الجمهورية العربية المتحدة

معهد التخطيط القومي

Memo. No. 435

LECTURE NOTES ON REGRESSION ANALYSIS
PART II

by

Mr. C.J. van Rees

May 18,1964

القاهرة
٣ شارع محمد مظهر - بالزمالك

CONTENTS

1. Introduction.

This second exploration in the field of regression will add one or more dimensions to our range of vision. For in practice it will not be possible in general to explain one phenomenon simply by one other phenomenon . Due to the intricate pattern of our community often many explanatory variables are necessary to provide an insight in the development of another variable . Thus the consumption of limonade by a person can be explained by his income. Often this is not enough and the price of limonade has to be taken into account. If this person is living in a country with strongly fluctuating temperatures, we may expect an influence from the scope of the temperature on the consumption of limonade. In this way it will be possible to show even more factors playing a role in the explanation of this man's consumption.

If we are convinced that a causal relation exists, it will be desirable to draw a scatter in order to obtain an idea about the extent of the relation. However, practical objections limit us in the execution of our task. After that, we have to specify the relation by use of the available mathematical techniques, See Section 2. Because we have more than one explanatory variable here, this is called <u>multiple</u> regression. The techniques will be outlined in Section 3. This section will also give formulae to show the fit of the relation. Not only the resulting regression coefficients, but also the values of the variables give us an indication of the importance of the several explanatory variables. With the help of a so-called regression chart this will be demonstrated in section 4.

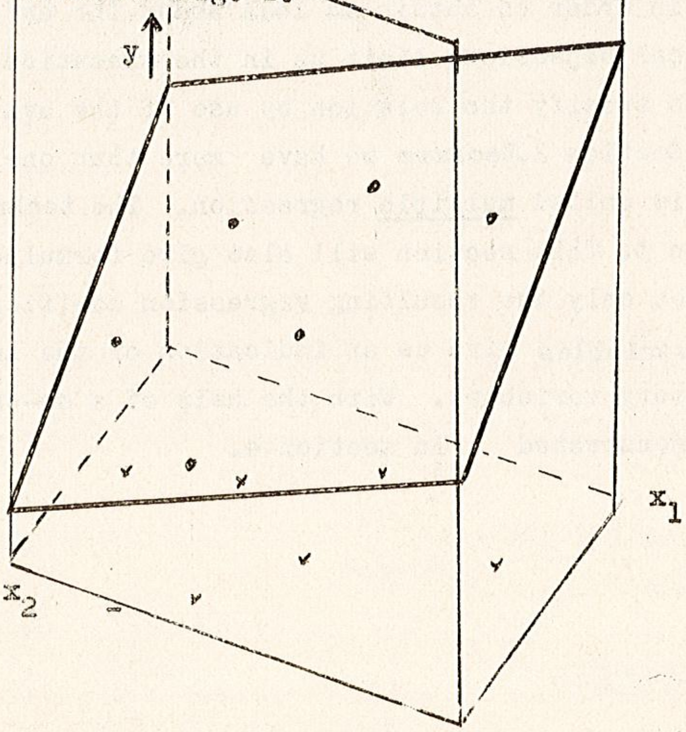
The use of matrix notation in regression analysis is widespread . An introduction to it will be given in the Appendix.

2. Preparatory work for the multiple regression

2.1. The scatter diagram:

Our first care in the estimation of a relation between more than two variables will be similar to the simple regression case. It leads us to the determination of the variables playing a role in the economic process under hand. After considering the dependent variable we enter the second step of our procedure, i.e., drawing a scatter diagram.

In the introduction we already pointed to the fact that practical considerations limit our possibilities here. Simple regression presented to us a scatter diagram in a two dimensional space. Adding one dependent variable to the relation brings us to the simplest case of multiple regression. The three dimensional scatter diagram can be shown in a box. A sketch is shown in fig. 1.



Two explanatory variables (x_1 and x_2) are shown in a horizontal space. The dependent variable (z) in reality gives the scatter its three-dimensional shape. The dotted lines show the intersecting-lines of the plane through the points of the scatter with the walls of the box.

