XXI IMEKO World Congress "Measurement in Research and Industry"
August 30 – September 4, 2015, Prague, Czech Republic

REGRESSION AND CONFLUENT ALGORITHMS AND MODELS FOR CONSTRUCTION OF FUNCTIONS IN MEASUREMENTS

T. N. Siraya

Concern CSRI Elektropribor, JSC, St. Petersburg, Russia
E-mail: tnsiraya@gmail.com

Abstract – Construction of functions is considered as an important part of measurement, which is essential for its quality. Since the traditional least-squares method is only applicable within the strict regression model of data, other methods are required in practice. In the paper, some proper extensions of regression models are considered, which enable consistent estimates for functions. The confluent estimates based on these models are studied; the accuracy characteristics are primarily considered.

Keywords: measurement, function, calibration, algorithm, fitting

1. INTRODUCTION

Construction of functions based on experimental data is a widespread and important procedure in measurement problems. For instance, we can mention such cases as fitting of calibration curves for measuring instruments, or deduction of equations for indirect measurement methods. In these cases the quality of the empirical function is essential for ensuring the integral quality of measurement. In particular, we should obtain an adequate empirical function, which is sufficiently precise for the practical purposes. Moreover, we should also find reliable estimates of the accuracy characteristics for this empirical function.

The traditional method for fitting of functions is the classical least squares (LS) method [1]. Unfortunately, the LS estimates are optimal only under strict conditions upon errors of data, that is, within a classic regression model. So, when it does not hold in practice, some other methods are required, so-called, confluent methods [2 - 4].

However, a general confluent model is not appropriate for practice either, since neither LS nor other data processing methods can provide statistically consistent estimates for functions. Therefore, the key problem is to construct proper extensions of regression models, which allow obtaining consistent estimates for functions. That is, the proper model should include additional information on data or data errors, which makes it possible to construct consistent estimates.

Confluent estimates should be constructed in accordance with the type of the extended model. However, the most important problem for measurements is to study the accuracy characteristics of these estimates, taking into account both random and systematic components of data errors [5].

In this paper, several ways for expansion of regression models are outlined, and associated groups of estimates are studied. First of all, the accuracy properties of estimates are considered, which may be compared with the corresponding characteristics of classical LS estimates. This way of reasoning is certainly heuristic, but the optimisation approach does not work in this case.

The main criterion for confluent methods deriving is consistency as an asymptotic property of estimates. However, confluent estimates are used in the cases of small or modest volume of data. Therefore, the accuracy characteristics of various confluent estimates should be also compared with the classical LS fitting under practical conditions. For this purpose, the certification scheme for calibration algorithms may be useful as a basis for studying estimates [6]. Some preliminary recommendations based on error estimates may be formulated in this regard.

2. REGRESSION AND CONFLUENT MODELS

It is assumed that a function (in particular, calibration curve) is presented in analytic form:

$$Y = f(X) = f(X, a_1, \dots, a_k), \tag{1}$$

where X and Y are input and output values;
f is p-parameter function of a certain analytical type (selected on the basis of a priori data);
$a_1, \dots, a_p$ are the parameters to be estimated.

The set of experimental data used to construct function (1) may be generally presented as follows:

$$\{ u_{ij}, v_{ij}, \ i=1 \dots n, \ j=1 \dots m_i \}, \tag{2}$$

where u_{ij} are the data obtained by measuring input values X_i, and v_{ij} are the data obtained by measuring output values Y_i.

This paper is essentially concerned with the linear model case, when function (1) is linearly dependent on the parameters $a_1, \dots, a_p$, and it can be presented in form:

$$Y = f(X) = \Sigma a_k g_k(X). \tag{3}$$

For this case, the data model can be presented in matrix form:

$$\boldsymbol{y} = \boldsymbol{H} \boldsymbol{a}, \tag{4}$$

$$\boldsymbol{u} = \boldsymbol{X} + \boldsymbol{\varepsilon}\,, \qquad (5)$$

$$\boldsymbol{v} = \boldsymbol{y} + \boldsymbol{\xi}. \qquad (6)$$

Here (4) is the relation for the true values (non-observable variables), so the vectors look as follows:

