NORMAL REGRESSION THEORY IN THE PRESENCE OF INTRA-CLASS CORRELATION

By Max Halperin[1]

USAF School of Aviation Medicine[2]

1. Summary. In this paper we prove that certain estimators and tests of significance used in regression analysis when observations are independent are equally valid in the presence of intra-class correlation. An application of this result is presented for the situation in which several replications of the correlated set of observations are available. As a special case of this application, it is shown that the usual test of "column effects" in the analysis of variance for a two-way classification remains valid when rows are independent and columns are uniformly correlated. This latter fact is also pointed out in [3].

2. Introduction. In the usual treatment of regression theory, as in [1] (Chapters VIII and IX), it is assumed that we have a sample of n independent observations, $y_1, \cdots, y_n$, where y_α arises from a normal distribution with mean $\sum_{p=1}^{k} C_p x_{p\alpha}$, and variance σ^2. Here, the $x_{p\alpha}$ are taken to be fixed variates. On the basis of these assumptions, unbiased estimates of $C_1, C_2, \cdots, C_k$ are obtained, and two theorems are proved, one concerning the joint distribution of the estimates of the C_p and the sum of squares of deviations from regression, the other concerning tests of significance of the C_p.

Now, on the one hand, it may happen that the results given in [1] are applied when, unknown to the experimenter, the observations are actually correlated. On the other hand, it may be clear, a priori, that the observations are correlated and that estimates and tests of the C_p are required in the light of the particular kind of correlation assumed to hold. In either case an investigation of estimates and distributions is called for. We consider these questions in Section 3 for the case that $y_1, \cdots, y_n$ have a variance matrix

$$R_n = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \rho & \rho & \cdots & 1 \end{pmatrix}. \tag{2.1}$$

In Section 4 we consider an application of our result to several replications of the correlated set of observations.

3. Estimates and significance tests in normal regression theory for correlated observations. We slightly modify the regression model indicated in Section 2

[1] Now at National Heart Institute, Bethesda, Md.

[2] This paper represents the views of the author and not necessarily those of the Department of the Air Force.

by supposing that the expected values of the y_α are given by

$$Ey_\alpha = \mu + \sum_{p=1}^{k} C_p x_{p\alpha}, \qquad \alpha = 1, 2, \cdots, n. \tag{3.1}$$

The reason for this modification will be apparent later. Assuming then that the y_α have the covariance matrix, R_n, the appropriate sample likelihood of $y_1, \cdots, y_n$ is readily seen to be

$$p(y_1, \cdots, y_n) = \frac{|R_n|^{-1/2}}{(2\pi)^{n/2}} \exp - \frac{1}{2}\{y - Ey\} R_n^{-1} \{y - Ey\}', \tag{3.2}$$

where

$$\begin{aligned} y &= (y_1, \cdots, y_n), \\ Ey &= \mu(1, \cdots, 1) + \sum_{p=1}^{k} C_p(x_{p1}, \cdots, x_{pn}). \end{aligned} \tag{3.21}$$

The maximum likelihood equations for the estimation of parameters from (3.2) are of such a formidable character that an explicit solution does not appear possible. As alternative estimates for μ, $C_1, \cdots, C_k$, one can use

$$\begin{aligned} \hat{\mu} &= \bar{y} - \sum_{p=1}^{k} \hat{C}_p \bar{x}_p, \\ \hat{C}_p &= \sum_{r=1}^{k} s_{ry} S_{rp}, \qquad p = 1, 2, \cdots, k, \end{aligned} \tag{3.3}$$

where

$$s_{ry} = \sum_{\alpha=1}^{n} (x_{r\alpha} - \bar{x}_r)(y_\alpha - \bar{y}), \qquad r = 1, 2, \cdots, k, \tag{3.31}$$

and the S_{rp} are elements of the inverse of

$$S = \begin{Vmatrix} s_{11} & \cdots & s_{1k} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ s_{k1} & \cdots & s_{kk} \end{Vmatrix}, \tag{3.32}$$

where

$$s_{ij} = \sum_{\alpha=1}^{n} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j), \qquad i, j = 1, 2, \cdots, k. \tag{3.33}$$

We go on now to investigate the distribution of $\hat{C}_1, \cdots, \hat{C}_p$, when (3.2) holds. We have the following

THEOREM A. *Let $y_1, \cdots, y_n$ be a sample of one from a multivariate normal population with covariance matrix R_n and means $\mu + \sum_{p=1}^{k} C_p x_{p\alpha}$, $\alpha = 1, \cdots, n$. Let estimates of μ and the C_p be $\hat{\mu}$ and the $\hat{C}_p$ as defined in* (3.3). *Then* (*a*) *the* $(\hat{C}_p - C_p)$ *have a multivariate normal distribution with zero means and covariance matrix* $(1 - \rho)\sigma^2 S^{-1}$, *and*

(*b*) *the quantity* $\sum_{\alpha=1}^{n}(y_\alpha - \hat{\mu} - \sum_{p=1}^{k}\hat{C}_p x_{p\alpha})^2 (= V)$ *is distributed as* $(1 - \rho)\sigma^2\chi^2$ *with* $(n - k - 1)$ *degrees of freedom, and independently of the* $\hat{C}_p$.

