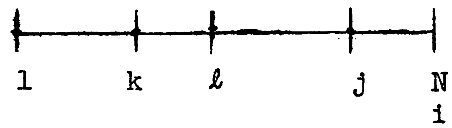
(11) becomes $(P_{\Omega_i}\, e_k, D\, P_{V_j - \Omega_j}\, e_\ell) = (P_{\Omega_i}\, e_k, D(I - P_{\Omega_j})\, e_\ell)$

$$= (P_{\Omega_i}\, e_k, D\, e_\ell) - (P_{\Omega_i}\, e_k, D\, P_{\Omega_j}\, e_\ell) \qquad (12)$$

Evaluation of (12).

Assume $D = \sigma^2 I$.

Case 1.



$1 \le k, \ell, j \le N$
$i = N, \ell \le j$

(12) becomes

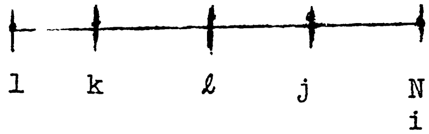
$$\sigma^2 \left\{ (P_{\Omega_N} e_k, e_\ell) - (P_{\Omega_N} e_k, P_{\Omega_j} e_\ell) \right\} = \sigma^2 (P_{\Omega_N} e_k, P_{V_j - \Omega_j} e_\ell)$$

But $V_j - \Omega_j \perp \Omega_j$ and $\perp V_N - V_j$. And $\Omega_N \subset \Omega_j \oplus V_N - V_j$. So $V_j - \Omega_j \perp \Omega_N$.
Hence above equals 0.

Evaluation of (12).

Assume $D = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_N^2 \end{pmatrix}$.

Case 1.



$1 \leq k, \ell, j \leq N$
$i = N, \ell \leq j.$

(12) becomes

$$\sigma_\ell^2 (P_{\Omega_N} e_k, e_\ell) - (P_{\Omega_N} e_k, D\, P_{\Omega_j} e_\ell),$$

which is not zero, in general.

Let $f_n = (e_n, P_{V_n - \Omega_n} P_{V_n} y)$

Var $f_n = \left\{1 - \|P_{\Omega_n} e_n\|^2\right\} \sigma^2 = C_n\, \sigma^2$, under homoscedasticity assumption.

Var $f_n = \sigma_n^2 - 2\sigma_n^2 (e_n, P_{\Omega_n} e_n) + (P_{\Omega_n} e_n, D\, P_{\Omega_n} e_n) = C_n' \sigma_n^2$, under heteroscedasticity assumption.

Let $d_n = \dfrac{f_n}{\sqrt{c_n}}$.

Var $d_n = \sigma^2$, under homoscedasticity assumption.

$= \sigma_n^2 \dfrac{c_n'}{c_n}$, under heteroscedasticity assumption.

Under homoscedasticity assumption, the d_n are uncorrelated, with constant var. σ^2, $n = r + 1, \cdots, N$, where r = rank of X. Under heteroscedasticity assumption, the d_n are correlated, with $\text{cov}(d_n, d_{n+1}) = \dfrac{1}{\sqrt{c_n c_{n+1}}} \text{cov}(f_n, f_{n+1})$, and var $d_n = \sigma_n^2 \frac{C_n'}{C_n}$.

The d's have expectation 0, under both hypotheses. If heteroscedasticity holds,

$P\{|d_{n+1}| > |d_n| > \frac{1}{2}$ if

$$\left| \operatorname{cor} (d_{n+1}, d_n) \sqrt{\frac{\operatorname{var} d_{n+1}}{\operatorname{var} d_n}} \right| > 1.$$

Above equals $\left| \frac{\operatorname{cov} (d_{n+1}, d_n)}{\operatorname{var} d_n} \right| = \left| \frac{\operatorname{cov} (f_{n+1}, f_n)}{\sigma_n^2 \frac{C_n'}{C_n} \sqrt{C_{n+1} C_n}} \right|$

$$= \left| \frac{\operatorname{cov} (f_{n+1}, f_n)}{\operatorname{var} f_n \sqrt{\frac{C_{n+1}}{C_n}}} \right| .$$

Thus, a sufficient condition that $P\{|d_{n+1}| > |d_n|\} > \frac{1}{2}$ $n = r + 1, \cdots, N$

is that the absolute value of

$$\frac{(D P_{\Omega_n} e_n, P_{\Omega_{n+1}} e_{n+1}) - \sigma_n^2 (e_n, P_{\Omega_{n+1}} e_{n+1})}{\left[\sigma_n^2 - 2\sigma_n^2 (e_n, P_{\Omega_n} e_n) + (P_{\Omega_n} e_n, D P_{\Omega_n} e_n)\right] \left[\frac{1 - \|P_{\Omega_{n+1}} e_{n+1}\|^2}{1 - \|P_{\Omega_n} e_n\|^2} \right]^{\frac{1}{2}}}$$

be ≥ 1.

This condition could then be used to insure power against alternatives in the Goldfeld, Quandt peak-test.

References

[1] Hedayat, Abdossamad (1966). Homoscedasticity in Linear Regression Analysis with Equally Spaced x's. M.S. Thesis, Cornell University, Ithaca, New York.