REMARKS ON THE APPLICATION OF WEIGHTED REGRESSION

By

K. TETTAMANTI, R. STOMFAI,* S. KEMÉNY and J. MANCZINGER

Department of Chemical Unit Operations, Technical University, Budapest

Received February 2, 1977

To determine the relationship between variables methods of the regression analysis are generally used. Distinction can be made (but it is not necessary) between dependent and independent variables; there can be one or more dependent and/or independent variables.

In this paper the case of a single dependent and one independent variables will be discussed but our results are valid for cases of more than one independent variables, as well.

Let us denote the functional relationship between the dependent and the independent variables y and x $[Y = Y(X, \underline{\Theta})]$, where $\underline{\Theta}$ is the vector of parameters** (constants). Instead of the exact value of the dependent variable Y_i, measurements yield a datum point y_i subject to error ε_i; $y_i = Y_i + \varepsilon_i$. Data are processed by fitting an estimated regression function $\hat{Y} = Y(X, \underline{\hat{\Theta}})$ of the same type.

Discussion concerns only cases where the estimated regression function is of the same form as the theoretical regression function $Y = Y(X, \underline{\Theta})$.

I.

Let us consider first the case where the exact value of the independent variable can be measured, that is, it is no subject to error (it is a non-stochastic variable) $x_i \equiv X_i$. If the variance of the dependent variable is constant, method of least squares can be directly applied. If the dependent variable is of normal distribution, the estimates of parameters will be unbiased and efficient [1 ,2]. If the variance of the dependent variable differs at different measurement points, the unweighted regression remains unbiased but it will not be efficient any more. In such cases the method of weighted least

* L. Eötvös Institute of Geophysics, Budapest

** The term "parameter" is commonly used in mathematical statistics for the constants of a distribution function and of a theoretical regression function or a model. In chemistry "parameter" means the measured (or controlled) variables; parameters of a function have to be distinguished from these latter.

squares is to be applied with the same criterion as that of maximum likelihood if the dependent variable is of normal distribution, leading to efficient estimates [2]. Therefore the criterion to estimate the parameters is:

$$S_y = \sum_i \frac{[y_i - Y(x_i, \hat{\underline{\Theta}})]^2}{\sigma_{y_i}^2} = \sum_i \frac{(\Delta_y)_i^2}{(\sigma_y)_i^2} = \min \tag{1}$$

where

y_i measured value of the dependent variable;
$\hat{Y}_i = Y(x_i, \hat{\underline{\Theta}})$ its estimated value;
x_i measured value of the independent variable;
$(\sigma_y)_i^0$ variance characterizing the accuracy of the measured dependent variable;
$\hat{\underline{\Theta}}$ vector of estimated parameters;
$\Delta_y = y - \hat{Y}$ the residual for the dependent variable.

Often motivated by the form of function $Y(x, \underline{\Theta})$ or upon any other practical consideration a function $F(y)$ of, rather than the directly measured dependent variable is minimized, permitting e.g. to treat an exponential function after logarithmization as linear.

In such cases the right criterion for estimation [2, 3] is:

$$S_F = \sum_i \frac{[F(y_i) - F(\hat{Y}_i)]^2}{(\sigma_{F(y)})_i^2} = \sum_i \frac{[F(y_i) - \hat{\Phi}(x_i)]^2}{(\sigma_{F(y)})_i^2} = \sum_i \frac{(\Delta_F)_i^2}{(\sigma_{F(y)})_i^2} = \min \tag{2}$$

where F is the actual transform;

$\Phi(x)$ the transformed function explicite to x;
$\hat{\Phi}(x) = F(\hat{Y}) = F[Y(x, \hat{\underline{\Theta}})]$ its estimation;
$(\sigma_{F(y)})_i^2$ variance of the transformed dependent variable;
$\Delta_F = F(y) - F(\hat{Y}) = F(y) - \hat{\Phi}(x)$ residual for the transformed dependent variable.

The variance $\sigma_{F(y)}^2$ of the transformed dependent variable derived from the "propagation of error law" [3]:

$$\sigma_{F(y)}^2 \simeq \left(\frac{dF}{dY}\right)^2 \sigma_y^2 \tag{3}$$

Thus the criterion (2) can be written as:

$$\sum_i \frac{[F(y_i) - \Phi(x_i)]^2}{\left(\frac{dF}{dY}\right)_i^2 (\sigma_y)_i^2} = \sum_i \frac{(\Delta_F)_i^2}{\left(\frac{dF}{dY}\right)_i^2 (\sigma_y)_i^2} = \min \tag{4}$$

The weight is seen not to be constant (due to the occurrence of derivatives) even if σ_y^2 is constant.