PROOF. Conclusion (a) of the theorem follows readily from the fact that the $\hat{C}_p$ are linear functions of variables obeying a multivariate normal law and from some simple calculations to verify that the $\hat{C}_p$ are unbiased and have the indicated covariance matrix. The details are omitted.

Now let

$$L_n = \begin{pmatrix} l_{11} & \cdots & l_{1,n-1} & \frac{1}{\sqrt{n}} \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ l_{n1} & \cdots & l_{n,n-1} & \frac{1}{\sqrt{n}} \end{pmatrix}$$

be an $n \times n$ orthogonal matrix, and let

$$\begin{aligned} z &= yL_n , \\ w_p &= x_p L_n , \qquad p = 1, 2, \cdots, k \end{aligned} \tag{3.5}$$

By this transformation (3.2) becomes

$$\begin{aligned} p(z_1, \cdots, z_n) = {}& \frac{1}{\{2\pi\sigma^2(1 - \rho)\}^{n-1/2}} \exp\left[- \frac{1}{2\sigma^2(1 - \rho)} \sum_{\alpha=1}^{n-1} (z_\alpha - Ez_\alpha)^2 \right] \\ & \cdot \frac{1}{\sigma\sqrt{2\pi}\,\{1 + (n - 1)\rho\}^{1/2}} \exp\left[- \frac{(z_n - Ez_n)^2}{2\sigma^2\{1 + (n - 1)\rho\}} \right] \end{aligned} \tag{3.51}$$

while

$$\begin{aligned} s_{ij} &= \sum_{\alpha=1}^{n} w_{i\alpha} w_{j\alpha} - w_{in} w_{jn} \\ &= \sum_{\alpha=1}^{n-1} w_{i\alpha} w_{j\alpha} , \\ s_{ry} &= \sum_{\alpha=1}^{n-1} w_{r\alpha} z_\alpha = s_{rz} . \end{aligned}$$

Applying the transformation (3.5) to the $\hat{C}_p$ and V, it is easy to show that

$$\begin{aligned} \hat{C}_p &= \sum_{r=1}^{k} s_{rz} S_{rp} , \\ V &= \sum_{\alpha=1}^{n-1} \left(z_\alpha - \sum_{p=1}^{k} \hat{C}_p w_{p\alpha} \right)^2 . \end{aligned}$$

Since it can also be shown that

$$Ez_\alpha = \sum_{p=1}^{k} C_p w_{p\alpha}, \qquad \alpha = 1, 2, \cdots, n - 1,$$

it is clear that the transformation (3.5) has reduced the problem to the standard one indicated in Section 2, with $(n - 1)$ variables instead of n, and the theorem follows by the arguments given in [1].

Theorem B. *Let $y_1, \cdots, y_n$ be as specified in Theorem A. Let H_0 be the statistical hypothesis that $C_{r+1} = C_{r+1,0}, \cdots, C_k = C_{k,0}$, regardless of the values of $C_1, \cdots, C_r$. When H_0 is true, the quantities*

$$V = \sum_{\alpha=1}^{n} \left(y_\alpha - \hat{\mu} - \sum_{p=1}^{k} \hat{C}_p x_{p\alpha} \right)$$

and

$$q = \sum_{g,h=r+1}^{k} b_{gh}(\hat{C}_g - C_{g,0})(\hat{C}_h - C_{h,0})$$

are independently distributed as $(1 - \rho)\sigma^2\chi^2$ with $(n - k - 1)$ and $(k - r)$ degrees of freedom respectively. Here $\hat{C}_p$ is defined by (3.3) and the b_{gh} are defined by the matrix equation

$$\begin{pmatrix} b_{r+1,r+1} & \cdots & b_{r+1,k} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ b_{k,r+1} & \cdots & b_{kk} \end{pmatrix} = \begin{pmatrix} S_{r+1,r+1} & \cdots & S_{r+1,k} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ S_{k,r+1} & \cdots & S_{kk} \end{pmatrix}^{-1}. \tag{3.6}$$

Also

$$F = \frac{(n - k - 1)\, q}{(k - r)\, V}$$

provides a test of H_0 for $1 > \rho > - 1/n - 1$.