$$\boldsymbol{y} = (f(X_i)), \qquad \boldsymbol{a} = (a_k), \qquad (7)$$

and matrix H takes the form:

$$\boldsymbol{H} = (h_{ik}) = (g_k(X_i)). \qquad (8)$$

Therefore, relations (5-6) describe the experimental values (observable variables) of the form:

$$u_{ij} = X_i + \eta_{ij}, \qquad (9)$$

$$v_{ij} = Y_i + \xi_{ij}, \qquad (10)$$

where η_{ij} and ξ_{ij} are the errors of measured X_i and Y_i .

The experimental data and especially errors may be described in terms of various probabilistic (or statistical) models.

The initial model is the classical regression model:

$$\Omega(R_0) = \{ u_{ij} = X_i,\ v_{ij} = f(X_i) + \xi_{ij}\,,\ i=1\ldots n,\ j=1\ldots m_i\}, \quad (11)$$

where input values X_i are known exactly, and output values $Y_i = f(X_i)$ are measured with errors ξ_{ij} .

In statistics, the errors ξ_{ij} are usually assumed to be random values, but in measurement problems, one should take into account the fact that errors ξ_{ij} contain both random ε^y_{ij} and systematic θ_{yi} components:

$$\xi_{ij} = \theta_{yi} + \varepsilon^y_{ij}\ . \qquad (12)$$

It is well known that in the case of regression model $\Omega(R_0)$ of the form (11), with Gaussian distributions of random errors ξ_{ij}, the classical LS method should be used. It ensures the optimal estimates, which are also statistically consistent. Moreover, LS estimates usually remain efficient in presence of systematic errors θ_{yi}.

Nevertheless, the regression models do not always hold in practice, so we should consider the extended models with errors in arguments X_i. Thus, the general confluent model may be presented as follows:

$$\Omega(C_0) = \{u_{ij} = X_i + \eta_{ij},\ v_{ij} = f(X_i) + \xi_{ij},\ ,\ i=1\ldots n,\ j=1\ldots m_i\}. \quad (13)$$

However, under this general model $\Omega(C_0)$, the classical LS estimates turn out statistically inconsistent, that is, LS estimates are systematically biased.

For instance, in the simple, but practically important case of the linear function

$$Y = a + b\,X\,, \qquad (14)$$

the LS estimate of coefficient b has the asymptotic bias of the form

$$\mathrm{B(b)=M(b)} - \mathrm{b}_0 \approx - \mathrm{b}\ \sigma^2_{\mathrm{x}} / (\tau_{\mathrm{n}} + \sigma^2_{\mathrm{x}}\ ), \qquad (15)$$

where parameter τ_n defines the location of the points X_i within the range:

$$\tau_n = \Sigma(X_i - \bar{x})^2/n\,, \qquad (16)$$

and it is also assumed that points X_i are not concentrated around one point:

$$\tau_n \to \tau > 0. \qquad (17)$$

Thus, even in the absence of systematic errors of data, there is a significant systematic bias in the function coefficient.

Unfortunately, under the general confluent model $\Omega(C_0)$, neither classical LS estimates nor any other estimates are statistically consistent. Thus, it is reasonable to construct various confluent models that allow obtaining consistent estimates. However, there are two points to emphasize.

Firstly, these models and confluent methods demand for some additional information on data (and data errors) in order to ensure consistent estimates.

Secondly, these methods are usually developed for some particular cases of application. So they frequently employ the forms of information which are specific to these fields, but are not convenient for measurements at all. Thus, our task is to select the cases and kinds of information that are available for measurement practice.

Therefore, it is necessary to construct a family of confluent models which are proper, or correct, in a sense, that it is possible to determine the consistent estimates.

