ESTIMATION OF VARIANCES OF AVERAGES BASED ON OVERLAPPING SAMPLES IN REPEATED SURVEYS

B. Quenneville and K.P. Srinath, Statistics Canada

1. INTRODUCTION

Surveys carried out by government agencies to obtain information on economic and social trends are usually repeated either monthly or quarterly. Since the surveys are expected to provide both estimates of levels and changes, they are based on overlapping samples. The amount of overlap between two periods of time is usually specified in advance, but also depends on births and deaths in the population. Quite often there is a need to produce annual averages of the monthly estimates, especially if the results of a single sample are subject to large sampling errors.

In this paper, a method of estimating the variances of annual averages and ratios of these averages is suggested. The method is based on fitting a linear model to the logarithm of the correlation coefficient between the estimates of monthly totals which are 'd' months apart, where d = 1, 2, ... 11. In section 2 a description of the problem is provided. In section 3 a model to estimate the correlation coefficients is proposed. We present in section 4 an application to the variance of ratios and to stratified samples. The usefulness of this model for estimating the variance of the annual averages of monthly totals of the Canadian Survey of Employment, Payrolls and Hours is examined in section 5. Finally, section 6 presents the conclusion of the study.

2. THE VARIANCE OF THE ANNUAL AVERAGE

The problem of deriving the variance of the annual average is straightforward. Let $\hat{Y} = \frac{N}{n} \sum_{i \in s} y_i$ be the estimator of total Y for a a simple random sample without replacement (s) of size n.

Let $\hat{Y}_i$ (i = 1, 2, ..., 12) be the estimators of total Y for months i = 1, 2, ..., 12.

Define $\bar{\hat{Y}} = \frac{1}{12} \sum_{i=1}^{12} \hat{Y}_i$ as the annual mean of the $\hat{Y}_i$'s.

We have:

$$V(\bar{\hat{Y}}) = V\left[\frac{1}{12} \sum_{i=1}^{12} \hat{Y}_i\right]$$

$$= \frac{1}{144}\left[\sum_{i=1}^{12} V(\hat{Y}_i) + \sum\sum_{i \neq j} \mathrm{Cov}(\hat{Y}_i, \hat{Y}_j)\right]$$

and we use $v(\bar{\hat{Y}})$ to estimate $V(\bar{\hat{Y}})$, where

$$v(\bar{\hat{Y}}) = \frac{1}{144}\left[\sum_{i=1}^{12} v(\hat{Y}_i) + \sum\sum_{i \neq j} \mathrm{cov}\,(\hat{Y}_i, \hat{Y}_j)\right] \quad (1)$$

If the samples for two months i and j are selected independently, then the $\mathrm{cov}(\hat{Y}_i, \hat{Y}_j)$ terms are zero.

In a partial replacement survey, part of the sample for a given month overlaps the samples of other months. As a result, the covariance terms are non-zero, possibly even quite large.

There are 66 covariance terms in $v(\bar{\hat{Y}})$. To calculate these 66 covariances, we require information on the units common to months i and j for each $\mathrm{cov}(\hat{Y}_i, \hat{Y}_j)$ term, which would require processing 12 months data, two months at a time - a formidable computing burden.

A method of approximating $v(\bar{\hat{Y}})$ was developed in order to: i) avoid the problem of identifying the common units for every pair of months, ii) use monthly information that is already available, and iii) comply with the constraints of existing computer software.

In (1), $v(\hat{Y}_i)$ are available monthly. Only the covariance terms have to be computed. Now,

$$\mathrm{cov}(\hat{Y}_i, \hat{Y}_j) = \hat{\rho}_{\hat{Y}_i, \hat{Y}_j} \cdot \sqrt{v(\hat{Y}_i)} \cdot \sqrt{v(\hat{Y}_j)} \,. \quad (2)$$

Again, $v(\hat{Y}_i)$ and $v(\hat{Y}_j)$ are available on a monthly basis. Therefore only the estimate of the coefficient of correlation between $\hat{Y}_i$ and $\hat{Y}_j$ has to be computed.

There are 66 correlation coefficients $\hat{\rho}_{\hat{Y}_i, \hat{Y}_j}$ to calculate. As already noted above, computation of the correlation coefficients involves the identification of common units between every pair of months. In some surveys it may be computationally less expensive to identify common units between months which are 1 or 2 months apart than to identify common units which are 'd' months apart where 'd' is greater than 2, since the amount of processing is greatly reduced.