Often both kinds of weighting (for $\sigma_y^2 \neq$ const and for $\left(\frac{dF}{dY}\right)^2$) are omitted. Minimizing a function other than (1) or (2) also the estimated parameters will differ and that will be inefficient [2, 8].

It will be shown, that provided the independent variable is free of random error, minimization of any transform of the estimated parameters — with right weighting — yields closely identical values for the estimated constants; that is, criteria (1) and (2) are practically identical [2].

Expanding the function $F(y)$ into Taylor series at $\hat{Y}$ and omitting terms higher than first degree:

$$F(y) \simeq F(\hat{Y}) + \left(\frac{dF}{dY}\right)(y - \hat{Y}). \tag{5}$$

This approximation is allowed if second and higher derivatives of F(y) with respect to y are not very high and the residual $(y - \hat{Y})$ from the measurement error is small enough. For measurements of physicochemical type these conditions are usually fulfilled. Rearranging Eq. (5) yields for the residual of the transformed dependent variable:

$$F(y) - F(\hat{Y}) \simeq \left(\frac{dF}{dY}\right)(y - \hat{Y}) \tag{6}$$

or concisely:

$$\Delta_F \simeq \left(\frac{dF}{dY}\right)\Delta_y$$

Substituting into (4) leads to Eq. (1). If neglects are allowable, criteria (1) and (2) are seen fairly to be identical.

Example 1: Determination of vapour pressure function of liquids [8].

Vapour pressure value p_i are read at a temperature t_i supposed to be measured at an extreme precision. (This means that not only $\sigma_t \simeq 0$ in comparison with σ_p but also $\left(\frac{dp}{dt}\right)\sigma_t$ is negligible, see error propagation law; cf. Example 2, expression (g)). The relationship is generally characterized by the Antoine equation:

$$\ln p = a - \frac{b}{C + t}. \tag{a}$$

The criterion for estimation of parameters from Eq. (1):

$$\sum_i \frac{\left[p_i - \exp\left(a - \frac{b}{C + t_i}\right)\right]^2}{\sigma_{p_i}^2} = \min. \tag{b}$$

Taking logarithm of dependent variable p leads to a more convenient form:

$$\sum_i \frac{\left(\ln p_i - a + \frac{b}{C + t_i}\right)^2}{(\sigma_{\ln p})_i^2} = \min \tag{c}$$

where

$$(\sigma_{\ln p})_i^2 = \left(\frac{d \ln p}{dp}\right)_i^2 (\sigma_p)_i^2 = \frac{1}{p_i^2} (\sigma_p)_i^2 \tag{d}$$

That is, (c) yields an estimation criterion corresponding to Eq. (4):

$$\sum_i \frac{\left(\ln p_i - a + \frac{b}{C + t_i}\right)^2}{\left(\frac{\sigma_p}{p}\right)_i^2} = \min . \tag{e}$$

Criterion (e) replaces well the tedious expression (b) inappropriate for direct computation. (e) is often replaced by:

$$\sum_i \left(\ln p_i - a + \frac{b}{C + t_i}\right)^2 = \min \tag{f}$$

This latter is right only for $\frac{\sigma_{p_i}}{p_i} = \text{const}$, that is, if the relative variance accuracy) of pressure measurement is constant.

II.

In cases where the independent variable is also subject to random error, the situation is more complicated.

Measurements not only change Y_i to $y_i = Y_i + \varepsilon_i$ but also the independent variable X_i will be error-laden: $x_i = X_i + \delta_i$. According to Vincze [4] the method of least squares cannot be applied without any consideration. Kendall and Stuart [5] give the following estimation criterion based on the principle of maximum likelihood:

$$\sum_i \frac{[y_i - Y(x_i, \hat{\Theta})]^2}{(\sigma_y)_i^2} + \sum_i \frac{(x_i - \hat{X}_i)^2}{(\sigma_x)_i^2} = \min$$

or

$$\sum_i \frac{(\Delta_y)_i^2}{(\sigma_y)_i^2} + \sum_i \frac{(\Delta_x)_i^2}{(\sigma_x)_i^2} = \min \tag{7}$$

where

x_i measured value of the independent variable $x_i = X_i + \delta_i$;
X_i its true value;
$\hat{X}_i$ its estimated value;
σ_x^2 variance characterizing the measurement accuracy of x;
$\Delta_x = x - \hat{X}$ residual of the independent variable.

