# On Systematic Errors in the Least Squares Regression Analysis, with Application to the Atmospheric Effects on the Cosmic Radiation

By HARALD TREFALL and JACK NORDÖ[1], Universitetet i Bergen, Bergen, Norway

(Manuscript received January 2, 1959)

## Abstract

It is shown that if the variables employed in a least squares regression analysis are subject to random errors of measurement, the expectation values of the partial regression coefficients, of the partial correlation coefficients and of the multiple correlation coefficient may all differ from those which would have existed, had no errors been present. If there is no intercorrelation between the errors of different variables, random errors in a given variable always reduce the numerical expectation values of the corresponding partial regression and correlation coefficients. Coefficients corresponding to other variables may, however, be influenced in either direction depending on the intercorrelations between the variables. The expectation value of the multiple correlation coefficient is reduced by errors in any variable. The general case, in which the errors of different variables are intercorrelated, has also been briefly discussed.

The problem of determining the atmospheric effects on the cosmic radiation is then discussed. It is shown that some previously unexplained discrepancies between empirical and theoretical estimates, and also between empirical estimates obtained from the study of day-to-day variations, and such obtained from the seasonal variations of the cosmic-ray intensity, are probably due to systematic effects of random errors in the aerological data employed in the regression analysis.

Estimates of error variances and covariances of aerological data from the upper troposphere and the lower stratosphere have been obtained by analysing differences between data from two closely situated stations on Spitzbergen. They have then been used to obtain corrected estimates of the cosmic-ray atmospheric effects, which are now found to agree fairly well with the theoretical ones.

## I. General considerations

### 1. *Introduction*

A least squares regression analysis is frequently used when a possible linear relationship between two or more measured quantities is investigated. The simplicity of the method is, perhaps, the main reason why it is so commonly used. It does not appear, however, to be generally appreciated that random errors in the independent variables will not only increase the variances of the correlation and regression coefficients, but will also change their expectation values.

This effect is very easily seen in the case of only one independent variable. The relation between the measured values $x_1$ of the independent variable and $x_0$ of the dependent variable is then given by

$$x_0 = b_1 x_1 + \xi \tag{1}$$

where $b_1$ is the regression coefficient and $\xi$ is the residual in the regression equation between the measured quantities. If the measurements are subject to random errors we may write

$$x_i = x_i' + \varepsilon_i \tag{2}$$

[1] Now at Det Norske Meteorologiske Institutt, Oslo.

where $x_i'$ is the true value of $x_i$ and $\varepsilon_i$ is the random error of measurement, the corresponding regression equation between the true values being

$$x_0' = b_1' x_1' + \xi'. \tag{3}$$

For the sake of simplicity, but without loss of generality we may assume that all variables have zero mean.

We shall assume that the errors are always random with respect to the true variables, i.e.

$$\overline{\varepsilon_i x_j'} = 0 \tag{4}$$

for all $i$ and $j$, which implies that

$$\overline{x_i x_j} = \overline{x_i' x_j'} + \overline{\varepsilon_i \varepsilon_j}. \tag{5}$$

A bar denotes the mean value of an infinitely large sample.

The expectation values of the regression coefficients $b_1$ and $b_1'$ are then

$$b_1' = \frac{\overline{x_0' x_1'}}{\overline{x_1'^2}} \tag{6}$$

and

$$b_1 = \frac{\overline{x_0' x_1'} + \overline{\varepsilon_0 \varepsilon_1}}{\overline{x_1'^2} + \overline{\varepsilon_1^2}}. \tag{7}$$

For the corresponding correlation coefficients they are

$$R' = \frac{\overline{x_0' x_1'}}{\left(\overline{x_0'^2} \cdot \overline{x_1'^2}\right)^{\frac{1}{2}}} \tag{8}$$

and

$$R = \frac{\overline{x_0' x_1'} + \overline{\varepsilon_0 \varepsilon_1}}{\left[\left(\overline{x_0'^2} + \overline{\varepsilon_0^2}\right)\left(\overline{x_1'^2} + \overline{\varepsilon_1^2}\right)\right]^{\frac{1}{2}}}. \tag{9}$$

From these equations it follows that the expectation values of the regression and correlation coefficients are always affected by random errors in the independent variable $x_1$. Regarding errors in the dependent variable, the correlation coefficient $R$ is, of course, always influenced, whereas the regression coefficient $b_1'$ is affected only if this error is correlated with the error in the independent variable. We see that $b_1$ may be smaller or larger than $b_1'$, depending on the correlation which exists between $\varepsilon_0$ and $\varepsilon_1$. Even for the correlation coefficient it may happen that $R > R'$, which is rather surprising as one intuitively expects that the presence of errors would invariably reduce the magnitude of this statistical parameter. The case of $R > R'$ can, however, occur only in the rather unusual situation when the correlation between the errors is stronger than between the true variables themselves.

