Application of the experimental design algorithms to the construction of estimated subsets of minimum covariance determinant method

I O Usowa, E A Khailenko

1Novosibirsk State Technical University, 20, Karla Markska ave., Novosibirsk, 630073, Russia

E-mail: kemirinaou@gmail.com, xajlenko@corp.nstu.ru

Abstract. The problem of stable estimation of errors in variables models is considered in this paper. The basic concepts of experimental design methods, orthogonal regression method, the minimum covariance determinant method are investigated. Experiment planning methods use to form estimated subsets in order to achieve more accurate estimation results. Designs of experiments will allow one to receive the most informative points on which estimation will be made. Research, displaying change in the condition of the D-optimal design when applying different spectra of the design and in the conditions of various kinds of emissions, has been conducted in this work. According to the research it is found that the most accurate estimates are obtained when applying the methods of experiment planning.

1. Introduction

Regression analysis involves many different approaches to modeling and analyzing the relationship between dependent and independent variables. There is no doubt that the least squares method (OLS) is one of the most used among approaches currently available, despite the fact that it is very sensitive to anomalous observations (outliers), which are found in the source data [7]. Besides, the least squares method gives inconsistent estimates for errors in variables models (EIV), which are quite common in practice. Therefore, the need to use special methods for estimating the parameters of structural dependencies appears which will be stable to various kinds of emissions. The use of robust estimation methods is relevant for solving various statistical problems because even a small amount of emissions can lead to estimates that are completely inconsistent with the initial data.

An important fact is the informativeness of the points on which measurements are made. Discarding the emissions, you can ignore the points with important information. Subsequently this will affect the quality of the parameter estimates. The methods of planning an experiment apply in order to control this situation.

In this paper, we will consider an algorithm for synthesizing a close to optimal plan and applying an algorithm to the minimum covariance determinant method.
2. Problem formulation
In this paper we consider multivariate errors in variables models given by [3, 4]:

\[ Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \]

(1)

where \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \) - vector of unknown parameters of this model; \( X = (x_1, x_2, \ldots, x_p) \) is multivariate observable vector; \( x_j, j = 1, p \) and \( Y \) - factor vector and response variable vector respectively, having the dimension \( n \); \( p \) is quantity of unknown parameters without a free member; \( n \) - number of observations.

The data \( x \) and \( y \) are components of the vectors \( x_{ij} \) and \( Y \), respectively. It is represented as noisy values of true variables \( x_{ij}^* \) and \( Y_i^* \):

\[
\begin{align*}
x &= x_{ij}^* + \delta_i, \\
y &= y_i^* + \epsilon_i,
\end{align*}
\]

(2)

where \( \delta_i \) and \( \epsilon_i \) - the vectors of measurement errors independent and identically distributed, that is, \( \delta_i \sim N(0, \sigma_\delta^2) \), \( \epsilon_i \sim N(0, \sigma_\epsilon^2) \), \( i = 1, n \). Measurements will be carried out according to the normalized optimal designs \( \xi = \left\{ x_1, x_2, \ldots, x_n \right\} \), where \( \omega_i = n_i/n, \ i = 1, \ldots, n \).

The problem is to most accurately estimate the unknown parameters \( \beta \) of equation (1).

3. Parameter estimation methods
There are a large number of methods for estimating the parameters of regression models. Each of them gives the best estimates under certain conditions. The orthogonal regression (OR) method is often used when errors occur in explanatory variables. It consists in minimizing the sum of squares of orthogonal distances. However, OR is quite sensitive to outliers. Therefore, the need to use methods that are resistant to anomalous observations appears, such as, for example, the method of the smallest determinant of the covariance matrix.

The concept of the minimum covariance determinant method is described in the articles [11]. Its objective is to find \( h \) observations (out of \( n \)) whose covariance matrix has the lowest determinant. The MCD method is based on repeated repetition of the C-step [9], where the estimated subset \( \hat{h} \) includes observations with the minimum distance:

\[ d_i = \sqrt{(x_i - T_i)^T S_i^{-1} (x_i - T_i)}, \ i = 1, n, \]

(3)

where \( T_i = \frac{1}{h \tilde{h}_i} \sum_{i \in H_i} x_{ij} \) and \( S_i = \frac{1}{h \tilde{h}_i} \sum_{i \in H_i} (x_{ij} - T_i)(x_{ij} - T_i)^T \).

After all the steps of the MCD method have been performed and the estimated subset has been obtained, the parameters are evaluated by any classical method. In our work, all parameter estimates were obtained by the OR method.