Then our possibilities are exhausted. We have no physical means of expressing a scatter in four or even more dimensions. Also it is not possible to draw a scatter between x_1 and y only, because the result will generally be disturbed by the influence of the remaining explanatory variable x_2 . This will be even worse with more than two explanatory variables.

After the estimation of the regression coefficients of the relation there are some ways to research the influence of each explanatory variable and the linearity of the relation. This is shown in section 4.

2.2 <u>The Mathematical form of the relation:</u>

In part I we discovered the complexity of the problem to fix the mathematical form of the relation. The numerous possibilities did not allow us to give a complete picture. The explanation of the consumption of food illustrated the method to be followed.

Sometimes an explanatory variable has to be included in the postulated relation not only in its linear form, but also as a square. This **serves** to account for a non-linearity in the relation. The easiest example is a quadratic relation in two variables:

$$y = a + b x + c x^2$$

In the computations of the regression coefficients as shown below this relation is interpreted as:

$$y = a + bx_1 + cx_2$$

i.e., a relation in two explanatory variables x_1 and x_2, where $x_1 = x$ and $x_2 = x^2$. For the mathematical handling it makes no difference if we work with the square of an economic variable!

3. The techniques of multiple regression:

3.1. The estimation of the multiple regression coefficient:

In order to be able to find values for the regression coefficients we need n observations for each of the explanatory variables and the same condition goes for the dependent variable.
To estimate the regression coefficients for the x and y values we use the same principle of least squares as we used in Part I, section 4. The postulated relation is written as.

$$y = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

In general this relation will not suit for all values of x and y. Therefore we should write.

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki} + v_i$$

where x_{ji} is the i-th observation of the j-th explanatory variable and v_i is the disturbance of the i-th observation. The subscript i goes, from 1 to n. According to the principle of least squares we should minimize the sum of squares of all disturbances. This sum is written as.

$$\sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} (y_1 - a - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki})^2$$

According to our principle this function of a and $b_1 \ldots b_k$ should be minimized with respect to a and $b_1 \ldots b_k$. The partial derivative with respect to a is

$$\frac{\partial (v_i^2)}{\partial a} = -2 \sum_{i=1}^{n} (y_1 - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})$$

we obtain a minimum by equalizing this form to zero.

$$0 = -2 \sum_{i=1}^{n} (y_1 - a - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_h x_{ki})$$

Dividing the relation by - 2n, implies

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \ldots - b_k \bar{x}_k$$

Where y, x_1, x_2, ... x_k are the averages of y_i, x_{1i}, x_{2i}, .. x_{ki}. This means we can compute a similar to the method used for simple regression. So, when all values for b are estimated, we have a simple formula to compute a. Further it shows that the least square plane goes through the point of avarages. To simplify our argument we shall estimate the values for b using a relation with two explanatory variables only, Now $\sum v_i^2$ is equal to.

$$\sum v_i^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2$$

Differentiating this relation with respect to b_1 gives

$$\frac{\partial (\sum v_i^2)}{\partial b_1} = -2 \sum_{i=1}^{n} x_{1i} (y_1 - a - b_1 x_{1i} - b_2 x_{2i}) = 0$$

When we substitute our relation for a in the last relation we see

$$\sum_{i=1}^{n} x_{1i} \left[y_i - \bar{y} - b_i (x_{1i} - \bar{x}_1) - b_2 (x_{2i} - \bar{x}_2) \right] = 0$$

We may substitute $Y_i = y_i - \bar{y}$, $X_{1i} = x_{1i} - \bar{x}_1$ and $X_{2i} = x_{2i} - \bar{x}_2$. This gives:

$$\sum_{i=1}^{n} x_{1i} \left\{ Y_i - b_1 X_{1i} - b_2 X_{2i} \right\} = 0.$$

$$\text{or} \sum_{i=1}^{n} x_{1i} y_1 - b_1 \sum_{i=1}^{n} x_{1i} X_{1i} - b_2 \sum_{i=1}^{n} x_{1i} X_{2i} = 0 .$$

In Part I we saw that $\sum X_1 y_i$ can be written as $\sum X_i y_i$. Applying this here, and writing part of the relation on the left hand side gives:

$$\sum_{i=1}^{n} X_{1i} y_i = b_1 \sum_{i=1}^{n} X_{1i} + b_2 \sum_{i=1}^{n} X_{1i} X_{2i}$$