If the errors are not correlated with each other, i.e.

$$\overline{\varepsilon_i \varepsilon_j} = 0 \text{ for all } i \neq j, \tag{10}$$

it is easily seen that the expectation value of $R$ is always less than that of $R'$, and that the numerical expectation value of $b_1$ is always less than that of $b_1'$. This effect of random errors of measurement on the correlation coefficient was first noted by SPEARMAN (1904).

## 2. *Several variables with uncorrelated errors*

We shall now consider regression equations with several independent variables. As these variables may well be correlated with each other, they are often only formally independent. If the errors of the different variables are also correlated with each other, the situation becomes so complicated that very few conclusions can be drawn regarding the effect of the errors on the regression and correlation coefficients. However, in many (and perhaps in the majority) of the cases of practical interest the errors are not correlated with each other. We shall, therefore, begin with a discussion of this restricted case. On the other hand, the regression between cosmic-ray intensity and atmospheric conditions, which is to be discussed in the second part of this paper, presents a case in which the errors *are* intercorrelated.

Let $x_0$ be the dependent variable, and $x_i$, $i = 1, 2, \dots n$, be the "independent" ones. Let us denote by $A$ the determinant with elements

$$a_{ij} = \overline{x_i x_j}, \quad i = 0, 1, \dots . n, \text{ and}$$
$$j = 0, 1, \dots . n, \tag{11}$$

and by $A_{ij}$ the corresponding cofactors. The empirical multiple regression coefficients are then given by

$$b_i = -A_{0i}/A_{00}, \tag{12}$$

the corresponding partial correlation coefficients by

$$r_i^2 = A_{0i}^2/(A_{00}A_{ii}), \tag{13}$$

and the multiple correlation coefficient by

$$R^2 = 1 - A/(a_{00}A_{00}). \tag{14}$$

The corresponding true values $b'_i$, $r'_i$ and $R'$ are obtained by substituting for $a_{ij}$ the true covariances

$$a'_{ij} = \overline{x'_i x'_j}, \tag{15}$$

based on the true values of the variables.

For the variances of errors we shall find it convenient to introduce the notation $\overline{\varepsilon_i^2} = \varepsilon_{ii}$. In order to see how the errors influence the expectation values of regression and correlation coefficients, we first note that in the restricted case under consideration we have, according to (5) and (10), that

$$\partial a_{kk}/\partial\varepsilon_{kk} = 1, \tag{16}$$

whereas

$$\partial a_{ij}/\partial\varepsilon_{kk} = 0 \text{ if either } i \neq k \text{ and/or } j \neq k. \tag{17}$$

Eq. (16) holds, however, also in the most general case, but not (17). By means of (16) and (17) we now find that

$$\partial A/\partial\varepsilon_{kk} = A_{kk} \tag{18}$$

and

$$\partial A_{ij}/\partial\varepsilon_{kk} = (A_{ij})_{kk}, \tag{19}$$

where $(A_{ij})_{kk}$ denotes the cofactor of the element $a_{kk}$ as it appears in the determinant $A_{ij}$. If $a_{kk}$ is not an element of $A_{ij}$, $(A_{ij})_{kk}$ is identically zero, and it should be noted that this occurs whenever $i$ or $j$ equals $k$.

If we differentiate (12) we find, by means of (18) and (19), that

$$\frac{\partial b_i}{\partial\varepsilon_{kk}} = -b_i\left[\frac{(A_{00})_{kk}}{A_{00}} - \frac{(A_{0i})_{kk}}{A_{0i}}\right]. \tag{20}$$

If $k=0$, we have $(A_{00})_{kk} = (A_{0i})_{kk} = 0$, and thus $\partial b_i/\partial\varepsilon_{00} = 0$. Consequently, *the expectation values of the regression coefficients are not influenced by errors in the dependent variable.* If $k=i$ we still have $(A_{0i})_{kk}=0$, but as $i$ cannot here be zero, $(A_{00})_{kk} \neq 0$, and we get

$$\frac{1}{b_i}\frac{\partial b_i}{\partial\varepsilon_{ii}} = -\frac{(A_{00})_{ii}}{A_{00}}. \tag{21}$$

As $(A_{00})_{ii}$ and $A_{00}$ are both positive definite we conclude that *random errors in an independent variable always tend to reduce the numerical value of the corresponding regression coefficient.* However, as nothing definite can be said about the sign of $\partial b_i/\partial\varepsilon_{kk}$ when $0 \neq k \neq i$, the regression coefficients corresponding to other independent variables may be influenced in either direction depending not only on the intercorrelation between the "independent" variables, but also on their correlations with the dependent variable. Consequently, *if several variables contain random errors, no general rule can be formulated regarding the combined effect of these errors on a given regression coefficient.*

Differentiating (13), we obtain

$$\frac{\partial r_i}{\partial\varepsilon_{kk}} = -\frac{r_i}{2}\left[\frac{(A_{00})_{kk}}{A_{00}} + \frac{(A_{ii})_{kk}}{A_{ii}} - 2\frac{(A_{0i})_{kk}}{A_{0i}}\right] \tag{22}$$