First of all, a proper confluent model $\Omega\ (C_p)$ is an expansion of the regression model that allows for the errors in arguments X_i . However, it is a restriction of the general confluent model $\Omega(C_0)$ as it should contain some additional information on data or limitations on the parameters of data errors. Then, it must be a relation of the form:

$$\Omega(R_0) \subset \Omega(C_p) \subset \Omega(C_0). \qquad (18)$$

For instance, the formula of systematic bias (15) suggests that we can obtain proper models by specifying the variances of data errors. Really, the model of the form

$$\Omega\{C_p | D(\varepsilon_x) = \sigma^2_x\} =$$

$$= \{ u_{ij} = X_i + \varepsilon_{ij}^x,\ v_{ij} = f(X_i) + \xi_{ij},\ D(\varepsilon_{ij}^x) = \sigma^2_x \} \qquad (19)$$

is a proper confluent model for which the modified LS estimates are consistent.

Another kind of information is associated with the location of points X_i, that is, with designing of the measurement experiment. Generally, the case of an active (designed) experiment is more promising for confluent analysis, as it gives additional information on data, and thus, more efficient estimates.

Further, the main variants of additional information on data, available for measurement problems, are presented, and the corresponding confluent estimates are studied.

3. MAIN GROUPS OF CONFLUENT MODELS AND ESTIMATES

The following groups of proper confluent models may be considered which employ the available additional information on data and provide consistent confluent estimates for functions.

Model A (M_A)

Multiple measurements of input values X_i , $i=1...n,$ and corresponding output values Y_i are realized

$$\Omega_A = \Omega\{C_p \mid u_{ij}, v_{ij}, \ i=1...n, \ j=1...m_i>1\}. \quad (20)$$

This case is rather widespread both in active and passive experiments.

Model B (M_B)

Variances $D(\xi)$ or $D(\eta)$ of data errors are known:

$$\Omega_{BX} = \Omega\{C_p \mid D(\eta)=\sigma^2_x\}, \quad (21)$$

$$\Omega_{BY} = \Omega(C_p \mid D(\xi)=\sigma^2_y). \quad (22)$$

These parameters may be given a priori or estimated in independent experiments.

Model C (M_C)

The ratio of data error variances $\lambda = D(\xi) / D(\eta)$ is known:

$$\Omega_C = \Omega\{C_p \mid \lambda = D(\xi) / D(\eta)\}. \quad (23)$$

Likewise, this ratio may be given a priori (or estimated in independent experiments).

Model D (M_D)

The increasing order of input values X_i is known:

$$\Omega_D = \Omega\{C_p \mid X_1 \le X_2 \le ... \le X_n\}. \quad (24)$$

This order may be given a priori or deduced from the physical conditions of the experiment.

The corresponding confluent estimates may be constructed under the proper confluent models. In a general case, the confluent estimates cannot be presented in explicit form. Thus, some numerical or iterative methods should be used [3, 4]. This also remains valid for general linear models of the form (3). Note that for the case of a simple linear function (14) some confluent estimates may be obtained in explicit form.

The confluent estimates are of rather various forms. Many of them are just modifications of LS estimates, which are improved by using the known parameters of data errors, but others are quite different from LS estimates. The common features of all the confluent estimates are the following:

– these estimates are statistically consistent;

– when the input values X_i are known exactly, they coincide with LS estimates.

For instance, the following groups of consistent confluent estimates for the coefficient b of linear function (14) may be used.

Estimate A (for Model A)

Variance analysis estimates are based on multiple observations at points:

$$b_1 = \pm[(\Sigma_y - (n-1)\, S^2_y)\, /(\Sigma_x - (n-1)\, S^2_x)]^{1/2}, \quad (25)$$

where the sums of data are denoted as follows:

$$\Sigma_x = \Sigma\, (x_i - \bar{x})^2 , \qquad \Sigma_y = \Sigma\, (y_i - \bar{y})^2 . \quad (26)$$

Here S^2_x, S^2_y are the estimates of data variances based on dispersion within the groups of data at points X_i, Y_i ;
x_i, y_i are means of data groups at points X_i, Y_i ;
$\bar{x}$, $\bar{y}$ are the general means.