The so-called normal equations (obtained by zeroing the partial derivatives of (7) with respect to elements of estimated parameter vector $\hat{\underline{\Theta}}$ and to $\hat{X}$ will not be linear even in the case of a linear $Y(X)$ function. For cases of more complicated functions the normal equation system cannot hope to have an analytic solution.

From different starting points, Guest [2]. Klepikow and Sokolow [7] and Clutton-Brock [6] get the following identical result for the advisable minimization criterion instead of (1):

$$S_y = \sum_i \frac{[y_i - Y(x_i, \hat{\underline{\Theta}})]^2}{(\sigma_y)_i^2 + \left(\frac{dY}{dX}\right)_i^2 (\sigma_x)_i^2} = \sum_i \frac{(\Delta_y)_i^2}{(\sigma_{\Delta_y})_i^2} = \min \tag{8}$$

The term $(\sigma_{\Delta_y})_i^2$ is obtained from the error propagation law:

$$(\sigma_{\Delta_y})_i^2 = \left(\frac{\partial \Delta_y}{\partial y}\right)_i^2 (\sigma_y)_i^2 + \left(\frac{\partial \Delta_y}{\partial x}\right)_i^2 (\sigma_x)_i^2 = (\sigma_y)_i^2 + \left(\frac{dY}{dx}\right)_i^2 (\sigma_x)_i^2 \tag{9}$$

Function $Y = Y(X, \underline{\Theta})$ being not exactly known, computation involves approximative (iterative) values of the derivatives:

$$\left(\frac{dY}{dX}\right) \simeq \left(\frac{d\hat{Y}}{dx}\right) \tag{10}$$

Estimates obtained in this way are generally not unbiased [2], except the case of linear function $Y(X)$, but this criterion is consistent with our approach and expectation, as seen in Fig. 1. Effect of an error δ_i in measuring the independent variable will less affect the dependent variable on the gently sloping part of the curve than on the steeper part. The error propagation law implies the uncertainty of the dependent variable to have two constituents: the measurement error ε_i of the dependent variable itself and the consequence of the measurement error δ_i in the independent variable. Points affected by a greater uncertainty from these two effects combined have to be assigned lower values.

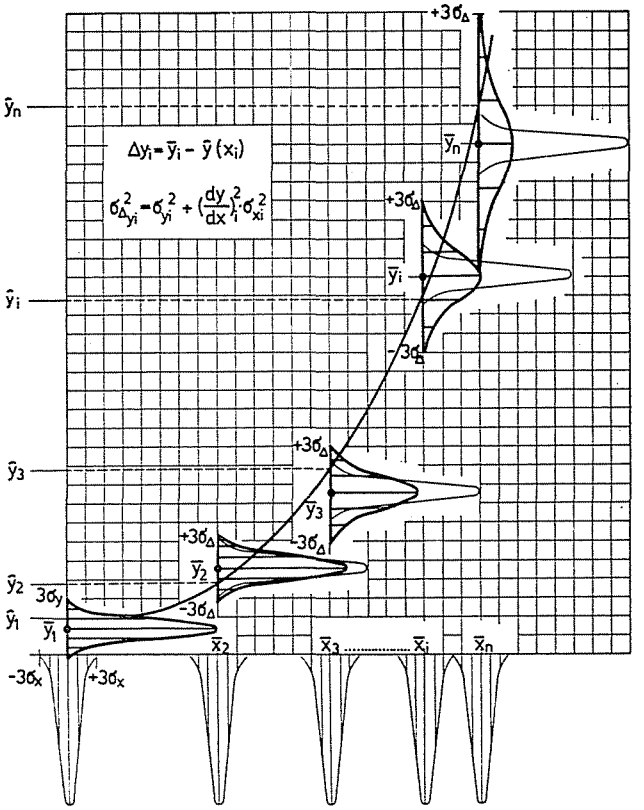

Fig. 1.

Let us consider now whether in cases of an error-laden independent variable (stochastic variable), it is true that appropriate minimization of SSQ of whichever function (transform) of the dependent variable leads to the same estimated parameters (constants).

Let the relationship between the two variables

$$Y = Y(X, \underline{\Theta}) \tag{11}$$

with expression (8) as the right criterion for parameter estimation.

The regression is to be made for some transform $F(Y)$ of Y. Two cases will be distinguished:

1. The transform F depends on the independent variable X only indirectly, through Y : $F = F[Y(X)]$; that is, the formula of transformation does not contain any operation with X. Such kinds of transformation are e.g. $F = \ln Y$ or $F = 1/Y$.