If $k=0$ the first and the third quotients within the brackets vanish, but as $i$ cannot be zero the second term remains, and we find that $(1/r_i)\cdot$ $\cdot(\partial r_i/\partial\varepsilon_{00})$ is always negative. Consequently, *errors in the dependent variable reduce the numerical expectation values of all partial correlation coefficients.* If $k=i$ only the first term remains, and we find that also $(1/r_i)$ $(\partial r_i/\partial\varepsilon_{ii})$ is negative, whereas if $0 \neq k \neq i$ the sign depends on the correlations between all variables. We thus see that *random errors in an independent variable also tend to reduce the numerical value of the corresponding partial coefficient, whereas the expectation values of the other partial correlation coefficients may be changed in either direction.*

Finally, differentiation of (14) gives for $k \neq 0$

$$2R\frac{\partial R}{\partial\varepsilon_{kk}} = -\frac{(A_{00})_{kk}}{A_{00}}\left[\frac{A_{kk}}{a_{00}(A_{00})_{kk}} - \frac{A}{a_{00}A_{00}}\right] =$$
$$= -[(A_{00})_{kk}/A_{00}]\,[(1-R_k^2)-(1-R^2)], \tag{23}$$

where $R_k$ denotes the multiple correlation coefficient which would have been obtained without the variable $x_k$. As the correlation cannot improve if one of the independent variables is discarded, we conclude that $\partial R/\partial\varepsilon_{kk} < 0$ for all $k \neq 0$. (It might have been zero if $x_0$ had not been correlated with $x_k$, but then there would have been no need for $x_k$ as an independent variable.) Consequently, *random errors in any independent variable reduce the expectation value of the multiple correlation coefficient.* It is obvious that the same applies to errors in the dependent variable when these are not correlated with the errors in the independent variables.

If the independent variables are not intercorrelated, the situation is much simpler. We then have $a_{ij}=0$ whenever $ij(i-j)\neq 0$, and find that (12), (13) and (14) reduce to

$$b_i = a_{0i}/a_{ii}, \tag{24}$$

$$r_i^2 = \frac{a_{0i}^2}{a_{00}a_{ii}\left[1 - \sum\limits_{j\neq 0,1} a_{0j}^2/(a_{00}a_{jj})\right]}, \tag{25}$$

and

$$R^2 = \sum_{i\neq 0} a_{0i}^2/(a_{00}a_{ii}). \tag{26}$$

We thus see that $b_i$ can only be influenced by $\varepsilon_{ii}$, whereas $r_i$ and $R$ are still influenced by random errors in any variable.

### 3. *Several variables with correlated errors*

In the more general case when the errors in different variables are correlated and (10) is no longer satisfied, (18) and (19) do not hold and the rules formulated on the basis of the equations (20), (22) and (23) no longer apply. The following can, however, be said:

A. If, but only if, the errors in the dependent variable are correlated with the errors in one or more of the independent variables, the expectation values of all regression coefficients will be influenced also by the errors in the dependent variable, but not necessarily in such a way as to reduce their numerical values.

B. The numerical expectation value of a partial regression or correlation coefficient is no longer necessarily reduced by errors in the corresponding independent variable, nor will errors in the dependent variable always reduce the numerical expectation values of all partial correlation coefficients.

C. Even the multiple correlation coefficient may have its expectation value increased.

It is not difficult to convince oneself about the truth of the statements (A) and (B), but (C) is at first sight rather surprising. It can, however, be shown that the residual variance has still an absolute minimum when random errors of measurement are absent.

The residual $\xi$ of the empirical regression equation

$$x_0 = \sum_{i\neq 0} b_i x_i + \xi \tag{27}$$

can be written in the form

$$\xi = \xi' + (\varepsilon_0 - \sum_{i\neq 0} b_i\varepsilon_i) + \sum_{i\neq 0}(b_i' - b_i)x_i', \tag{28}$$

where $\xi'$ is the residual of the regression equation

$$x_0' = \sum_{i\neq 0} b_i' x_i' + \xi' \tag{29}$$

between the true variables. As $\overline{\xi' x_i'} = 0$ for all $i \neq 0$ and, according to (4), $\overline{\xi'\varepsilon_i} = 0$ for all $i$, we find that

$$\overline{\xi^2} = \overline{\xi'^2} + \overline{[\varepsilon_0 - \sum_{i\neq 0} b_i\varepsilon_i]^2} + \overline{[\sum_{i\neq 0}(b_i' - b_i)x_i']^2}, \tag{30}$$

which clearly demonstrates that in all cases the residual variance has an absolute minimum at zero errors.