The same procedure can be followed for the differentiation with respect to b_2. This gives.

$$\sum_{i=1}^{n} X_{2i} y_i = b_1 \sum_{i=1}^{n} X_{1i} X_{2i} + b_2 X_{2i}^2$$

We can now extend our argument for the relation with k explanatory variables. This gives us a set of k equations, written as:

$$\sum X_{1i} y_2 = b_1 X_{1i}^2 + b_2 \sum X_{1i}^2 X_{2i} + \dots + b_k \sum X_{1i} X_{ki}$$

$$\sum X_{2i} y_i = b_1 \sum X_{2i} X_{1i} + b_2 \sum X_{2i}^2 + \dots + b_k \sum X_{2i} X_{ki}$$

$$\vdots$$

$$\sum X_{ki} y_i = b_1 \sum X_{ki} X_{1i} + b_2 \sum X_{ki} X_{2i} + \dots + b_k \sum X_{ki}^2$$

We can simplify the notation by omitting the subscript i, because it appears in every part of the system. The k equations above are called the k normal equations. They are linear in b and these values can be solved after computing $\sum X_1^2$, $\sum X_1X_2$, ... and $\sum X_k^2$. **When we** make use of matrix notation the computations are highly facilitated. The linear equation

$$y = a + b_1x_1 + \dots + b_k x_k$$

is called the <u>multiple regression equation</u>, b_1, .. b_k are called the partial regression coefficient.

Some special cases can now be distinguished. When. $k = 1$ the system has only one equation and the remaing coefficient (b_1) is the simple regression coefficient. **Sometimes all cross products are equal to zero:**

$$\sum X_1 X_2 = \sum X_1 X_3 = \dots = X_{k-1} X_k = 0.$$

This reduces the normal equations to:

$$\sum X_1 Y = b_1 \sum X_1^2, \sum X_2 Y = b_2 \sum X_2^2, \text{ etc.}$$

The first statement means, that we have no correlation between the explanatory variables[1]. The second statement means that the values for b_1 ... b_k obtained here are the same as the values of b_1... b_k estimated by consecutive simple regression of y with all explanatory variables separately.

When we obtained the formulas for the simple regression coefficients a and b, we started by taking the value for a equal to zero. This can be **done** with multiple regression as well. It will be clear, we have to replace the values in deviations by the values in absolute terms. E.g. the first of the normal equations takes the form.

$$\sum x_1 y = b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \dots + b_h \sum x_1 x_k.$$

So, the difference between the formulas is analogous to the difference existing in the case of simple regression.

The last method can also be used, when we have a multiple regression with a

1) E.g. the sum $\sum X_1X_2$ is the nominator of the simple correlation coefficient of a regression between x_1 and x_2.

constant term $a \neq o$. Therefore we write $a = b_o x_{oi}$, where $b_o = a$ and $x_{oi} = 1$ for $i = 1 . n$. Then the general relation is written as:

$$y_i = b_o x_{oi} + b_1 x_{1i} + \dots + b_k x_{ki} + v_i$$

We have replaced our constant term by an extra variable, which is advantageous, because the computations will not be in deviations now. On the other hand one extra normal equation has been added to the system.

3.2 The multiple correlation coefficient:

Anologous to the simple correlation coefficient we have to define a coefficient which gives us any idea about the fit of the relation. We define the actual value for y_1 as.

$$y_1 = a + b_1 x_{1i} + \dots + b_k x_{ki} + v_i$$

Further, the regression value for y is equal to:

$$\hat{y}_i = a + b_1 x_{1i} + \dots + b_k x_{ki}$$

Using the same notation as in the preceding section we may write

$$\hat{Y}_i = b_1 X_{1i} + b_2 X_{2i} + \dots b_k X_{ki}$$

The multiple correlation coefficient is defined as..

$$R = \frac{\sum Y_1 \hat{Y}_1}{\sqrt{\sum Y_1^2 \sum \hat{Y}_1^2}}$$

In words: The multiple correlation coefficient is equal to the simple correlation coefficient between the actual value of the dependent variable and the regression value of the dependent variable