In the proposed technique only correlation coefficients between months which are 1 or 2 months apart are actually computed whereas the remaining correlation coefficients are estimated by means of a model. For example, January and February are 1 month apart, March and May are 2 months apart.

In the case where the distance between months i and j is either 1 or 2, we have:

$$\hat{\rho}_{\hat{Y}_i, \hat{Y}_j} = \frac{n_{c_{ij}}}{\sqrt{n_i}\sqrt{n_j}} * \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \quad (3)$$

where

$n_{c_{ij}}$: the number of sample units that months i and j have in common

n_i : the number of units in the sample for month i

n_j : the number of units in the sample for month j

$$s_{ij} = \sum_{k=1}^{n_{c_{ij}}} y_{ik}\, y_{jk} - \frac{\left(\sum_{k=1}^{n_{c_{ij}}} y_{ik}\right)\left(\sum_{k=1}^{n_{c_{ij}}} y_{jk}\right)}{n_{c_{ij}}}$$

$$s_{ii} = \sum_{k=1}^{n_{c_{ij}}} y_{ik}^2 - \frac{\left(\sum_{k=1}^{n_{c_{ij}}} y_{ik}\right)^2}{n_{c_{ij}}}$$

$$s_{jj} = \sum_{k=1}^{n_{c_{ij}}} y_{jk}^2 - \frac{\left(\sum_{k=1}^{n_{c_{ij}}} y_{jk}\right)^2}{n_{c_{ij}}}$$

$\sum_{k=1}^{n_{c_{ij}}}$: is the summation over the units that months i and j have in common

y_{ik} : value of the characteristic for the k-th unit in month i

y_{jk} : value of the characteristic for the same k-th unit (y_{ik}) in month j.

Equation (3) is a modification of the formula given in Kish, Survey Sampling, page 458.

3. THE MODEL

The model used to estimate the remaining correlation coefficients, for which $j-i \geq 3$ and $j > i$ is

$$E\left[\rho_{\hat{Y}_i, \hat{Y}_j} \;\middle|\; j-i\right] = e^{-\theta(j-i)} \qquad (4)$$

where θ is estimated by $\hat{\theta}$ using the 21 coefficients calculated in (3), where

$$\theta = \frac{-1}{51}\left[\sum_{i=1}^{11} \log\left(\hat{\rho}_{\hat{Y}_i, \hat{Y}_{i+1}}\right) + \sum_{i=1}^{10} \log\left(\hat{\rho}_{\hat{Y}_i, \hat{Y}_{i+2}}\right)\right] . \qquad (5)$$

Note that $\hat{\theta}$ is the least-squares estimator of θ for the model described above.

4. APPLICATION TO THE VARIANCE OF RATIOS AND TO STRATIFIED SAMPLES

4.1 Ratio of Averages

Let $\bar{\hat{X}} = \frac{1}{12}\sum_{i=1}^{12} \hat{X}_i$ and $\bar{\hat{Y}} = \frac{1}{12}\sum_{i=1}^{12} \hat{Y}_i$.

Define $\hat{R} = \frac{\bar{\hat{Y}}}{\bar{\hat{X}}}$, for example $\bar{\hat{X}}$ could be the average monthly total employment and $\bar{\hat{Y}}$ be the average monthly total earnings.

Again, the problem is to estimate $V(\hat{R})$ using $v(\hat{R})$.

We have

$$V(\hat{R}) = \frac{1}{\bar{\hat{X}}^2}\left\{V(\bar{\hat{Y}}) + R^2\, V(\bar{\hat{X}}) - 2R\, \mathrm{Cov}(\bar{\hat{X}}, \bar{\hat{Y}})\right\}$$

and

$$v(\hat{R}) = \frac{1}{\bar{\hat{X}}^2}\left\{v(\bar{\hat{Y}}) + \hat{R}^2\, v(\bar{\hat{X}}) - 2\,\hat{R}\, \mathrm{cov}(\bar{\hat{X}}, \bar{\hat{Y}})\right\}$$

Now, $v(\bar{\hat{Y}})$ and $v(\bar{\hat{X}})$ are computed as indicated in section 2. The only new term is $\mathrm{cov}(\bar{\hat{X}}, \bar{\hat{Y}})$.