If we express the residual variances $\overline{\xi^2}$ and $\overline{\xi'^2}$ by means of the respective multiple correlation coefficients and total variances of the dependent variables, and use the relation $\overline{x_0'^2} = \overline{x_0^2} - \overline{\varepsilon_0^2}$, we get from (30) that

$$(R'^2 - R^2)\overline{x_0^2} = -(1 - R'^2)\overline{\varepsilon_0^2} + \overline{[\varepsilon_0 - \sum_{i\neq 0} b_i\varepsilon_i]^2} + \\ + \overline{[\sum_{i\neq 0}(b_i' - b_i)x_i']^2}. \tag{31}$$

If $\overline{\varepsilon_0\varepsilon_i} = 0$ for all $i \neq 0$, (31) reduces to

$$(R'^2 - R^2)\overline{x_0^2} = R'^2\overline{\varepsilon_0^2} + \overline{[\sum_{i\neq 0} b_i\varepsilon_i]^2} + \\ + \overline{[\sum_{i\neq 0}(b_i' - b_i)x_i']^2}. \tag{32}$$

From the equations (31) and (32) we can now draw the following conclusions:

A′. If the errors in the dependent variable are either absent or uncorrelated with the errors in the independent variables, the multiple correlation coefficient has an absolute maximum at zero errors. This does not, however, imply that all partial derivatives of $R$ with respect to the error variances will be negative or zero everywhere in error space.

B′. If the errors in the dependent variable are correlated with the errors in the independent variables, $R$ may in special cases be greater than $R'$, provided that the correlation between the errors is stronger than that between the true variables. It would, for example, happen if the partial regression coefficients between the errors are approximately equal to those between the corresponding true variables, as the third term of (31) is then close to zero and the second term approximately equal to $(1 - R_\varepsilon^2)\overline{\varepsilon_0^2}$, where $R_\varepsilon$ denotes the multiple correlation coefficient between $\varepsilon_0$ as dependent variable and the other errors as independent variables. It would also happen if all error variances are large compared to the variances of the corresponding true variables, as the empirical regression coefficients will also then be approximately equal to those between the errors, again making the first and second terms of (31) dominant and their sum approximately equal to $(R_\varepsilon^2 - R'^2)\,\overline{\varepsilon_0^2}$.

Even though peculiar cases such as those discussed in (B′) may be very rare in practical applications, they serve to demonstrate how careful one should be when judging the quality of empirical relationships established between variables with unknown errors.

The important consequence of these considerations is that when some relationship is expected to exist between two or more physical quantities for theoretical reasons, empirically determined regression coefficients are never strictly comparable to the theoretically estimated ones. The expected discrepancies may, of course, be small or even negligible if the measurements are very accurate, but it should never be forgotten that significant and otherwise unexplainable discrepancies may arise in this way. In the second part of this paper we shall present what we believe to be such a case.

It should especially be noted that if some of the "independent" variables are very strongly correlated with each other, the empirical regression coefficients may be widely different from the true ones if one of those variables contains even relatively small errors.

However, if we can in some way or other obtain estimates of the variances and covariances of errors which we know to be present in empirical data, we can use eq. (5) to compute true variances and covariances of the variables. Having thus removed the influence of the errors, the true regression and correlation coefficients can be found. This we shall now try to do in the case of the atmospheric effects on the hard component of the cosmic radiation.

## II. Application to the cosmic-ray atmospheric effects

### 1. *Introduction*

As an application of the preceding general considerations we shall discuss the problem of determining the regression coefficients representing the atmospheric effects on the hard component of the cosmic radiation at sea-level. This component is usually defined as that part of the radiation which is able to penetrate 10 cm of lead. It consists of nearly only $\mu$-mesons, and their number depends on the atmospheric conditions. The intensity of the hard component is usually correlated with the barometric pressure $B$ at sea-level, the height $H_1$ of the 100 mb level and the temperature $T_1$ at or near that level. The barometer effect is mainly due to the absorption of the radiation in the atmosphere. The correlation with the height of the 100 mb level is due to the dependence of the disintegration probability of the $\mu$-mesons on the distance between their production levels and sea-level, whereas the origin of the (positive) temperature effect is more complicated.

The interpretations of these effects have been discussed by TREFALL (1955 a, 1955 b and 1956), and theoretical estimates of the regres-

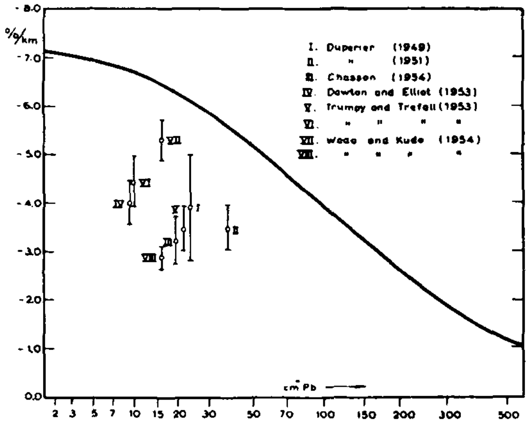

Fig. 1. In this figure are plotted theoretical estimates and empirical values of the height coefficient belonging to the set of independent variables $B$, $H_1$ and $T_1$, as functions of the amount of absorbing material in the cosmic-ray recording apparatus. The recordings in Manchester were performed with about 10 cm of lead.

sion coefficients have been calculated and compared with empirical results. It was then found (TREFALL 1956) that the empirically determined values of the height coefficient (the partial regression coefficient with respect to $H_1$, often also called the negative temperature coefficient) had usually too small numerical values to be compatible with the theoretical estimate. This is clearly shown in Figure 1, where curve $B$ gives the theoretically expected dependence of the height coefficient on the amount of absorbing material in the meson recorder. At that time no satisfactory explanation could be found for the discrepancy, but it now appears that it is probably due to errors in the aerological data which have been used as independent variables in the regression equation for the determination of the atmospheric effects.