To obtain upper and lower limits for the value of R it is necessary to simplify this formula. For this purpose we use the normal equations. They can be developed as follows:

$$0 = \sum_{i=1}^{n} X_{1i} Y_i - b_1 \sum_{i=1}^{n} X_{1i}^2 - \dots - b_k \sum_{i=1}^{n} X_{1i} X_{ki}$$

$$= \sum X_{1i} (Y_i - b_1 X_{1i} - \dots - b_k X_{ki})$$

$$= \sum X (Y_i - \hat{Y}_i)$$

$$= \sum X_{1i} (y_i - \bar{y} - \hat{y}_i + \bar{\hat{y}})$$

The first two relations of this section show us that $y_i - \hat{y}_i = v_i$. Further we can easily prove $\bar{y} = \bar{\hat{y}}$. Substituting both results in our relation gives $\sum X_{1i} v_i = 0$. This shows that the first dependent variable is not correlated with the disturbances. This can be proved for all dependent variables using the remaining normal equations as a starting point. It follows that no correlation exists between the disturbances and all dependent variables. This result can be used to substitute it in the nominator of the correlation coefficient:

$$\sum Y_1 \hat{Y}_i = \sum (\hat{Y}_1 + v_i) \hat{Y}_i$$

$$= \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \hat{Y}_i v_i$$

The last part of this relation is equal to zero. This is proved as follows.

$$\sum_{i=1}^{n} \hat{Y}_i v_i = \sum_{i=1}^{n} (b_1 X_{1i} + b_2 X_{2i} + \dots b_k X_{ki}) v_i$$

$$= b_1 \sum_{i=1}^{n} X_{1i} v_i + b_2 \sum X_{2i} v_i + \dots + b_k \sum_{i=1}^{n} X_{ki} v_i = 0$$

Combining this with the preceding result shows us.

$$\sum Y_1 \hat{Y}_1 = \sum \hat{Y}_1^2$$

As the nominator of the correlation coefficient is always greater than or equal to zero, the same goes for the correlation coefficient itself (the nominator is a sum of squares, i.e. always $\geqslant 0$; the denominator is the square root of two sums of squares multiplied with each other, i.e., always $\geqslant 0$). An upperlimit for the correlation coefficient can be found along the same lines as performed in part I for the simple correlation coefficient. So write the variance of the disturbances as.

$$S_v^2 = \frac{1}{n} \sum_{i=1}^{n} (v_i - \bar{v})^2 = \frac{1}{n} \sum v_i^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \frac{1}{n} \sum Y_i \hat{Y}_i$$

Also, the formula for R^2 can be changed by substituting $\sum Y_i \hat{Y}_i = \sum \hat{Y}_i^2$. This gives.

$$R^2 = \frac{(\sum \hat{Y}_i Y_i)^2}{\sum Y_i^2 \sum \hat{Y}_i^2} = \frac{\sum Y_i \hat{Y}_i}{\sum Y_i^2}$$

And $1 - R^2$ is equal to

$$1-R^2 = \frac{\sum Y_i^2}{\sum Y_i^2} - \frac{\sum Y_i \hat{Y}_i}{\sum Y_i^2}$$

Combining this result with the formula of S_v^2 we can write

$$S_v^2 = (1 - R^2) S_y^2$$

or

$$\frac{S_v^2}{S_y^2} = 1 - R^2$$

Again the highest value for R^2 to be obtained here is $R^2 = 1$, with $S_v^2 = 0$. Combining both upper and lower limit for R gives.

$$0 \leqslant R \leqslant 1$$

Are there reasons to consider a regression as giving a good or moderate fit in certain cases. It is very difficult to give a definite answer to this question. Each case has to be considered on its own merit. Some indication can be given. If we have a macro economic study based on time series $R = 0,8$ is a low value. A value $R > 0,95$ is no exception here. If we estimate Engel curves of the consumption of families, using households budget material, $R > 0,7$ in a very good result. It is not exceptional here to obtain a value $R < 0,5$.

Another point which has to be taken into consideration is the level of aggregation applied to the goods or groups. In general the correlation coefficient will show a larger value according as we have a more pronounced aggregation. E.g. this accounts for the difference found in the estimation of Englecurves for families and the estimation of macro economic function.