2. The transform F contains directly X : $F = F[Y(X), X]$; that is, the formula of transformation contains operations with X. Such kinds of transformation are e.g. $F = Y \cdot X$ or $F = Y/X$.

Case 1:

$$F = F[Y(X)] = \hat{\Phi}(X) \tag{12}$$

Here $F[Y(X)]$ is the transformation instruction, $\Phi(X)$ means the function explicite to X. The residual for the new variable is:

$$\Delta_F = F(y_i) - \hat{\Phi}(x_i) \tag{13}$$

with the variance:

$$(\sigma_{\Delta_F})_i^2 = \left(\frac{\partial \Delta_F}{\partial Y}\right)_i^2 (\sigma_y)_i^2 + \left(\frac{\partial \Delta_F}{\partial X}\right)_i^2 (\sigma_x)_i^2 \tag{14}$$

Since F depends only on Y and Φ only on X:

$$\begin{aligned}(\sigma_{\Delta_F})_i^2 = \left(\frac{dF}{dY}\right)_i^2 (\sigma_y)_i^2 + \left(\frac{d\Phi}{dX}\right)_i^2 (\sigma_x)_i^2 = \left(\frac{dF}{dY}\right)_i^2 (\sigma_y)_i^2 + \left[\left(\frac{dF}{dY}\right)\left(\frac{d\Phi}{dF}\right)\right. \\ \left.\left(\frac{dY}{dX}\right)\right]_i^2 (\sigma_x)_i^2 = \left(\frac{dF}{dY}\right)_i^2 \left[(\sigma_y)_i^2 + \left(\frac{dY}{dX}\right)_i^2 (\sigma_x)_i^2\right] = \left(\frac{dF}{dY}\right)_i^2 (\sigma_{\Delta_y})_i^2\end{aligned} \tag{15}$$

since $d\Phi/dF = 1$.

This expression is analogous with (3) except that σ_y is replaced by σ_{Δ_y} since x is also subject to error. Thus the right criterion for estimation is:

$$S_F = \sum_i \frac{[F(y_i) - \hat{\Phi}(x_i)]^2}{\left(\frac{dF}{dY}\right)_i^2 \left[(\sigma_y)_i^2 + \left(\frac{dY}{dX}\right)_i^2 (\sigma_x)_i^2\right]} = \min \tag{16}$$

Expression (6) shows criterion (16) to be identical with (8).

Thus if the new dependent variable given by transformation does not explicitely contain the stochastic idependent variable, then it is true that minimization of whichever properly weighted SSQ leads to the same estimated parameter values.

Example 2. Let us see again the problem of the measurement of vapour pressure. In fact thermometry cannot be stated to be always infinitely precise, thus in regression also the uncertainty of the independent variable has to be taken into consideration:

$$(\sigma_{\Delta_p})_i^2 = (\sigma_p)_i^2 + \left(\frac{dp}{dt}\right)_i^2 (\sigma_t)_i^2 \tag{g}$$

Even if the precision of the pressure measurement is constant, the uncertainty of p-values will be greater for higher values of pressure because of the effect

of thermometry uncertainty. The criterion for estimation analogous to Eq. (1) correcponding to Eq. (8) is:

$$\sum_i \frac{\left[p_i - \exp\left(a - \frac{b}{C + t_i}\right)^2\right]}{(\sigma_p)_i^2 + \left(\frac{dp}{dt}\right)_i^2 (\sigma_t)_i^2} = \min \tag{h}$$

Case 2.

$$F = F(Y, X) = \Phi(X) \tag{17}$$

Here $F(Y, X)$ is the transformation instruction, $\Phi(X)$ denotes function F explicite to X.

Let us expand function $F(x, y)$ into Taylor series at point $(x, \hat{Y})$:

$$\begin{aligned} F(x, y) &\simeq F(x, \hat{Y}) + \left(\frac{\partial F}{\partial x}\right)(x - x) + \left(\frac{\partial F}{\partial Y}\right)(y - \hat{Y}) = \\ &= F(x, \hat{Y}) + \left(\frac{\partial F}{\partial Y}\right)(y - \hat{Y}) = \hat{\Phi}(x) + \left(\frac{\partial F}{\partial Y}\right)(y - \hat{Y}) \end{aligned} \tag{18}$$

Expressing the residual for the transformed dependent variable:

$$\Delta_F = F(x, y) - \hat{\Phi}(x) \simeq \left(\frac{\partial F}{\partial Y}\right)(y - \hat{Y}) \tag{19}$$

or concisely:

$$\Delta_F \simeq \left(\frac{\partial F}{\partial Y}\right)\Delta_y \tag{20}$$

The variance of this residual;

$$(\sigma_{\Delta_F})^2 = \left(\frac{\partial F}{\partial Y}\right)^2 (\sigma_{\Delta_y})^2 \tag{22}$$

It is seen from (21) that criteria (8) and (22) are fairly identical: that is, minimizing any transform of the dependent variable in case of proper weighting practically the same criterion is obtained as that of the regression with the original dependent variable minimized, therefore values of the parameters are approximately identical.

