UDK 519.22

T. Ya. Yeleyko, O. A. Yarova

THE MIXTURE OF MULTIPLE REGRESSION EQUATIONS: OPEN PROBLEMS

T. Ya. Yeleyko, O. A. Yarova. *The mixture of multiple regression equations: open problems*, Mat. Stud. **63** (2025), 221–224.

In this article multiple regression equations are considered. The study is based on a sample that is influenced by the external environment. This external environment is represented in the form of factors that influence the main sample. The sample is divided into parts and a multiple regression equation is constructed for each part. We construct a mixture of regression equations. There are posed open problems concerning determinination of the coefficients of mixture of nonlinear regression equations via lasso, ridge and elastic regression estimators.

An external environment explaining some uncertainty has an important impact on a regression equation. Therefore, it should be taken into account when analyzing the sample. Let the external environment be given by events $A_1, A_2, \ldots, A_n$, which form a complete group of pairwise incompatible events. Recently, Mayboroda ([1–3]), Miroshnychenko ([4]), Grün ([5]) with their co-authors investigated mixtures of distributions, as linear, so nonlinear. In particular, there was described the dependencies between the observed variables by mixture of nonlinear regression models with unknown regression parameters and error terms distributions different for different components. It was suggested that the mixing probabilities (concentrations of the components in the mixture) vary from observation to observation. The authors considered estimators for quantiles of error terms distribution via weighted empirical distribution functions of the regression models residuals.

Let us consider sample $x_1, x_2, \ldots, x_n$ and it is known that the external environment is described by events $A_1, A_2, \ldots, A_n$, which form a sample space of mutually exclusive outcomes. We highlight statistical data that correspond to the occurrence of each of these events $x(A_1), \ldots, x(A_n)$ and write the regression equation $g(x(A_1))$. If the sample is large enough, then we can write the regression equations $g(x(A_1)), g(x(A_2)), \ldots, g(x(A_n))$. Assuming that $\frac{n(A_i)}{n} \to p_i$ as $n \to \infty$, we obtain the mixture of regression

$$g(x) = p_1 g(x(A_1)) + p_2 g(x(A_2)) + \ldots + p_n g(x(A_n)).$$

Next, consider in more detail the construction of a mixture based on a sample by analogy to [6].

Consider a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and the sample $Y = (y_1, y_2, \ldots, y_n)$ in this space, where $y_1, y_2, \ldots, y_n$ are independent identically distributed random variables. Let these random variables depend on the factors $X = (X_1, X_2, \ldots, X_m)$. Given this, each factor can

2020 *Mathematics Subject Classification:* 62J05, 62B15.
Keywords: multiple regression; mixture; regression statistic; regression coefficients.
doi:10.30970/ms.63.2.221-224

be represented as a sample of independent, identically distributed random variables. These factors describe the impact of the external environment on the sample. But not all factors have a large impact on each of the sample elements. Therefore, there is a need to "sift" the sample, that is, divide it into several parts depending on the impact of the factors.

Let divide the sample into parts $Y_1 = (y_{11}, y_{12}, \ldots, y_{1n_1})$, $Y_2 = (y_{21}, y_{22}, \ldots, y_{2n_2}), \ldots,$ $Y_k = (y_{k1}, y_{k2}, \ldots, y_{kn_k})$, where $n_1 + n_2 + \ldots + n_k = n$. Therefore, we obtain k samples and initial sample can be rewrite $Y = (Y_1, Y_2, \ldots, Y_k)$. Assume, Y_1 depend on the factors $x_{11}, \ldots, x_{1s_1}$, Y_2 depend on the factors $x_{21}, \ldots, x_{2s_2}$ and Y_k depend on the factors $x_{k1}, \ldots, x_{ks_k}$, where $x_{ij} \in X, i \in \{1, \ldots, k\}, j \in \{1, \ldots, s_k\}, s_k \leq m$. In this case, we obtain the next k regression equation

$$Y_1 = b_{10} + b_{11}x_{11} + b_{12}x_{12} + \ldots + b_{1s_1}x_{1s_1} + \varepsilon_1,$$
$$\ldots$$
$$Y_k = b_{k0} + b_{k1}x_{k1} + b_{k2}x_{k2} + \ldots + b_{ks_k}x_{ks_k} + \varepsilon_k.$$

Here ε_i is normal distributed random variables with mean $\mathrm{E}\,\varepsilon_i = 0$ and variance $\mathrm{Var}(\varepsilon_i) = 1$.