3.3 The Partial correlation coefficient

In the last section we outlined the coefficient R. This gives us an impression of the over-all correlation between the dependent variable and the explanatory variables. If we limit ourselves to the three-variable case, a high multiple correlation coefficient does not necessarily mean a clear association between y and x_1. This net association may merely be due to the common influence of x_2 on them. The partial correlation coefficient between y and x_1 tries to remove the influence of x_2 from each of the other two variables. The mathematical procedure is the following:
We take the linear regressions of y on x_1 and x_1 on y_1. This gives the system.

$$y_i = a' + b_{02} x_{2i} + u_i$$

$$x_{1i} = a'' + b_{12} x_{2i} + w_i$$

In the subscripts of b, the first variable indicates the variable on the left-hand side of the equation, the second indicates the variable to which it is attached. This system can easily be written in deviations. At the same time we write the disturbances on the left-hand side. This gives

$$u_i = Y_i - b_{o2} X_{2i}$$

$$w_i = X_{1i} - b_{12} X_{2i}$$

The partial correlation coefficient is defined as the correlation between the unexplained residuals that remain, after removing the influence of X_2. This means the partial correlation coefficient between Y and x_1 is equal to .

$$r_{01.2} = \frac{\Sigma u_i w_i}{\sqrt{\Sigma u_i^2 \Sigma w_i^2}}$$

In the subscript of r, the figures before the point denote the variables between which correlation is taken[1]. The subscript after the point denotes the variables which is kept constant. It is possible to substitute $S_v^2 / S_y^2 = 1 - r^2$ in this relation (see Part I,p.25). This results in:

$$r_{o1.2} = \frac{\Sigma (Y_i - b_{02} X_{2i}) (X_{1i} - b_{12} X_{2i})}{\sqrt{\Sigma Y_1^2 (1 - r_{o2}^2) . \Sigma X_{1i}^2 (1 - r_{12}^2)}}$$

---

1) The subscript o refers to y.

Where r_{o2} means the simple correlation coefficient of y on x_2. We can write the nominator of this coefficient in full:

$$r_{01.2} = \frac{\sum Y_i X_{1i} - b_{02} \sum X_{1i} X_{2i} - b_{12} \sum Y_i X_{2i} + b_{02} b_{12} \sum X_{2i}^2}{\sqrt{\sum Y_i^2 . (1 - r_{02}^2) . \sum X_{1i}^2 . (1 - r_{12}^2)}}$$

The coefficient b of the simple regression can be rearranged as follows.

$$b = \frac{\sum Y_i \; Y_i}{\sum X_i^2} = \frac{\sqrt{\sum Y_i^2}}{\sqrt{\sum X_i^2}} \; \frac{\sum Y_i \; X_i}{\sqrt{\sum X_i^2 \sum Y_i^2}} = \frac{S_y}{S_x} r$$

Substituting this in the coefficient $r_{12.3}$ gives:

$$r_{01.2} = \frac{\sum Y_i X_{1i} - r_{o2} \frac{S_y}{S_{x_2}} \sum X_{1i} X_{2i} - r_{12} \frac{S_{x_1}}{S_{x_2}} \sum Y_i X_{2i} + r_{02} r_{12} \frac{S_y S_{x_1}}{S_{x_2}^2} X_{2i}^2}{\sqrt{\sum Y_i^2} \sqrt{\sum X_{1i}^2} \sqrt{1 - r_{02}^2} \sqrt{1 - r_{12}^2}}$$

This can be changed into:

$$r_{01.2} \frac{n \, r_{01} S_y S_{x_1} - r_{02} \frac{S_y}{S_{x_2}} . n \, r_{12} S_{x_1} S_{x_2} - r_{12} \frac{S_{x1}}{S_{x_2}} . n . r_{02} S_y S_{x_2} + r_{02} r_{12} \frac{S_y S_{x1}}{S_{x_2}} n S_{x_2}}{n \; S_y \; S_{x_2} \sqrt{(1 - r_{02}^2)} \sqrt{(1 - r_{12}^2)}}$$

Hence

$$r_{o1.2} = \frac{n \, S_y S_{x_1} (r_{o1} - r_{o2} - r_{12})}{n \, S_y S_{x_1} \sqrt{1 - r_{o2}^2} \sqrt{1 - r_{12}^2}}$$