However, we know that

$$\mathrm{cov}(\bar{\hat{X}}, \bar{\hat{Y}}) = \frac{1}{144}\,\mathrm{cov}\left[\sum_{i=1}^{12} \hat{X}_i, \sum_{i=1}^{12} \hat{Y}_i\right]$$

$$= \frac{1}{144}\left[\sum_{i=1}^{12} \mathrm{cov}(\hat{X}_i, \hat{Y}_i) + \sum_{i\neq j}\sum \mathrm{cov}(\hat{X}_i, \hat{Y}_j)\right] .$$

The $\mathrm{cov}(\hat{X}_i, \hat{Y}_i)$ terms are calculated monthly.

The $\mathrm{cov}(\hat{X}_i, \hat{Y}_j)$ terms can be rewritten as

$$\mathrm{cov}(\hat{X}_i, \hat{Y}_j) = \hat{\rho}_{\hat{X}_i, \hat{Y}_j}\sqrt{v(\hat{X}_i)}\sqrt{v(\hat{Y}_j)} .$$

Again, $v(X_i)$ and $v(Y_j)$ are available on a monthly basis.

We compute $\hat{\rho}_{\hat{X}_i, \hat{Y}_j}$ by the same method as described in section 3, substituting X_i for Y_i Thus, characteristic X for month i is analyzed instead of characteristic Y.

4.2 Stratified Samples

Let $\hat{Y}_{i\ell}$ be the estimator of the total of characteristic Y for month i = 1, ..., 12 in stratum ℓ = 1, ..., L.

Define $\hat{Y}_i = \sum_{\ell=1}^{L} Y_{i\ell}$

and

$$\bar{\hat{Y}} = \frac{1}{12}\sum_{i=1}^{12} \hat{Y}_i = \frac{1}{12}\sum_{i=1}^{12}\sum_{\ell=1}^{L} \hat{Y}_{i\ell}$$

$$= \frac{1}{12}\sum_{\ell=1}^{L}\sum_{i=1}^{12} \hat{Y}_{i\ell} = \sum_{\ell=1}^{L} \bar{\hat{Y}}_\ell$$

There are two methods of estimating $v(\bar{\hat{Y}})$. The first is to take

$$\bar{\hat{Y}} = \sum_{\ell=1}^{L} \bar{\hat{Y}}_\ell \quad \text{and} \quad v(\bar{\hat{Y}}) = \sum_{\ell=1}^{L} v(\bar{\hat{Y}}_\ell)$$

where $v(\bar{\hat{Y}}_\ell)$ is obtained as in part 3.

The second technique is to model ρ at the strata aggregation level rather than at the stratum level, as $v(\bar{\hat{Y}}_\ell)$ does.

The procedure for modelling ρ at the strata aggregation level is as follows:

1) Compute the coefficient of correlation between the estimators for months i and j for j-i = 1, 2 at the stratum level.
2) Compute the coefficient of correlation between the estimators for months i and j for j-i = 1, 2 at aggregation level A, using

$$\hat{\rho}_{\hat{Y}_{i_A}\hat{Y}_{j_A}} = \frac{\sum_{\ell\in A} \hat{\rho}_{\hat{Y}_{i\ell}\hat{Y}_{j\ell}}\sqrt{v(\hat{Y}_{i\ell})}\sqrt{v(\hat{Y}_{j\ell})}}{\sqrt{v(\hat{Y}_{iA})}\sqrt{v(\hat{Y}_{jA})}}$$

where $v(\hat{Y}_{i\ell})$, $v(\hat{Y}_{j\ell})$ $v(\hat{Y}_{iA})$ $v(\hat{Y}_{jA})$ are obtained monthly and $\hat{\rho}_{\hat{Y}_{i\ell}, \hat{Y}_{j\ell}}$ is given

by 1) above.

3) Estimate θ by $\hat{\theta}$ using (5), replacing

$$\hat{\rho}_{\hat{Y}_i,\ \hat{Y}_{i+1}} \quad \text{with} \quad \hat{\rho}_{\hat{Y}_{iA},\ \hat{Y}_{(i+1)A}}$$

and $\hat{\rho}_{\hat{Y}_i,\ \hat{Y}_{i+2}}$ with $\hat{\rho}_{\hat{Y}_{iA},\ \hat{Y}_{(i+2)A}}$

4) Estimate the coefficient of correlation between the estimators for months i and j for j-i > 2 using model (4).