Besides, we can construct a multiple regression equation for initial sample and factors

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m + \varepsilon.$$

Then, we can submit the last regression equation as a mixture of previous regression equations

$$Y = p_1Y_1 + p_2Y_2 + \ldots + p_kY_k + \varepsilon.$$

Consider p_1, which is the probability that the random variables from the sample Y_1 depend on $x_{11}, \ldots, x_{1s_1}$. Thus, $p_1 \to \frac{n_1}{n}$ as $n \to \infty$. Similarly for others, $p_i \to \frac{n_i}{n}$. So, the mixture of multiple regression equations has the next representation

$$Y = \frac{n_1}{n}Y_1 + \frac{n_2}{n}Y_2 + \ldots + \frac{n_k}{n}Y_k + \varepsilon.$$

Let's substitute the regression equations in mixture

$$Y = p_1(b_{10} + b_{11}x_{11} + b_{12}x_{12} + \ldots + b_{1s_1}x_{1s_1} + \varepsilon_1) + \ldots$$
$$\ldots + p_k(b_{k0} + b_{k1}x_{k1} + b_{k2}x_{k2} + \ldots + b_{ks_k}x_{ks_k} + \varepsilon_k).$$

Denote $b_0 = p_1b_{10} + p_2b_{20} + \ldots + p_kb_0$ and $\varepsilon = p_1\varepsilon_1 + p_2\varepsilon_2 + \ldots + p_k\varepsilon_k$. Then, rearrange the factor elements and obtain the multiple regression equations

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m + \varepsilon.$$

As the fact, we obtain such an elementary assertion.

Proposition 1. *Let the sample $Y = (Y_1, Y_2, \ldots, Y_n)$ are depend on $X = (X_1, X_2, \ldots, X_m)$ and $Y_i = b_{i0} + b_{i1}x_{i1} + b_{i2}x_{i2} + \ldots + b_{is_i}x_{is_i} + \varepsilon_i$, where $x_{ij} \in X, i \in \{1, \ldots, k\}, j \in \{1, \ldots, s_k\}$. Then Y can be represented as a mixture of multiple regression equations*

$$Y = p_1Y_1 + p_2Y_2 + \ldots + p_kY_k + \varepsilon,$$

where $p_i \to \frac{n_i}{n}$, $i \in \{1, \ldots, k\}$, $n \to \infty$ and $\varepsilon \sim N(0; 1)$

Let us consider an example. A sample of 100 observations for profit modeling are given. All observations are divided into three economic periods: Stable (50 observations), Crisis (30 observations), and Recovery (20 observations). Let's consider 6 factors that affect profit in

different economic periods: demand, price, inflation, costs, investment and subsidies. They are denoted by $X_1, \ldots, X_6$, respectively. In the Stable period, profit is mainly driven by demand and price, reflecting a typical market equilibrium. For the Crisis period, inflation and costs play a crucial role, showing how financial instability affects profitability. And in the Recovery period, investment and subsidies become key, as external support and reinvestment contribute to economic recovery. So, we have the most important factors for each period. Then we present the data in a table, where each observation includes the following variables: Period — The economic period in which the data was recorded (Stable, Crisis, or Recovery).

Factor 1 — The main factor of influence, which varies depending on the period. In the stable period, it is demand (measured in units), in the crisis period — inflation. (percentage), and in the recovery period — investment (monetary units).

Factor 2 — The secondary influencing factor, which also depends on the period. In the stable period, it is price (monetary units per unit), in the crisis period — costs (monetary units), and in the recovery period — subsidy (monetary units).

Profit — The resulting profit (monetary units), influenced by the two factors above.

A fragment of the table is given below:

Period	Factor1	Factor1 Name	Factor2	Factor2 Name	Profit
Stable	98.49	Demand	29.46	Price	282.96
Stable	107.91	Demand	26.02	Price	292.38
Stable	103.98	Demand	27.43	Price	281.74
Stable	97.47	Demand	24.14	Price	235.71
Stable	111.94	Demand	21.62	Price	241.68
Crisis	10.42	Inflation	72.46	Costs	161.37
Crisis	12.31	Inflation	83.51	Costs	142.98
Crisis	11.54	Inflation	91.32	Costs	131.47
Crisis	6.85	Inflation	51.27	Costs	178.32
Crisis	15.24	Inflation	87.96	Costs	124.57
Recovery	45.68	Investment	7.12	Subsidy	50.47
Recovery	50.14	Investment	14.23	Subsidy	53.78
Recovery	73.65	Investment	5.87	Subsidy	74.92
Recovery	46.78	Investment	12.14	Subsidy	71.32
Recovery	39.21	Investment	13.65	Subsidy	58.94