Thus

$$r_{01.2} = \frac{r_{01} - r_{02} \; r_{12}}{\sqrt{1 - r_{02}^2} \; \sqrt{1 - r_{12}^2}}$$

Similarly we can develop the partial correlation coefficient between y and x_2, and x_1 and x_2. This gives the following farmulae.

$$r_{12.1} = \frac{r_{02} - r_{01} \; r_{12}}{\sqrt{1 - r_{01}^2} \; \sqrt{1 - r_{12}^2}}$$

and $$r_{12.0} = \frac{r_{12} - r_{01} \; r_{02}}{\sqrt{1 - r_{01}^2} \; \sqrt{1 - r_{02}^2}}$$

Without proof we present another formulation of the partial correlation. coefficient. This will help us in understanding its meaning. We find

$$r_{01.2}^2 = \frac{S_y^2 \; (R^2 - r_{02}^2)}{S_y^2 \; (1 - r_{02}^2)} = \frac{R^2 - r_{02}^2}{1 - r_{02}^2}$$

Now the denominator $S_y^2 (1 - r_{02}^2)$ shows us the variation in Y unexplained by x_2(see above). In the same may it applies to the multiple correlation coefficients, i.e., $S_y^2 R^2$ is the variation in y explained by x_1 and x_2 Combining the two last statements gives us that $S_y^2 (R^2 - r_{02}^2)$ is the increase in the explained variation in y due to x_1. From this we derive that the partial correlation coefficient between y and x_1 measures the proportion of the variation in y unaccounted for by x_2, that has been explained by the addition of a variable x_1.

In the same way the partial correlation cofficient between y and x_2 can be formulated as.

$$r^2_{02.1} = \frac{R^2 - r^2_{01}}{1 - r^2_{01}}$$

4. <u>Alternative geometrical representations.</u>

4.1. <u>The partial scatter diagram.</u>

As we have seen in Section 2 it is generally not possible to draw a scatter diagram when we have more than two explanatory variables in our regression. We can partially overcome this difficulty by drawing so-called partial scatter-diagrams.
After computing the regression coefficients and the constant term a, we can arbitrarily choose one of the explanatory variables, e.g. x_1 , and correct the dependent variable for all the remaining explanatory variables. After correction the dependent variable is equal to

$$y - b_2x_2 - \ldots\ldots - b_k x_k$$

According to the general equation for the multiple regression, this form is linear dependent on x_1;

$$y - b_2x_2 - \ldots\ldots - b_k x_k = a + b_1x_1$$

As this is a simple regression, we can construct a scatterdiagram. We have x on the horizontal ax and the corresponding " dependent" variable $(y - b_2x_2 - \ldots b_kx_k)$ on the vertical ax. This enables us to discover if this relation approximates a straight line[1]. This diagram is called a <u>partial scatterdiagram.</u>[2] We can deal with the remaining dependent variables

1- Due to the disturbances, not all of our observations will generally be on the straight line.

2- This diagram is called partial as we use only one explanatory variable, while the dependent variable is corrected for the other explanatory variable.

analogously. Altogether, this gives us k partial scatterdiagrams, one for each dependent variable. The aim of this procedure, is to find a linear relation in all cases. If we find a curved line in one of the scatters, we have to revise the functional relation. E.g., if the partial diagram for x_1 shows us that x_1^2 rather than x_1 may give a good approximation of a straightline in the partial scatter diagrams, we introduce x_1^2 in the multiple regression instead of x_1, and repeat the calculations for all the multiple regression coefficients. After doing so, we draw the partial scatter diagram for the new relation and correct possible remaining defects.

4.2. Regression Charts.

In many occasions, the observations are in the form of time series. In that case the index i relates to years, months etc. Then the regression can be illustrated with the use of regression charts. On the horizantal ax we present time, on the vertical ax the dependent variable is shown. The observations are plotted successively and connected with each other. This gives us the value of the dependent variable in the course of time. Further we do the same for $\hat{y}$, the regression value of the dependent variable. Then values are connected with a dotted line. Comparing the dotted line with the line representing $\hat{y}$, gives us an indea about the fit of the relation over time. In successive panels after the first, we present the numerical value of each explanatory variable multiplied by its respective coefficient. In this manner we obtain a picture of the contribution of each independent variable to the explanation of the dependent variable . In the bottom panel we present the residuals (v). As we mentioned above, these can also be ascertained from the top panel showing the regression and observed values of the dependent variable.