## 2. *Estimation of errors in aerological data*

The observational errors in radiosonde data have been studied in later years by several authors. Some of their results are given in Tables 1 a, b and c. RAAB and RODSKJER (1950) used twin soundings, and obtained the error standard deviations given in Table 1 a. From night soundings ROSSI (1952) obtained similar results (Table 1 b). From the Payerne comparison of radiosondes NYBERG (1952) computed the standard deviations given in Table 1 c for the errors in night releases. These values are probably lower than the errors of aerological data obtained under ordinary working conditions, when there is less careful inspection of the instruments and less reliable conversion of transmitted data. In addition all aerological records incorporate influences from small-scale systems such as cumulus convection, gravity waves and others. Such disturbances are "noise" on the synoptic scale, and effectively increase the errors beyond those obtained from twin soundings.

**Table 1 a. Estimates of error standard deviations for the Finnish radiosonde**
(RAAB and RODSKJER 1950)

| Meteorological element | Isobaric levels | |
|---|---|---|
| | 350 mb—150 mb | ≧ 150 mb |
| Temperature | 0.57 °C | 0.61 °C |
| Pressure..... | 6.3 mb | 4.5 mb |

**Table 1 b. Error standard deviations for night twin soundings with the Finnish radiosonde**
(ROSSI 1952)

| Meteorological element | Isobaric levels | | |
|---|---|---|---|
| | 275 mb | 162 mb | 96 mb |
| Temperature....... | 0.68 °C | 0.72 °C | 0.78 °C |
| Pressure........... | 4.1 mb | 3.0 mb | 3.2 mb |

**Table 1 c. Error standard deviations for the Payerne data of 1950**
(NYBERG 1952)

| Meteorological element | Radiosonde type | Isobaric levels | |
|---|---|---|---|
| | | 500 mb—300 mb | 300 mb—200 mb |
| Temperature | Finnish<br>All six tested | 0.54 °C<br>0.47 °C | 0.86 °C<br>0.81 °C |
| Pressure.... | Finnish<br>All six tested | 5.4 mb<br>5.2 mb | 5.9 mb<br>5.8 mb |

A better method for determining the total error variance would be to perform simultaneous or approximately simultaneous soundings at stations separated by a distance of the order of 10 km. Such a pair of stations now exist on Spitzbergen, where the Norwegian station at Isfjord Radio and the Russian station at Barentsburg are situated only 18 km apart. If $Y_i$ and $Y_j$ denote the deviations of two meteorological elements from their respective mean values, we have

$$Y_{i,I} - Y_{i,B} \approx \varepsilon_{i,I} - \varepsilon_{i,B} \tag{33}$$

and

$$Y_{j,I} - Y_{j,B} \approx \varepsilon_{j,I} - \varepsilon_{j,B}, \tag{34}$$

where subscripts $I$ and $B$ refer to Isfjord Radio or Barentsburg data, respectively. As there can be no significant correlation between the errors in corresponding data from the two stations, we find that the error covariances of the stations are related by the equation

$$\overline{\varepsilon_{i,I}\varepsilon_{j,I}} + \overline{\varepsilon_{i,B}\varepsilon_{j,B}} \approx \overline{(Y_{i,I} - Y_{i,B})\,(Y_{j,I} - Y_{j,B})}. \tag{35}$$

Putting $i=j$ we obtain an equally valid relation for error variances.

In order to separate the error variances of the two stations some assumption must be made regarding their relative reliability. As there is nothing to suggest that one station is more reliable than the other, our best estimate of the error covariance of a pair of meteorological elements observed at any station is, according to the method of maximum likelihood,

$$\overline{\varepsilon_i\varepsilon_j} = \frac{1}{n}\sum (Y_{i,I} - Y_{i,B})\,(Y_{j,I} - Y_{j,B}) \tag{36}$$

where the summation has been performed over the $n$ available data. Error variances are again obtained by putting $i=j$. Having thus estimated both variances and covariances of errors, correlation coefficients between errors may also be computed. The results are presented in Tables 2 and 3. In order to check the stability of these statistical parameters the data were divided in two groups of approximately equal sizes. The variances are apparently very stable, as was confirmed by means of the $F$-test. The correlation coefficients are not so stable, even though none of the differences are statistically significant at the 5 % level.