Estimate B (for Model B)

LS estimates may be improved by using the known data error variances in order to obtain the following estimates:

$$b_2 = \Sigma_{xy}/(\Sigma_x-(n-1)\, \sigma^2_x), \quad (27)$$

$$b_3 = (\Sigma_y-(n-1)\, \sigma^2_y)/ \Sigma_{xy}, \quad (28)$$

where the sums Σ_x , Σ_y of the data are as in (26), and

$$\Sigma_{xy} = \Sigma\, (x_i - \bar{x})(y_i - \bar{y}). \quad (29)$$

Estimate C (for Model C)

If the ratio of data error variances $\lambda = D(\xi) / D(\eta)$ is known, it is possible to use generalized orthogonal regression estimates, which are rather different from LS estimates:

$$b_o = r \pm (r^2 + \lambda)^{1/2}, \qquad r= (\Sigma_y - \lambda\Sigma_x) / 2\Sigma_{xy}. \quad (30)$$

Here, the sums of data are denoted just as in (26), (29).

The orthogonal regression estimates (with $\lambda = 1$) were derived in statistics quite a long time ago, but they were not popular as the case of $\lambda = 1$ was rather uncommon. But the introduction of the generalized orthogonal regression, with arbitrary value of scale parameter λ, makes it more promising for practice.

This estimate may be also obtained as an optimal one, from the condition of a minimum weighted sum of squared distances of experimental points from the straight line (14):

$$min\ Q_{ort} = \Sigma\, [y_i -(a + bx_i)]^2/ (\lambda+b^2). \quad (31)$$

In the case $\lambda = 1$, the distances are determined in the orthogonal direction to the line, but in general case the scale parameter λ is taken into account.

Estimate D (for Model D)

When the increasing order of input values X_i is known, the homographic, or linear fractional estimates, may be used. They are presented as follows:

$$b_w = \Sigma\, w_i\, y_i \,/\, \Sigma\, w_i\, x_i , \quad (32)$$

where w_i are constant weights, and $\Sigma w_i = 0$.

Generally, the form of these estimates is quite different from LS estimates, and the choice of weights may be rather broad. But if we try to determine optimal weights (in order to minimize the variance of b_w), we obtain:

$$w_i = c\, (x_i - \bar{x}); \quad (33)$$

so it is just the approximation of LS estimate.

The similar estimates corresponding to more general (non-linear) classes of functions cannot be presented in explicit form; thus, some numerical or iterative methods should be used. But the ways to obtain confluent estimates under the proper confluent models stated above remain valid.

There are also some other kinds of confluent estimates [2-4], for instance, instrumental variables methods. But they need some kinds of extra information, which are less suitable for measurement problems.

4. COMPARISON OF CONFLUENT METHODS

First of all, we should consider the accuracy characteristics of the confluent estimates, taking into account both random and systematic components of data errors.

The main accuracy characteristics used to compare confluent estimates are bias and variance.

There are two main cases to study:

a) asymptotic case, when the number of points is large;

b) practical case, when the number of points is small or modest.

The asymptotic properties of confluent methods are studied most frequently, and the main criterion is statistical consistency, that is, bias of estimate should tend to zero. Certainly, the variance (and standard deviation) remain significant. Sometimes, the integral characteristic of estimate is used, which is the second moment of the estimate relative to the true value of parameter.

For instance, the asymptotic biases and variances of several confluent estimates given above are presented in Table 1.

Table 1. Asymptotic characteristics of confluent estimates for coefficient *b* of linear function