4. Experiment design
Experiment planning methods can be used to form estimated subsets in order to achieve more accurate estimation results. Designs of experiments will allow one to receive the most informative points on which estimation will be made.
One of the most famous and frequently used criteria is the criterion of D-optimality. A design \( \xi^* \) is called D-optimal if it maximizes the determinant of the information matrix \( M(\xi) \):

\[
\xi^* = \text{Arg} \max [M(\xi)]
\]

or minimizes the determinant of the dispersion matrix \( D(\xi) \):

\[
\xi^* = \text{Arg} \min [D(\xi)].
\]

The information matrix \( M(\xi) \) is determined by [12]:

\[
M(\xi) = \sum_{i=1}^{n} \frac{\omega_i}{\sigma^2_i(x_i, \theta)} \left( \frac{\partial \hat{m}(x_i, \theta)}{\partial \theta} \right)^T \left( \frac{\partial \hat{m}(x_i, \theta)}{\partial \theta} \right),
\]

where \( \sigma^2_i(x_i, \theta) = \left( \frac{1}{\frac{\partial \hat{m}(x_i, \theta)}{\partial x_i}} \right)^2 \sum_{\theta'} \left( \frac{\partial \hat{m}(x_i, \theta)}{\partial x_i} \right)^T \).

It is known that the dispersion ellipsoid of parameters estimation of the D-optimal plan has a minimal volume.

As a condition for D-optimality, an expression of the form will be used:

\[
d_{\text{Dopt}}(x, \xi^*) = f^T(x)M^{-1}(\xi^*)f(x) = (p+1)\sigma(x, \theta).
\]

5. Algorithm for obtaining a design close to the optimal

In this paper, we propose to use an algorithm for obtaining a design that is close to optimal. This will allow selecting in the estimated subsets points which will be the most informative and as close as possible to the true values.

The algorithm for constructing a plan is as follows:

1. First, the MCD method is run until convergence to obtain an estimated subset of size \( h \).
2. In the resulting estimated subset \( n_1 \) points are left. It corresponds to the minimum distances defined by the formula (3). The number of points that will be included in the estimated subset is determined by:

\[
n_i = \eta h, \ \eta \in [0,1].
\]

3. The points gather to the estimated subset from the remaining \( n - n_i \) observations as follows: the value is set \( k = n_i \). Let us assume that \( \xi_{i_k} \) is design containing \( k \) points, \( \xi_{i_0} \) - design containing \( n - k \) points.

The following is repeated until \( k < h \):

- The search of the most optimal point is made:

\[
x^* = \text{Arg} \max_{x \in \xi_0} \phi(x, \xi^*),
\]

where \( x^* \) - point from design \( \xi_{i_k} \), \( \phi(x, \xi) = f^T(x)M^{-1}(\xi)f(x) \) - functional of the D-optimality criterion.

- The found optimal point is added to the design \( \xi_{i_{k+1}} = \left( 1 - \frac{1}{k} \right) \xi_{i_k} + \frac{1}{k} \xi^* \).

- Value \( k \) increases by 1.
6. The results of the investigation of experimental design

For studies, a linear regression model with errors in the explanatory variables was chosen. It has the form:

\[ Y = \beta_0 + \beta_1 x_i + \varepsilon, \quad x_j = x_j^* + \delta. \]  

(10)

The equilibrium two-point plan of the form \( \xi_1^* = \begin{bmatrix} -1 \\ 0.5 \\ 0.5 \end{bmatrix} \) will be D-optimal for this model. The values of input factors were chosen according to the plan \( \xi_1 \). True Value Parameter Vector \( \beta_{\text{vec}} = (5, 6)^T \), number of observations \( n = 200 \), quantity of regressors \( p = 1 \). The measurement errors of factor \( x_j - \delta \), and response variable \( Y - \varepsilon \) have an identically distributed with zero expectation function and variance \( \sigma_\delta^2 = \sigma_\varepsilon^2 = 0.01 \). When conducting research in cases of anomalous observations (emissions) in the sample, the following conditions were applied. The measurement errors \( \delta \) and \( \varepsilon \) are independent and law distributed:

\[ F(x) = (1 - \mu) F_i(x, m_i, \sigma_i) + \mu F_2(x, m_2, \sigma_2), \]  

(11)

where \( F_i(x, m_i, \sigma_i) \) – the normal distribution function with expected value \( m_i \) and variance \( \sigma_i^2 \); \( \mu \) – proportion of emissions that vary between \([0,1]\). Variance is \( \sigma_\delta^2 \leq \sigma_\varepsilon^2 \) in all computational experiments; variance \( \sigma_\varepsilon^2 \) is selected depending on the experimental conditions, and \( \sigma_\delta^2 = \sigma_\varepsilon^2 = 0.01 \); \( \sigma_\varepsilon^2 = 0.5 \), \( \sigma_\delta^2 = 0.5 \) expected value \( m_i = 0 \). The number of computational experiments was 500. As an indicator of accuracy, the expression was used:

\[ \Psi = \sum_{i=0}^{p} \left| \frac{\beta_{\text{vec}} - \beta_i}{\beta_{\text{vec}}} \right| \cdot 100\%. \]  

(12)

Figures 1-3 present the results of the dependence of the accuracy indicator on the size of the estimated subset and on the proportion of points included in the estimated subset for the three types of emissions:

- vertical, emissions are present in observation errors in response;
- horizontal, emissions are present only in errors in the explanatory variables;
- mixed, outliers are present in both the response and explanatory variables.

Figure 1. Dependence of the estimation accuracy \( \Psi \) on the size of the estimated subset and on the share \( \eta \), if there are vertical emissions in the sample.
Figure 2. Dependence of the estimation accuracy $\psi$ on the size of the estimated subset and on the share $\eta$, if there are horizontal emissions in the sample.