The existence of correlations between errors is, of course, due to the fact that the heights of the isobaric levels are not measured directly, but are computed from records of the atmospheric temperature distribution as a function of pressure. Any error in the temperature measured at a given pressure level will, therefore, contribute to the error in the heights of

**Table 2. Estimated error variances for heights and temperatures**

| Period | 300 mb | | | 200 mb | | | 100 mb | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\varepsilon_T^2}$ (C°)² | $\overline{\varepsilon_H^2}$ m² | $n$ | $\overline{\varepsilon_T^2}$ (C°)² | $\overline{\varepsilon_H^2}$ m² | $n$ | $\overline{\varepsilon_T^2}$ (C°)² | $\overline{\varepsilon_H^2}$ m² | $n$ |
| 17 Sept.—29 Oct. 1957 ........... | 2.74 | 1743 | 40 | 5.09 | 2862 | 33 | 5.25 | 5229 | 19 |
| 1 Nov. 57—6 Jan. 1958 ......... | 2.42 | 1991 | 46 | 4.39 | 2434 | 36 | 4.51 | 4986 | 17 |
| Both combined.................... | 2.56 | 1898 | 86 | 4.69 | 2772 | 69 | 5.04 | 5074 | 36 |

**Table 3. Correlation coefficients between errors**

| Period | Errors of: | | |
|---|---|---|---|
| | $H_1$ and $T_1$ | $H_1$ and $H_2$ | $H_1$ and $H_3$ |
| 17 Sept.—29 Oct. 1957 ...................... | 0.589 | 0.766 | 0.728 |
| 1 Nov. 57—6 Jan. 1958 .................... | 0.495 | 0.508 | 0.252 |
| Both combined............................ | 0.560 | 0.628 | 0.465 |

all higher levels. Only errors in the thicknesses of not overlapping atmospheric layers are expected to be uncorrelated. A further check on the correlations obtained could therefore be performed by computing correlation coefficients between the errors in the height $H_2$ of the 200 mb level and the thickness $H_1 - H_2$ of the 100 - 200 mb layer, and between the errors in the height $H_3$ of the 300 mb level and the thickness $H_1 - H_3$. The values obtained for the combined periods were - 0.050 and + 0.188, respectively, none of which are significantly different from zero.

ELIASSEN (1954), who analysed data from British stations, estimated by means of a different method the error variance of the height of the 300 mb level to be 2,092 $m^2$. This agrees very well with the results that are given in Table 2, and indicates that the error estimates in Table 2 can be used in the following discussion.

## 3. *Evaluation of corrected cosmic-ray regression coefficients*

After having thus obtained estimates of the error variances and covariances of the necessary meteorological elements, corrections were applied according to (5) to the aerological data used in the analysis of some cosmic-ray records from Manchester. The errors in the sea-level pressure records could be neglected because their variance was negligible compared to the total pressure variance. The aerological data came from the Liverpool station some 50 km away, whereas the pressure data were recorded in Manchester. As the cosmic-ray data represented the integrated intensity throughout a day, corresponding mean values were used for the meteorological elements. The daily mean of the barometric pressure was based on 24 hourly readings, and the aerological data on 4 radiosonde ascents per day. Only days on which all soundings reached the 100 mb level were used. When the corrections for errors were applied, allowance was made, of course, for the fact that we were using daily mean values of the meteorological elements.

Two periods were selected for analysis, one from July 13th to September 11th 1950, and another from September 1st to November 7th 1951, containing 34 and 40 useful days respectively. The total variances of the meteorological elements were quite different for the two periods, being very low in 1950 and rather high in 1951. This is clearly shown in Table 4, where we have given ratios between total variances of aerological data for each of the two periods and our estimated error variances. We therefore expect the regression coefficients from the 1950 period to be much more affected by the random errors of measurement than those from 1951.

In Table 5 we have given uncorrected and corrected regression and correlation coefficients corresponding to three different sets of independent variables. The reasons for introducing the second and third sets of variables will be given later. It should here only be noted that the theoretical estimates of the height coefficient $b_2'$ are nearly equal for all three sets of variables, its numerical value increasing only slightly as we go from the first to the third set. The temperature coefficient $b_3'$ will, however, depend critically on the choice of variables and is, in fact, expected to change from positive to negative values as we go from the first to the third set. The corresponding change in $b_1'$ is negligible.

The differences between the uncorrected and corrected height coefficients for the 1950 period are remarkable. Even though the values of $b_2$ obtained with the three different sets of variables are on the average only about 60 % of the corresponding theoretical values, the corrected coefficients $b_2'$ agree very well

**Table 4. Ratios between total variances and estimated error variances for the Liverpool aerological data**

| Period | Variable | | | | | |
|---|---|---|---|---|---|---|
| | $T_1$ | $H_1$ | $H_2$ | $H_3$ | $H_1$—$H_2$ | $H_1$—$H_3$ |
| 1950............ | 5.6 | 2.1 | 11.9 | 21.2 | 6.4 | 6.6 |
| 1951............ | 11.4 | 18.7 | 45.0 | 58.8 | 10.3 | 11.0 |