A priori information		Estimate	Bias $B(b)$	Variance
Known parameters or estimates	$\lambda = \sigma_y^2 / \sigma_x^2$ (estimate)	b_0	$\frac{2\sigma_y^2 + b^2 \sigma_x^2}{b \sum_x}$	$\frac{\sigma_y^2 + b^2 \sigma_x^2}{\sum_x}$
	σ_x^2 or estimate	b_2	$2\, b\, \sigma^2_x / \Sigma_x$	
	σ_y^2 or estimate	b_3	$(b^2 - \lambda)\, b\, \sigma^2_x / b^2\, \Sigma_x$	
Multiple observations	$S_x^2,\ S_y^2$	b_1	$\left(1 - \frac{\lambda}{b^2}\right) \frac{b\sigma_x^2}{2\sum_x}$	$\frac{\sigma_y^2 + b^2 \sigma_x^2}{\sum_x}$
	S_x^2	b_2	$B(b_2)$	
	S_y^2	b_3	$B(b_3)$	
	$\lambda^* = S_y^2 / S_x^2$	b_0	$B(b_0)$	
Ordering of true values	order: $X_1 \le \ldots \le X_n$	b_w	$\frac{b\sigma_x^2 W^2}{W(X)}$	$\frac{\sigma_y^2 + b^2 \sigma_x^2}{W^2(X)} W^2$
No extra information		LS estimate $b_{LS} = \Sigma_{xy} / \Sigma_x$	$-\frac{b\sigma_x^2 (n-3)}{\sum_x}$	$\frac{\sigma_y^2 + b^2 \sigma_x^2}{\sum_x}$

The listed confluent estimates have slightly different biases, but they all are consistent. The asymptotic variances of all these estimates are approximately the same as the variance of LS estimate, that is,

$$\sigma^2(b) = (\sigma^2_y + b^2\ \sigma^2_x) / \Sigma_x\ . \qquad (34)$$

The homographic, or linear fractional estimates, are rather specific (different from the other estimates). The corresponding sums in Table 1 are denoted as follows:

$$W^2 = \Sigma\, w_i^{\,2}\ , \quad W(X) = \Sigma\, w_i\, x_i\ . \qquad (35)$$

Since in practice confluent estimates should be used in the cases of small (or modest) volumes of data, they must be also studied in these cases.

The certification scheme for calibration algorithms may be used for this aim [6]. According to this scheme, the main characteristics of the confluent estimates should be evaluated for the typical models of data. The main accuracy characteristics are bias and variance, and the typical models of data are just the main confluent models stated above.

The characteristics of the confluent algorithms may be evaluated either by analytical methods or by computer simulation. The first way gives asymptotic characteristics, which is important for obtaining consistent estimates. But the second way is also essential for the practical use of estimates.

Simulation of data and fitting of functions by using confluent estimates show that in particular cases (for small volumes of data) classical LS estimates turn out to be more accurate than the corresponding confluent estimates. This depends on several factors, including the following:

- the quality of additional information used in confluent estimates;
- the location of the input values X_i within the range.

This research should be continued in order to provide practical recommendations. The study of homographic estimates seems to be very promising.

5. CONCLUSIONS

Since the traditional LS method only gives optimal estimates under a rigid regression model, other methods are required in practice.

In this paper, several ways for expansion of regression models are considered, which allow obtaining statistically consistent estimates for functions. The related groups of confluent estimates are studied, and the estimates for the linear case are presented in explicit form.

The accuracy properties of confluent estimates are considered, which may be compared with the corresponding characteristics of the classical LS fitting.

Since the main criterion for confluent methods is statistical consistency, confluent estimates should be also studied on modest volumes of data. For this aim, the certification scheme for calibration algorithms may be used.

ACKNOWLEDGMENTS

This work is supported by the Russian Foundation for Basic Research, project no. 13-08-00778.

REFERENCES

[1] N. R. Draper, and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, New York, 1981.

[2] M. G. Kendall, A. Stuart, *The Advanced Theory of Statistics, v. 2: Inference and relationship*, Charles Griffin & Co, London, 1966.

[3] E. Z. Demidenko*, Linear and Nonlinear Regressions,* Finance & Statistics, Moscow, 1981 (in Russian).

[4] I. Vuchkov, L. Boyadjieva, and E. Solakov, *Applied Linear Regression Analysis,* Finance & Statistics, Moscow, 1985 (in Russian).

[5] V. A. Granovsky, and T. N. Siraya, *Methods for Data Processing in Measurements,* Energoatomizdat, Leningrad, 1990 (in Russian).

[6] T. N. Siraya, "Certification of algorithms for constructing calibration curves of measuring instruments", *Advances in Mathematical and Computational Tools in Metrology and Testing X* (vol.10), Series on Advances in Mathematics for Applied Sciences, vol. 86, World Scientific, Singapore, 2015, pp 367-374.