5. EVALUATION OF THE MODEL

To evaluate the model

$$E\left[\rho_{\hat{Y}_i,\ \hat{Y}_j} \;\middle|\; j-1\right] = e^{-\theta(j-i)} \qquad (4)$$

data from the Canadian Survey of Employment, Payrolls and Hours were used.

The test involved computing all the 66 $\hat{\rho}_{\hat{Y}_i,\ \hat{Y}_j}$ terms and checking whether exponential model in j-i is a good fit.

Taking the logarithm on both sides of the equation $\hat{\rho}_{\hat{Y}_i,\ \hat{Y}_j} = e^{-\theta(j-i)} + \varepsilon$, we obtain

$$\log\left(\hat{\rho}_{\hat{Y}_i,\ \hat{Y}_j}\right) = -\theta(j-i) + \varepsilon \ .$$

This equation was fitted to the data on estimates of employment from May 82 through April 1983. 14 cases are evaluated. However graphical results are given for only 3 cases (printed at the end of the paper).

The left-hand graphs, LINEAR REGRESSION, show the actual data (X) and the expected values (.). The expected values are connected by a straight line with slope $-\theta$. The X-axis is the distance between months and the Y-axis is the logarithm of the coefficient of correlation between the estimators of total employment.

The right-hand graphs show the standardized residuals by the distance between months.

Examination of the graphs leads to the following observations:

1. In case 5 as in five other cases (2, 6, 12, 13, 14) the standardized residuals appear to be randomly distributed between -2 and 2. There are few residuals outside this range. Hence, the model appears correct and it may be assumed that the errors are normally distributed.
2. In case 3, as in case 1, the residuals appear to be a linear function of the distance. This is evident in the diagonal line they form.
3. In case 8, as in five other cases (4, 7, 9, 10, 11), the residuals appear to be a quadratic or cubic function of the distance. This is indicated by their curved distribution.

Precise values of R^2 were computed; they are all close to 1. They are listed by case under R_1^2 in TABLE 1.

TABLE 1

CASE	R_1^2	R_2^2
1	.9553	.8109
2	.9243	.9258
3	.7588	.4463
4	.9481	.9472
5	.9858	.9810
6	.9677	.9614
7	.9768	.9777
8	.9594	.9510
9	.9309	.9202
10	.9266	.8620
11	.9795	.8165
12	.9691	.9716
13	.9706	.9120
14	.9810	.9733

These R^2 values indicate that the model is reasonably good and that a linear, quadratic or cubic term for the distance would improve the fit only marginally.

Once we accept this model, the next step is to validate the proposed variance estimator, i.e. if we estimate θ by the least square estimator based only on the correlation coefficients for which the distance between the months is either 1 or 2, how do we predict the remaining correlation coefficients?

To answer this question, we first estimate θ based on the correlation coefficients between monthly totals which are 1 or 2 months apart. Second, we predict the correlation coefficients between estimators of monthly totals which are 'd' months apart, where 'd' = 3, ..., 11 by using the model (4), i.e.

$$\hat{\rho}_{\hat{Y}_i,\ \hat{Y}_j} = e^{-\theta(j-i)}$$

The graphical results of these 2 steps are again given for the 3 previous cases (printed at the end of the paper).

The left-hand graphs, PLOT OF THE MODEL, show the actual data (x) and the predicted data (.). The predicted data are connected by the function $e^{-\theta(j-i)}$. The X-axis is the distance between months and the Y-axis is the correlation coefficient between the estimators of total employment.

The right-hand graphs show the residuals by the distance between months. It can be seen that the residuals are zero when the distance between months is either 1 or 2. This is because we actually used the computed values in those cases.

Next, we compute R^2 again but based on only the observations for which 'd' = 3, 4, ..., 11. That is, we try to measure how well our model explains the actual values.

Listed by case, the values are under R_2^2 in TABLE 1.

Except for case 3, those values are all greater than .80, which is quite good.

Examination of the graphs leads to the following observations.

In case 5, as in 5 other cases, there is no obvious patterns in the residuals and we note that the R^2 values are all close to 1.

In case 8, as in 6 other cases, a decreasing pattern can be detected as the distance increases. But again the R^2 values are quite good and it would not be worthwhile to complicate the model to improve those R^2 values marginally.