Figure 3. Dependence of the estimation accuracy $\psi$ on the size of the estimated subset and on the share $\eta$, if there are mixed emissions in the sample.

In Figures 1 and 2, you can see the greater the proportion of points remaining in the estimated subset by the MCD method, the lower the accuracy. From Figures 1 and 3, it can be seen that the larger the size of the estimated subset, the worse the accuracy of parameter estimation. For each fraction of points $\eta$, there are different optimal sizes of the estimated subsets, but it can be seen that the greatest number of the best precisions accumulate in the neighborhood of $h = 140$. Also from the figures one can see that the proportion of points at which the highest accuracy is observed at almost all sizes of the estimated subset is $\eta = 0.5$.

Let us consider the accuracy of estimating the parameters of the OR, MCD methods and the MCD method using experimental design under approximately optimal conditions: the size of the estimated subset for the MCD method $h = 140$, is the proportion of points is $\eta = 0.5$. The results are presented in Table 1.
Table 1. The condition of the D-optimality design.

|                | True value | OR   | MCD  | MCD with planning |
|----------------|------------|------|------|-------------------|
| $\beta_0$      | 5          | 5.032| 5.032| 5.032             |
| $\beta_1$      | 6          | 5.858| 5.889| 5.979             |
| $\psi$         | 6.26E-02   | 3.83E-02| 5.58E-03 |

In table 1 you can see that the estimates are the most accurate when applying the MCD method with planning. Estimates are slightly worse with the usual MCD method. The orthogonal regression method showed the lowest accuracy.

7. Conclusion
In conclusion, this paper we considered the orthogonal regression method, the minimum covariance determinant method, and methods for planning an experiment. All the considered algorithms were implemented in the Python programming language. Experimental planning methods were incorporated into the MCD method and studies were carried out.

According to the results of the conducted research, it was concluded that, when influencing the estimating sample of errors from the vector of factors, the accuracy of parameter estimation changes as follows. The larger the size of the estimated subset, the worse the accuracy. In the greatest cases, the highest accuracy was obtained at $\eta = 140$. When errors are influenced by the response vector, the accuracy of parameter estimation improves with decreasing $\eta$. In the greatest cases, the highest accuracy was obtained with $\eta = 0.5$.

When the most optimal parameters were found, the estimates were calculated using the OR, MCD and MCD methods using the algorithm for constructing a near-optimal experiment plan. It was noted that the OR method showed the worst estimate, the accuracy rating of the MCD method was better, and with the application of experimental design the estimates were the most accurate.

Acknowledgments
This research has been supported by the Ministry of Education and Science of the Russian Federation as part of the state task (project No 2.7996.2017/8.9).

References
[1] Ayvazyan S A, Enyukov I S and Meshalkin L D 1985 *Applied statistics* (Moscow: Finansy i statistica) p 488
[2] Borovkov A A 1984 *Mathematical statistics. Parameter estimation, hypothesis testing: textbook* (Moscow: Nauka) p 472
[3] Greshilov A A, Stakun V A and Stakun A A 1997 *Mathematical methods for prediction construction: scientific publication* (Moscow: Radio i svyaz) p 112
[4] Kendall M and Stuart A 1973 *Statistical findings and connection*, ed A N Kolmogorov, (Moscow: Nauka) p 899
[5] Kramer G 1975 *Mathematical methods of statistics*, ed A N Kolmogorov, (Moscow: Mir) p 648
[6] Hampel F, Ronchetti E, Rausseu E and Stael V 1989 *Robustness in statistics. A hike based on influence functions* (Moscow: Mir) p 512
[7] Croux C, Fekri M and Ruiz-Gazen A *Fast and robust estimation of the multivariate errors in variables model* Retrieved from: https://link.springer.com/article/10.1007/s11749-009-0155-9
[8] Hao Han and Wei Zhu *RCR: Robust Compound Regression for Robust Estimation of Errors-in-Variables Model* Retrieved from: [https://arxiv.org/ftp/arxiv/papers/1508/1508.02925.pdf](https://arxiv.org/ftp/arxiv/papers/1508/1508.02925.pdf).

[9] Hao Han and Wei Zhu, *Least Sine Squares and Robust Compound Regression Analysis* Retrieved from: [https://ir.stonybrook.edu/spui/bitstream/11401/71254/1/Han_grad.sunysb_0771E_10819.pdf](https://ir.stonybrook.edu/spui/bitstream/11401/71254/1/Han_grad.sunysb_0771E_10819.pdf).

[10] Rousseeuw Peter J and Van Driessen K *A Fast Algorithm for the Minimum Covariance Determinant* Retrieved from: [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.5870&rep=rep1&type=pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.5870&rep=rep1&type=pdf).

[11] Konstantinou Maria and Dette Holger *Locally optimal designs for errors-in-variables models* Retrieved from: [http://www.ruhr-uni-bochum.de/imperia/md/content/mathematik3/publications/mem-unblinded.pdf](http://www.ruhr-uni-bochum.de/imperia/md/content/mathematik3/publications/mem-unblinded.pdf).