**Table 5. Regression and correlation coefficients for cosmic-ray atmospheric effects**

| Variables | Period | Uncorrected | | | | Corrected | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $b_1$ %/km | $b_2$ %/km | $b_3$ %/°C | $R$ | $b_1'$ %/km | $b_2'$ %/km | $b_3'$ %/°C | $R'$ |
| $B, H_1, T_1$ | 1950 | — 1.46 | — 3.48 | + 0.079 | 0.849 | — 1.37 | — 6.38 | + 0.067 | 0.860 |
| | 1951 | — 0.84 | — 5.95 | + 0.116 | 0.966 | — 0.80 | — 6.28 | + 0.124 | 0.971 |
| $B, H_2, (H_1—H_2)$ | 1950 | — 1.34 | — 4.13 | — 0.003* | 0.865 | — 1.25 | — 6.70 | — 0.059* | 0.875 |
| | 1951 | — 0.86 | — 6.00 | — 0.032* | 0.961 | — 0.83 | — 6.27 | — 0.035* | 0.965 |
| $B, H_3, (H_1—H_3)$ | 1950 | — 1.15 | — 4.72 | — 0.019* | 0.870 | — 1.08 | — 6.99 | — 0.099* | 0.879 |
| | 1951 | — 0.85 | — 6.52 | — 0.091* | 0.958 | — 0.85 | — 6.78 | — 0.102* | 0.963 |

* In order to facilitate comparisons between the regression coefficients $b_3$ or $b_3'$ obtained with different sets of independent variables, all coefficients marked by asterisks have been given in terms of equivalent variations of the mean temperature of the atmospheric layer under consideration.

with the theoretical estimate presented in Figure 1. Further, the rather high values of the barometer coefficient $b_1$ obtained with the first and second sets of variables are somewhat reduced when the corrections are applied. The temperature coefficient $b_3$ obtained with the first set seems, however, to be adversely affected by the correction.

The results from 1951 are not quite so satisfactory. This period is characterized by abnormally low values of the barometer coefficient $b_1$, and the even lower values of $b_1'$ are not encouraging. The height coefficients $b_2$ behave better, as they already before the correction agree fairly well with the theoretical estimates and the slight changes after the correction bring them into very good agreement with the corresponding coefficients for the 1950 period. The reason why the height coefficients of the 1951 period are so much less affected by our correction is, no doubt, that the total variances of the heights of all isobaric levels are greater in 1951 than in 1950. The temperature coefficients $b_3$ have quite reasonable values both before and after the correction.

Our correction has in all cases led to an increase of the multiple correlation coefficient. However, it was shown in the first part of this paper that only under very special circumstances would such a correction not lead to an increase in the multiple correlation coefficient. This would hold even if error variances and covariances were arbitrarily chosen within the range permitted by the general laws of statistics. The observed differences between $R'$ and $R$ cannot, therefore, be made the basis for any conclusion regarding the correctness of the estimated error variances. The real tests are that the variability of the regression coefficients obtained for different periods should be reduced and the agreement between empirical values and theoretical estimates should be improved.

The view that the low empirical height coefficients obtained by most workers may have been caused by the large errors in the height $H_1$ of the 100 mb level, is supported by the fact that higher values are usually found when monthly means of cosmic-ray intensity are correlated with corresponding meteorological data than when daily means are used (Bachelet and Conforto, 1956). As we expect the ratio of error variance to total variance to be smaller for monthly means than for daily means, such a trend in the empirically determined coefficients is to be expected. It must, however, be remembered that the characteristics of the day-to-day variations of the atmospheric temperature distribution differ from those of the seasonal variations. As the atmospheric temperature distribution is far from uniquely determined by the chosen variables, differences between "seasonal" and "day-to-day" regression coefficients may arise also in this way.

## III. Conclusion

The results obtained demonstrate how drastically empirical regression coefficients may be influenced by random errors of measurement. Therefore, if one out of a number of measured quantities can select different sets of variables, which would have contained approximately equal amounts of information had no errors been present, one should use that set for which the ratios between the error variances and the true variances of the variables are the lowest.

In Table 4 we gave ratios between the total variance and the estimated error variance of the quantities which have been employed as independent variables in the present analysis. It is not surprising to find that this ratio is much smaller for $H_1$ than for $H_2$ or $H_3$. The decrease of the ratio with increasing height is mainly due to the negative correlation between stratospheric and tropospheric temperatures, which makes the true variance of $H_1$ less than the corresponding variances for levels closer to the tropopause. As the error variance steadily increases with increasing height as shown in Table 2, the relative error variance (the ratio between error variance and total variance) increases very rapidly as we go from the troposphere to the stratosphere.