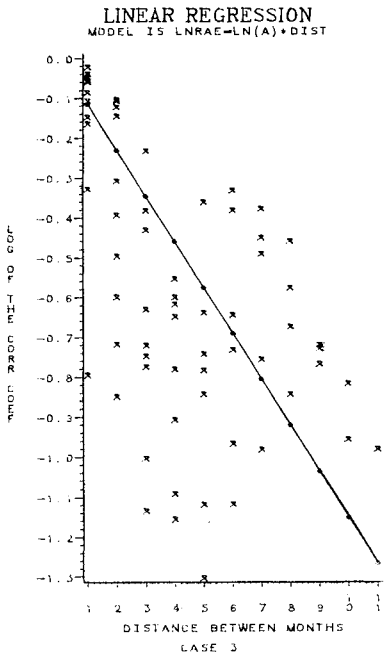
Only in the case 3 the model does not predict the values very well. It tends to underestimate the correlation coefficients.

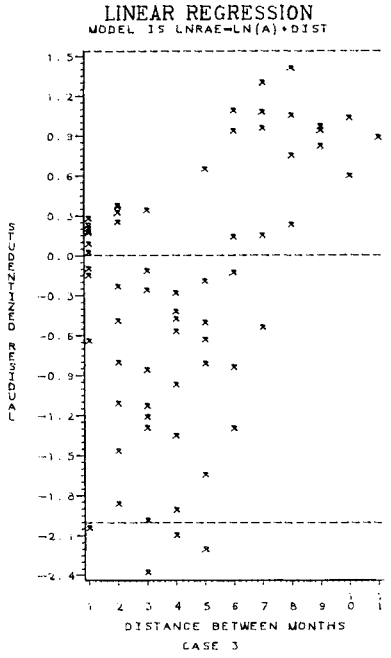
6. CONCLUSIONS

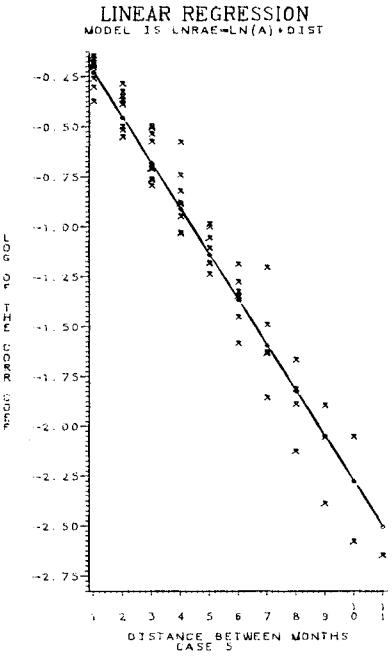
From the observations made in section 5 we can reasonably conclude that the assumed model is a good fit and predicts in most of the cases fairly accurately the correlation coefficients.

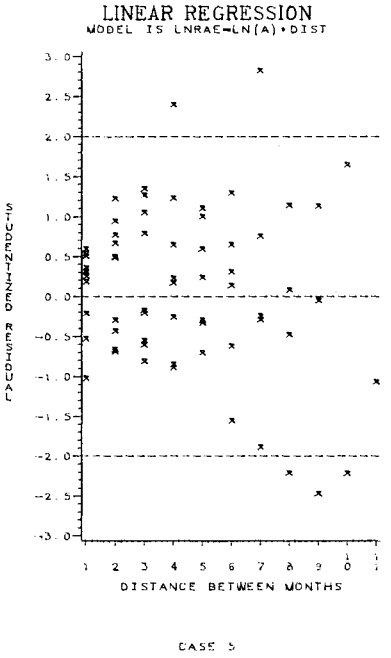
It seems worthwhile to estimate the correlation coefficients using the model as this involves computing only correlation coefficients between monthly totals which are 1 or 2 months apart without approximating the actual variance estimate to a great extent.

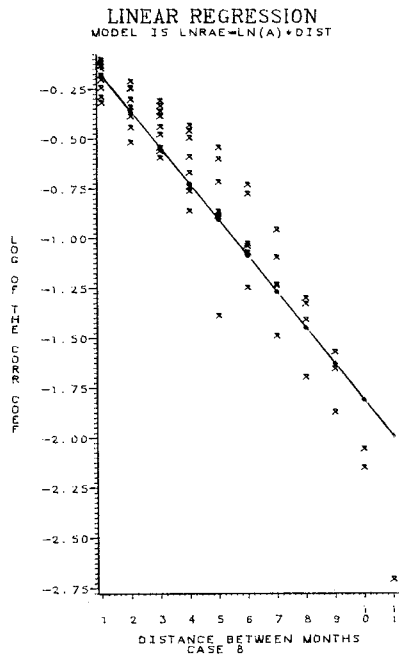
The saving in computation more than offsets the loss in precision of the variance estimates.