Table 1: Profit

Then, we construct a regression equation for each of the periods. Here Y_1 is "Profit" in Stable, Y_2 is "Profit" in Crisis, Y_3 is "Profit" in Recovery After calculation, we obtain three multiple regression equations with the corresponding coefficients of determination

$$Y_1 = -240 + 2.7X_1 + 9.45X_2, \quad R^2 = 0,9775,$$
$$Y_2 = 239.45 + 0.096X_2 - 1.586X_3 - 0.97X_4, \quad R^2 = 0,987,$$
$$Y_3 = 6.98 - 0.024X_4 + 0.94X_5 + 0.11X_6, \quad R^2 = 0,995.$$

Then, we need to calculate probability

$$p_1 = \frac{50}{100} = 0.5, p_2 = \frac{30}{100} = 0.3, p_3 = \frac{20}{100} = 0.2$$

Now, we can construct the mixture of this regression equations

$$Y = p_1Y_1 + p_2Y_2 + p_3Y_3 = -46.78 - 1.35X_1 + 4.75X_2 - 0.48X_3 - 0.29X_4 + 0.19X_5 + 0.02X_6.$$

Open Problems. In this article, mixture of multiple regression equations are considered. But there are possible cases, when regression estimators are defined other methods then the least squares method. In particular, the most powerful methods of regression shrinkage and selection of estimators are the lasso method ([7–9]), the ridge estimation ([10]), the elastic net ([11,12]). For these cases mixture of multiple regression equations are not considered yet.

REFERENCES

1. R. Maiboroda, V. Miroshnychenko, O. Sugakova, *Quantile estimators for regression errors in mixture models with varying concentrations*, Bulletin of Taras Shevchenko National University of Kyiv. Physical and Mathematical Sciences, **78** (2024), №1, 45–50. https://doi.org/10.17721/1812-5409.2024/1.8
2. R. Maiboroda, V. Miroshnychenko, *Asymptotic normality of modified LS estimator for mixture of nonlinear regressions*, Modern Stochastics: Theory and Applications, **7** (2020) №4, 435–448. https://doi.org/10.15559/20-VMSTA167
3. R. Maiboroda, O. Sugakova, Estimation and classification by observations from a mixture, Kyiv University, Kyiv, 2008. (in Ukrainian)
4. V.O. Miroshnychenko, *Residual analysis in regression mixture model*, Bulletin of Taras Shevchenko National University of Kyiv, Series. Physics and Mathematics, **3** (2019), №3, 8–16. https://doi.org/10.17721/1812-5409.2019/3.1
5. B. Grün, F. Leisch, *Fitting finite mixtures of linear regression models with varying & fixed effects in* R. In Alfredo Rizzi and Maurizio Vichi (Eds.), Compstat 2006, Proceedings in Computational Statistics, Heidelberg: Physica Verlag, 2006, 853–860.
6. Ya.I. Yeleyko, O.A. Yarova, *Mixture of distributions based on the Markov chain*, Cybernetics and System Analysis, **58** (2022), №5, 754–757. https://doi.org/10.1007/s10559-022-00508-4
7. M. Gruber, Improving efficiency by shrinkage: The James–Stein and Ridge regression estimators, CRC Press, part 2, 1998.
8. Y. Jiang, *Variable selection with prior information for generalized linear models via the prior lasso method*, J. Amer. Stat. Assoc., **111** (2016), №513, 355–376. https://doi/10.1080/01621459.2015.1008363
9. R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal Stat. Soc. Series B (methodological), **58** (1996), №1, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
10. A.K. Md. Ehsanes Saleh, M. Arashi, B.M. Kibria, Theory of Ridge regression estimation with applications, New York: John Wiley & Sons, 2019.
11. J.K. Tay, B. Narasimhan, T. Hastie, *Elastic net regularization paths for all generalized linear models*, J. Stat. Software, **106** (2023), №1. https://doi.org/10.18637/jss.v106.i01
12. H. Zou, T. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Stat. Soc. Series B (statistical Methodology), Wiley, **67** (2025), №2, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Lutsk National Technical University
Lutsk, Ukraine
yeleykot@gmail.com

Ivan Franko National University of Lviv
Lviv, Ukraine
oksana.yarova@lnu.edu.ua

Received 04.03.2025
Revised 20.06.2025