From the statistical point of view the height $H_1$ of the 100 mb level is, therefore, a poor variable. Actually, the extensive use of $H_1$ in the study of the atmospheric effects on the cosmic radiation seems to be somewhat fortuitous. It is true that Duperier, who was the first one to correlate cosmic-ray intensity with heights of isobaric levels, obtained a better correlation with $H_1$ than with heights of lower levels (DUPERIER 1945). However, later results by Duperier and others (DUPERIER 1949, WADA 1951, BACHELET and CONFORTO 1956) do not generally show this preference for the 100 mb level, and actually indicate that on the average the correlation is better with the 200 mb level than with the 100 mb level. This is really not surprising, as the mean level of meson production appears to lie near 150 mb.

Even though it might be desirable to avoid the 100 mb level altogether and use only data from lower levels, theoretical considerations show that some information about higher levels is also needed for a really satisfactory representation of the cosmic-ray atmospheric effects. This is the reason why $H_1$ has been retained in our second and third sets of variables, and only the temperature $T_1$ at the 100 mb level has been replaced by the height of some lower isobaric level. From the statistical point of view and according to Table 4 these sets should certainly be better than the first one. The results given in Table 5 show that the corrected regression coefficients are most stable when the third set of variables is employed, but no final conclusion should be based on the results obtained from these two small samples only.

The reasons why we have used the thickness of a certain layer (the difference between two heights) as the third variable of the second and third sets rather than the single height $H_1$ are threefold: Firstly, these sets of variables are essentially the same type as the first one because the third variable is a measure of the mean temperature of that layer, and this makes the comparison of regression coefficients very easy. Secondly, we now avoid the intercorrelation between errors which makes the detailed study of their effects so difficult. Finally, with this type of regression equation each term has a more direct physical significance than in the other case.

As a final remark we mention that the relative error variance of the atmospheric temperature has a minimum near the 200 mb level, where it is less than half the value at 100 mb (NORDÖ, 1958). Remembering that the mean level of meson production lies near 150 mb, one would expect the barometric pressure $B$, the height $H_2$ of the 200 mb level and the temperature $T_2$ at this level to be a very good set of variables for the representation of the cosmic-ray atmospheric effects. The reason why we have not tried this set is that the main purpose of the present paper was to point out the important effects which random errors of measurements generally may have on empirical regression coefficients. The cosmic-ray case should mainly serve as an example indicating the applicability of the mathematical considerations given above.

## IV. Acknowledgments

We are indebted to Mr. P. Breistein of the Arctic Group, Det Norske Meteorologiske Institutt, Oslo, for providing aerological data from Isfjord Radio and from Barentsburg, and to Dr. H. Elliot, Imperial College, London, for the use of cosmic-ray records from Manchester. The aerological data from Liverpool were taken from the Daily Aerological Record of the Meteorological Office, Londo:ı.

Finally, we wish to express our best thanks to Forskningsfondet av 1919 for financial support of this work obtained by one of the authors (Trefall).

## REFERENCES

BACHELET, F., and CONFORTO, A. M., 1956: Atmospheric Effects on the Cosmic Ray Total Intensity at Sea Level, *Il Nuovo Cimento, Series X*, **4**, p. 1479.

DUPERIER, A., 1945: The Geophysical Aspects of Cosmic Rays. *Proc. Phys. Soc. A*, **57**, p. 464.

DUPERIER, A., 1949: The Meson Intensity at the Surface of the Earth and the Temperature at the Production Level. *Proc. Phys. Soc. A*, **62**, p. 684.

ELIASSEN, A., 1954: Provisional Report on Calculation of Spatial Covariance and Autocorrelation of the Pressure Field. *Videnskapsakademiets Institutt for Vær- og klimaforskning, Oslo. Rapport nr* **5**, 1954.

NORDÖ, J., 1958: Unpublished manuscript.

NYBERG, A., 1952: On the Comparison of Radiosonde Data in Payerne, May 1950. *Sveriges Met. Hydr. Inst. Medd., Serie B, nr* **9**.

RAAB, L. and RODSKJER, N., 1950: A Study of the Accuracy of Measurements of the Väisälä Radiosonde. *Arkiv för Geofysik*, **1**, No. 2.

ROSSI, V., 1952: On the Accuracy of the Finnish Radiosonde. *Geophysica*, **4**, No. 2.

SPEARMAN, G., 1904: The Proof and Measurement of Association between two Things. *Amer. J. Psych.*, **15**, p. 88. (See also YULE and KENDALL: *An Introduction to the Theory of Statistics*. Charles Griffin Co., Ltd., London 1949.)

TREFALL, H., 1955 a: On the Positive Temperature Effect in the Cosmic Radiation and the $\mu$-$e$ Decay. *Proc. Phys. Soc. A*, **68**, p. 893.

— 1955 b: On the Barometer Effect on the Hard Component of the Cosmic Radiation. *Proc. Phys. Soc. A*, **68**, p. 953.

— 1956: On the Interpretation of the Atmospheric Effects on the Hard Component of the Cosmic Radiation. *Universitetet i Bergen, Årbok 1956, Naturvitenskapelig rekke, nr* **10**.

WADA, M., 1951: The Relation between Cosmic-Ray Intensities and Heights of Isobar Levels. *J. Sci. Res. Inst., Tokyo*, **45**, p. 77.