Proof. It is clear that application of the transformation (3.5) will reduce the problem to that of proving the corresponding theorem in standard regression theory with a sample of $(n - 1)$ independent observations. The theorem follows.

4. An application. We suppose we have m replications of the correlated sample of Section 3, generalizing slightly by further assuming that μ differs from replication to replication, assuming the value r_i for the ith replication. Thus, if $y_{i\alpha}$ is the αth measurement in the ith replication,

$$Ey_{i\alpha} = r_i + \sum_{p=1}^{k} C_p x_{p\alpha}, \qquad \begin{array}{l} i = 1, 2, \cdots, m, \\ \alpha = 1, 2, \cdots, n, \end{array} \tag{4.1}$$

and we ask for estimates of the r_i and C_p, and tests of significance for the C_p.

It follows easily from Section 3 that unbiased estimates of the r_i and the C_p are given by

$$(4.2)\qquad \begin{aligned} r_i &= \bar{y}_{i\cdot} - \sum_{p=1}^{k} \hat{C}_p \bar{x}_p, & i &= 1, 2, \cdots, m,\\ \hat{C}_p &= \sum_{r=1}^{k} s_{r\bar{y}} S_{rp}, & p &= 1, 2, \cdots, k, \end{aligned}$$

where

$$\begin{aligned} s_{r\bar{y}} &= \sum_{\alpha=1}^{n} (x_{r\alpha} - \bar{x}_r)(\bar{y}_{\cdot\alpha} - \bar{y}_{\cdot\cdot}),\\ \bar{y}_{i\cdot} &= \frac{1}{n} \sum_{\alpha=1}^{n} y_{i\alpha},\\ \bar{y}_{\cdot\alpha} &= \frac{1}{m} \sum_{i=1}^{m} y_{i\alpha},\\ \bar{y}_{\cdot\cdot} &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{\alpha=1}^{n} y_{i\alpha}, \end{aligned}$$

and S_{rp} is defined as in (3.33).

We now ask for the joint distribution of $\hat{C}_1, \cdots, \hat{C}_k$ and

$$(4.22)\qquad V = \sum_{i=1}^{m} \sum_{\alpha=1}^{n} \left\{ y_{i\alpha} - \bar{y}_{i\cdot} - \sum_{p=1}^{k} \hat{C}_p (x_{p\alpha} - \bar{x}_p) \right\}^2 .$$

It follows as in Section 3 that $\hat{C}_1, \cdots, \hat{C}_k$, have a multivariate normal distribution, and it is sufficient for our purposes to examine the joint distribution of V and W, where

$$W = m(\hat{C} - C)S(\hat{C} - C)', \qquad \hat{C} - C = (\hat{C}_1 - C_1, \cdots, \hat{C}_k - C_k).$$

By an application of the transformation $z_i = y_i L_n$ to the n observations of each replication, one obtains

THEOREM A'. *Let* $y_{i1}, \cdots, y_{in} (i = 1, 2, \cdots, m)$ *be a sample of one from a multivariate normal population with means given by* (4.1) *and the* $mn \times mn$ *variance matrix*

$$R_{nm} = \begin{vmatrix} R_n & 0 & \cdots & 0 \\ 0 & R_n & \cdots & 0 \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ 0 & 0 & \cdots & R_n \end{vmatrix}.$$

Then $(\hat{C}_1 - C_1), \cdots, (\hat{C}_k - C_k)$, *have a multivariate normal distribution with zero means and variance matrix* $[(1 - \rho)\sigma^2/m]S^{-1}$, *and* W *and* V *are independent* $(1 - \rho)\sigma^2\chi^2$ *variates with* k *and* $m(n - 1) - k$ *degrees of freedom respectively.*

We can also prove

THEOREM B′. *Let $y_{i1}, \cdots, y_{in} (i = 1, 2, \cdots, m)$ satisfy the conditions of Theorem A′. Let H_0 be as specified in Theorem B. Then the statistic*

$$F = \frac{[m(n-1) - k]\, q}{(k - r)\, V},$$

where

$$q = \sum_{g,h=r+1}^{k} b_{gh}(\hat{C}_g - C_{g,0})(\hat{C}_h - C_{h,0})$$

and b_{gh} is defined by the matrix equation

$$\begin{pmatrix} b_{r+1,r+1} & \cdots & b_{r+1,k} \\ \cdot & & \cdot \\ \cdot & & \\ \cdot & & \\ b_{k,r+1} & \cdots & b_{kk} \end{pmatrix} = m \begin{pmatrix} S_{r+1,r+1} & \cdots & S_{r+1,k} \\ \cdot & & \cdot \\ \cdot & & \\ \cdot & & \\ S_{k,r+1} & \cdots & S_{kk} \end{pmatrix}^{-1},$$

is distributed as Snedecor's F and provides a test of H_0.