Example 3. Determination of the distribution coefficient of a third component between two liquid phases.

Directly measured data are:

x: concentration of the component in the aqueous phase

y: the same in the organic phase.

The function approximating the relationship of the directly measured data be:

$$\hat{Y} = ax + bx^2 + cx^3 \tag{k}$$

The direct function to be minimized from (8):

$$S_y = \sum_i \frac{(\Delta_y)_i^2}{(\sigma_{\Delta_y})_i^2} = \sum_i \frac{(y_i - ax_i - bx_i^2 - cx_i^3)^2}{(\sigma_y)_i^2 + (a + 2bx_i + 3cx_i^2)^2 (\sigma_x)_i^2} \tag{l}$$

The directly measured data are often processed in form of indirectly calculable distribution coefficient $k(x) = y/x$. Here k is a transform of the dependent variable y, which contains x explicitly, too: $k(y, x) = y/x$ is the transformation instruction. The transformed function is approximated by:

$$\hat{k} = a + b\,x + c\,x^2 \tag{m}$$

The indirect function to be minimized according to (22) is:

$$S_k = \sum_i \frac{(\Delta_k)_i^2}{(\sigma_{\Delta_k})_i^2} = \sum_i \frac{(\Delta_k)_i^2}{\left(\frac{\partial k}{\partial Y}\right)_i^2 (\sigma_{\Delta_y})_i^2} = \sum_i \frac{[k(x_i, y_i) - \hat{k}(x_i)]^2}{\left(\frac{\partial k}{\partial Y}\right)_i^2 \left[(\sigma_y)_i^2 + \left(\frac{dY}{dx}\right)_i^2 (\sigma_x)_i^2\right]} =$$

$$= \sum_i \frac{(k_i - a - bx_i - cx_i^2)^2}{\left(\frac{1}{x_i}\right)^2 \left[(\sigma_y)_i^2 + (a + 2bx_i - 3cx_i^2)^2 (\sigma_x)_i^2\right]} =$$

$$= \sum_i \frac{x_i^2 (k_i - a - bx_i - cx_i^2)^2}{(\sigma_y)_i^2 + (a + 2bx_i + 3cx_i^2)^2 (\sigma_x)_i^2}$$

Since the transformation is very simple, (n) yields the expression (l) by purely algebraic manipulations, that is:

$$S_k = S_y \tag{o}$$

Naturally parameters $\hat{\underline{\Theta}} = (a, b, c)$ in denominators of criteria (l) and (n) are unknown a priori, the calculation is to be made by substituting iterative values. Identity (l) = (n) is only valid if the trial and error procedure is continued until appropriate precision.

Summary

It was shown that for an arbitrary transform of the dependent variable the estimated parameters obtained by regression analysis — with appropriate weighting — are practically identical.

References

1. DRAPER, N. R.—SMITH, H.: Applied Regression Analysis. J. Wiley and Sons, N. Y. 1966.
2. GUEST, P. G.: Numerical Methods of Curve Fitting. Cambridge, Univ. Press, 1961.
3. DEMING, W. E.: Statistical Adjustment of Data. Dover Publ. Inc. N. Y. 1964.
4. VINCZE, I.: Mathematical Statistics with Engineering Applications (in Hungarian). Műszaki Könyvkiadó, Budapest 1968.
5. KENDALL, M. G.—STUART, A.: The Advanced Theory of Statistics. Vol. 2.; Charles Griffin and Co. Ltd. London 1967.
6. CLUTTON-BROCK, M.: Technometrics **9**, 261 (1967).
7. KLEPIKOW, N. P.—SOKOLOW, S. N.: Analysis and Design of Experiments by Method of Maximum Likelihood (in Russian), Nauka, Moscow, 1964.
8. TETTAMANTI, K.—STOMFAI, R.—MANCZINGER, J.: Treatment of vapour pressure data of organic compounds (in Hungarian). Paper presented on Scientific Session of Technical University, Budapest, 1967.

Prof. Dr. Károly TETTAMANTI
Dr. József MANCZINGER
Dr. Sándor KEMÉNY
H-1521 Budapest

Róbert STOMFAI
H-1440 Budapest, POB 35.