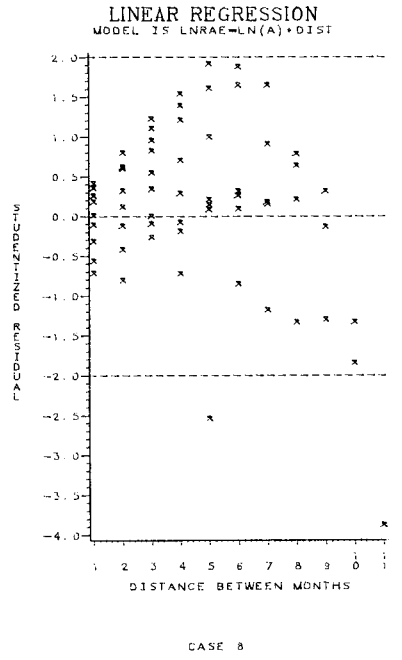


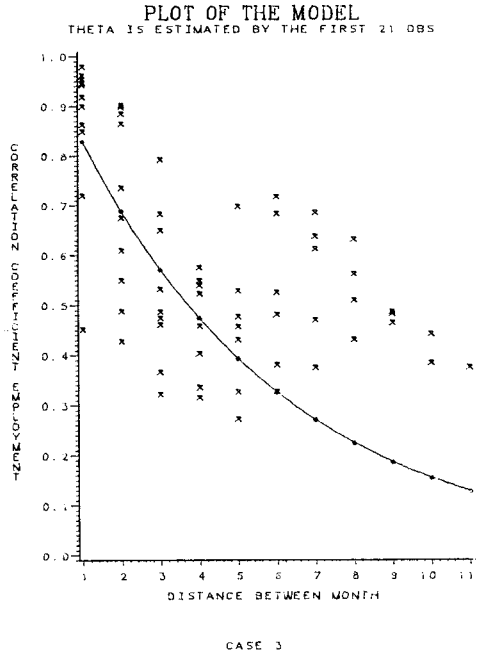


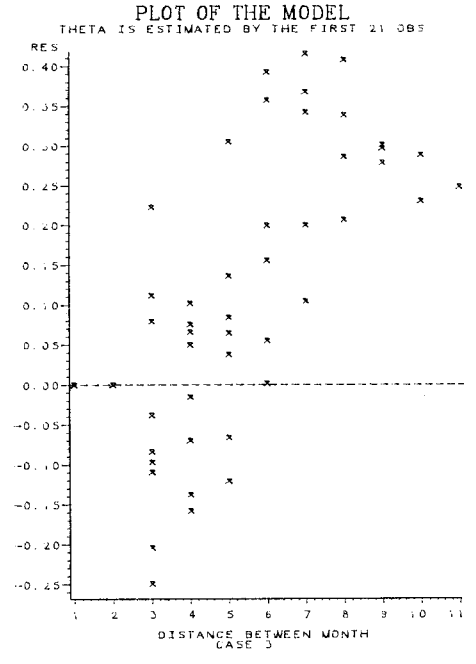


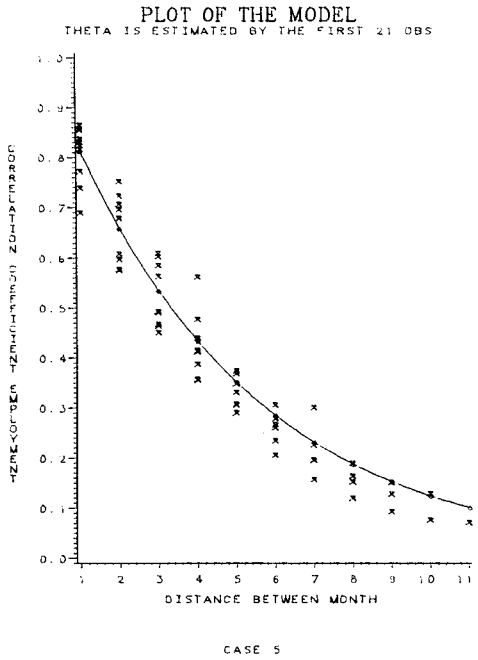


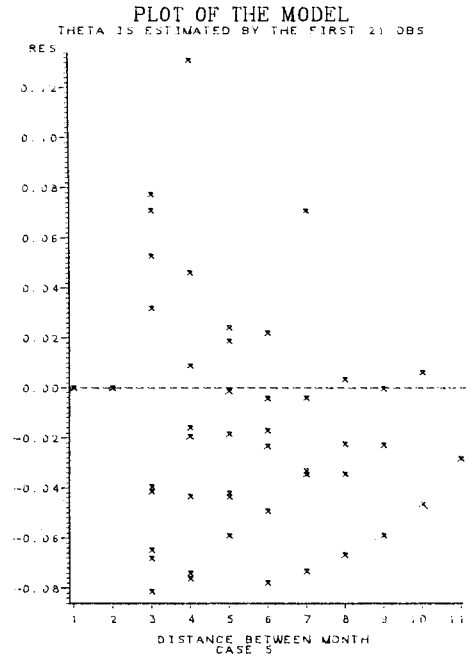

LINEAR REGRESSION
MODEL IS LNRAE=LN(A)+DIST
LOG OF THE GROUP MEAN
DISTANCE BETWEEN MONTHS
CASE 8

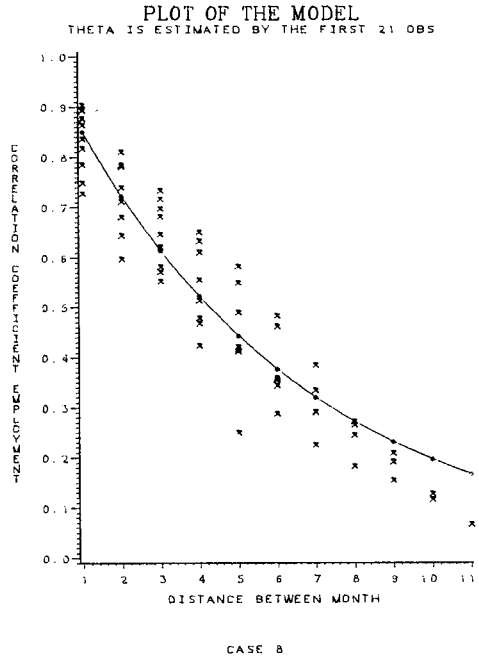

LINEAR REGRESSION
MODEL IS LNRAE=LN(A)+DIST
STUDENTIZED RESIDUAL
DISTANCE BETWEEN MONTHS
CASE 8

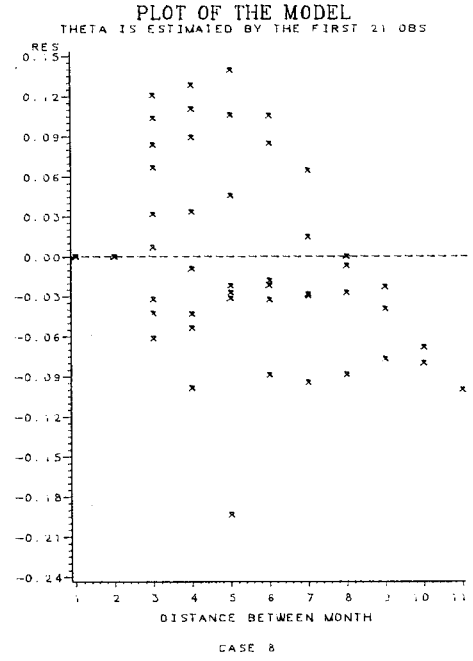

PLOT OF THE MODEL
THETA IS ESTIMATED BY THE FIRST 21 OBS
DISTANCE BETWEEN MONTH
CASE 3


PLOT OF THE MODEL
THETA IS ESTIMATED BY THE FIRST 21 OBS
RES
DISTANCE BETWEEN MONTH
CASE 3


PLOT OF THE MODEL
THETA IS ESTIMATED BY THE FIRST 21 OBS
DISTANCE BETWEEN MONTH
CASE 5


PLOT OF THE MODEL
THETA IS ESTIMATED BY THE FIRST 21 OBS
RES
DISTANCE BETWEEN MONTH
CASE 5


PLOT OF THE MODEL
THETA IS ESTIMATED BY THE FIRST 21 OBS
DISTANCE BETWEEN MONTH
CASE 8


PLOT OF THE MODEL
THETA IS ESTIMATED BY THE FIRST 21 OBS
RES
DISTANCE BETWEEN MONTH
CASE 8