The proof of Theorem B′ is along the same lines as that of Theorem A′ and is omitted.

We also remark that theorems akin to A′ and B′ hold if $r_i = r, i = 1, 2, \cdots, m$. We simply may take

$$\hat{r} = \bar{y}_{..} - \sum_{p=1}^{k} \hat{C}_p \overline{x}_p .$$

The estimates of the C_p need not be changed. If now we let

$$V' = \sum_{i=1}^{m} \sum_{j=1}^{n} \left\{ y_{ij} - \bar{y}_{..} - \sum_{p=1}^{k} \hat{C}_p (x_{pj} - \bar{x}_p) \right\}^2,$$

Theorems A′ and B′ hold with r_i and $\hat{r}_i$ replaced by r and $\hat{r}$, with V replaced by V' and $m(n - 1) - k$ degrees of freedom replaced by $nm - k - 1$ degrees of freedom.

As an example of the application of these notions we consider an analysis of variance problem. The same problem has been considered in [3]. Suppose we have mn observations,

$$\begin{matrix} y_{11} & \cdots & y_{1n} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ y_{m1} & \cdots & y_{mn} \end{matrix}$$

where the $y_{i\alpha}$ are jointly normal with covariance matrix R_{nm} and with means given by

$$Ey_{i\alpha} = r_i + C_\alpha . \tag{4.3}$$

In [3] it is shown that the F ratio for "columns" calculated in the usual way has the usual F distribution when the C_j are equal. To deduce this test from our results we write (4.3) as

$$(4.31) \qquad Ey_{i\alpha} = r_i + \sum_{p=1}^{n} C_p x_{p\alpha},$$

where

$$\begin{aligned} x_{p\alpha} &= 0, \quad p \neq \alpha, \\ &= 1, \quad p = \alpha. \end{aligned}$$

We have then

$$\begin{aligned} s_{pq} &= \sum_{\alpha=1}^{n} (x_{p\alpha} - \bar{x}_p)(x_{q\alpha} - \bar{x}_q) \\ &= -\frac{1}{n}, p \neq q, \\ &= \frac{n-1}{n}, p = q. \end{aligned}$$

The $n \times n$ matrix, S, is singular. To overcome this difficulty we can, since we are only interested in class differences rather than in the absolute values of the C_p, arbitrarily assign to one of the C_p, say C_n, the value zero. The test of column differences then becomes a test that $C_1 = C_2 = \cdots = C_{n-1} = 0$. It is then easy to see that $\hat{C}_p = \bar{y}_{\cdot p} - \bar{y}_{\cdot n}$, $p = 1, 2, \cdots, n-1$, and

$$\hat{r}_i = \bar{y}_{i\cdot} - \frac{1}{n} \sum_{p=1}^{n-1} \hat{C}_p.$$

If we substitute these values in $q = m\,\hat{C}\,S^*\,\hat{C}'$ and

$$V = \sum_{i=1}^{m} \sum_{\alpha=1}^{n} (y_{i\alpha} - \hat{r}_i - \sum_{p=1}^{n-1} \hat{C}_p x_{p\alpha})^2,$$

where

$$\hat{C} = (\hat{C}_1, \cdots, \hat{C}_{n-1})$$

and S^* is the minor of s_{nn} in S, we find after a little algebraic reduction that

$$F = \frac{(n-1)(m-1)q}{(n-1)V} = \frac{m(n-1)(m-1)\sum_{j=1}^{n} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2}{(n-1)\sum_{i=1}^{m}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2},$$

and this is the desired statistic.

Suggestions of the referee for simplifying the proofs are gratefully acknowledged.

REFERENCES

[1] S. S. Wilks, *Mathematical Statistics*, Princeton University Press, 1946.

[2] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946, pp. 490–496.

[3] D. F. Votaw, A. W. Kimball, and J. A. Rafferty, "Compound symmetry tests in the multivariate analysis of medical experiments," *Biometrics*, Vol. 6 (1950), pp. 259–281.

[4] J. E. Walsh, "Concerning the effect of intra-class correlation on certain significance tests," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 88–96.