This idea will be illustrated by an example taken from L.R. Klein and A.S. Goldberger: An Econometric Model of the United States 1929-1952. In this economic research the authors present a model for the United States . In Chapter IV we find regression charts for each of the estimated relations. For our example we chose the labor Market Adjustment Equation. This equation explains the increase in the index of hourly wages ($w_t - w_{t+1}$). The unemployment in number of persons (N_u) played a very important role during that time of recession. For that reason N_u was introduced as an explanatory variable. Generally, workers will bargain for a wage increase, because the general price level increased. To show the lag in wages and price level ($p_{t-1} - p_{t-2}$) was used as an explanatory variable (p_t = price in t) As we lived in a time of general in inflation, money wage rates showed a general upward trend. This resulted in the introduction of a trend factor (t). After estimation, the equation took the following form:

$$w_t - w_{t-1} = 4.11 - 0,75\ N_u + 0,56\ (p_{t-1} - p_{t-2}) + 0,56\ t.$$

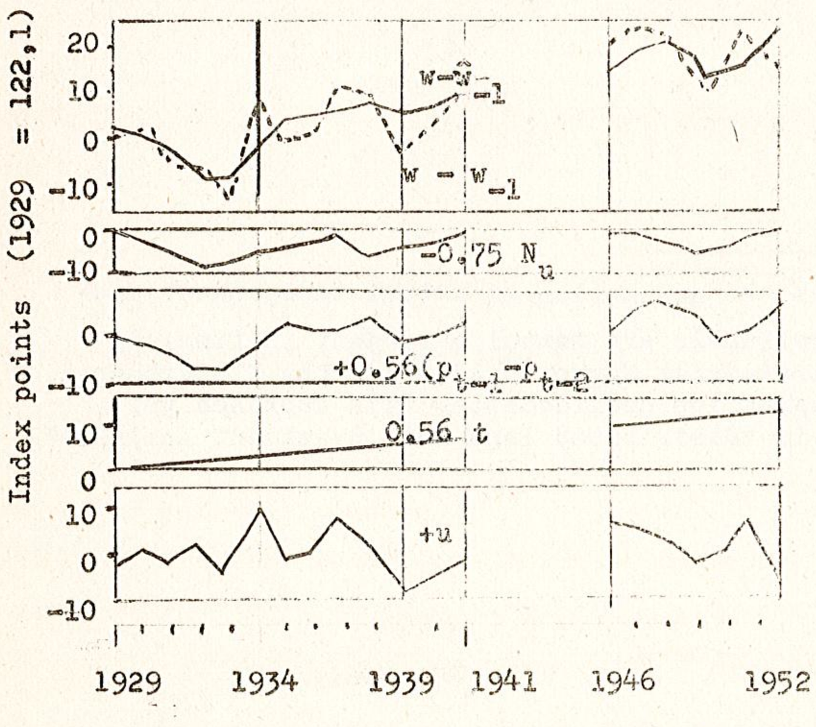


Fig. 2 Regression Chart of the Labor Market Adjustment Equation.

Figure 2 shows us time on the horizontal ax. Due to the exceptional circumstances during that period, observations for 1942-1946 were not included in the computations. Thus they are not represented in the chart. On the vertical ax the values are in index points (1939 = 122,1) The top panel gives the comparison between the actual and the regression value for the change in the index of hourly wages. Further, the influence of the number of unemployed, the change in price level and the trend factor are shown in successive parts of the chart. The bottom panel shows the disturtances. The disturbances show a stochastic pattern . This points to the absence of serial correlation[1]. In general serial correlation is the correlation between members of a time series and those members lagging behind or leading by a fixed distance in time .If the series is v_1 , b_2 , ... the serial correlation of order k is the correlation between the pairs (v_1 , v_{1+k}), (v_2 , v_{2+k}) , ...

1-) Serial correlation affects the assumption of mutual independent disturbances: In that case the residuals are mutual dependent in time. By making the assumption of a first-order Markov scheme for the disturbances, we find new values for the regression coefficients. This approach was first applied by L. M. Koyck in "Distributed lags and Investment Analysis" ch 2.

Appendix: The application of matrix notation to regression analysis

If we assume a linear relation between a variable y and (k-1) explanatory variables, the i-th observation of a sample of n observations can be written as

$$y_i = \alpha + \gamma_1 x_i + \gamma_2 x_{2i} + \dots + \gamma_{k-1} x_{k-1_i} + u_i$$

As we have seen in section 3 the constant α can be interpreted as the coefficient of a factor which has a constant value in all instances. Slightly changing the notation shows us the general formula in that case.

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + u_i$$

Introducing matrix notation shows for the same relation[1)]

$$Y = X\beta + u$$

where

$$Y = \begin{Bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{Bmatrix} \quad X = \begin{bmatrix} 1 & x_{2i} & \dots & x_{k1} \\ 1 & x_{22} & \dots & x_{k2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & x_{2n} & \dots & x_{kn} \end{bmatrix} \quad B = \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{Bmatrix} \quad y = \begin{Bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{Bmatrix}$$

This means that the rows of the system relate to the observations The columns of X show the values for a certain explanatory variable in successive observations. Usually we draw a sample of observations to estimate the coefficients of the relations. The estimates of the regression coefficients are denoted by $b = (b_1, b_2, \dots b_k)$ Now we may write the system as

$$Y = Xb + v$$

1- Usually in a book a matrix notation is denoted by bold letters i.e., X is printed as **X**. It will be clear, that this is impossible here.

where Y is a n x 1 vector of the dependent variable (n denotes the numbers of observations)[1], X is a matrix of order n x k , and v is a n x 1 sector of disturbances corresponding to the estimates for β . Applying the principle of least squares, we first show the matrix notation for the sum of squares:

$$\sum_{i=1}^{n} v_i^2 = v'v$$

$$= (Y-Xb)' \quad (Y-Xb)$$

$$= Y'Y - Y'Xb - b'X'Y + b'X'Xb$$

Both the second and the third term of the last relation are a scalar. This allows us to take the transpose without changing the value. Hence:

$$\sum_{i=1}^{n} v_i^2 = Y'Y - 2b'X'Y + b'X'Xb$$

To find the estimates for the regression coefficients we differentiate this relation with respect to the vector b. The necessary condition is

$$\frac{\partial (v'v)}{\partial b} = - 2\ X'Y + 2\ X'Xb = 0 .$$

This gives us

$$Y'Y = X'Xb.$$

or

$$b = (X'X)^{-1} \ X'Y$$

Notice, we used the assumption $k < n$, i.e., the number of observations exceeds the number of parameters to be estimated. This assumptions is needed in order that X'X is a non-singular matrix (X'X is of order k; the rank of X'X is equal to the rank of X; if $n < k$, X will be of rank n and hence

1- In estimating time series we use T instead of n for the number of observations.

X'X is of rank n; then X'X is singular and no solution exists). The initial system can be written as:

$$v = Y - Xb.$$

Premultiplying this relation by a matrix X' and using the condition for a minimum gives:

$$X'v = X'Y - X'Xb = 0.$$

We obtained the same result as shown in section 3, namely, the existence of a zero correlation between the explanatory variables and the disturbances.

References.

1) Hoel P.G. : Introduction to Mathematical Statistics, Wiley, New York, 1954.

2) Johnston J. : Econometric Methods, Mc Graw-Hill Book Company, Inc., New York, 1960

3) Klein L.R. : A Textbook of Econometrics, Harper and Row, New York, 1953.

4) Klein L.R. and Goldberger A.S. : An Econometric Model of the United States 1929-1952, North-Holland Publishing Company, Amsterdam, 1955.

5) Koyck L.M. : Distributed Lags and Investment Analysis, North - Holland Publishing Company, Amsterdam, 1954.

6) Theil H. : Economic Forecast and Policy, North-Holland Publishing Company, Amsterdam, 1958.

7) Wallis R.G.D. and Roberts H.V. : Statistics, A New Approach, Book Production Company, New York, 1956.

8) Yule G.U. and Kendall M.G. : An Introduction to the Theory of Statistics, Griffin, London, 1950.