National Research University Higher School of Economics

As a manuscript

Sergei Golovan

Econometric models for analysis of efficiency of economic agents

PhD Dissertation Summary
for the purpose of obtaining an academic degree
Doctor of Philosophy in Economics

Academic supervisor:
Professor Anatoly Peresetsky,
Doctor of Economic Sciences

JEL: C21, D22, G21

Moscow – 2024

Motivation

This dissertation, which is a summary of the author's scientific papers for the period from 2006 to 2023, is devoted to the development of parametric and non-parametric econometric methods of studying the efficiency of economic agents (enterprises, banks, hospitals, etc.) as well as to the application of these methods to the Russian banking system.

In this dissertation, efficiency of economic agents is investigated using the following methods of estimating production possibility frontier: data envelopment analysis, which estimates production possibility frontier non-parametrically by constructing a conical or convex hull in the space of outputs and inputs; stochastic frontier analysis, for which the distribution of deviation from the frontier is specified, so the frontier can be estimated by the maximum likelihood; quantile regression, which estimates production possibility frontier parametrically as a high-order quantile of output (usually considered $\tau = 0.9, 0.95$) conditional on given inputs.

In this dissertation, methods of constructing production possibility frontiers are applied to the analysis of issues in the banking industry of the Russian Federation, which are of current interest. The relevance of these methods comes from the fact that their results can be used by regulators for preliminary screening of banks in terms of efficiency and sustainability.

The theoretical part of the dissertation develops methods of quantile regression, specifically, it offers counterexamples to the main statements of Canay (2011) and proposes a consistent estimator of the covariance matrix of quantile regression coefficients with instrumental variables (IVQR) for clustered data. Quantile regressions make it possible to split the population of economic agents (enterprises, banks) into groups according to their efficiency, where high quantiles correspond to more efficient enterprises, low quantiles correspond to less efficient enterprises, and then estimate its own production function or cost function for each efficiency group.

Econometric methods of estimating production possibility frontiers and studying the efficiency of economic agents began to appear in the middle of the 20th century and continue their development to this day. The first works on Data Envelopment Analysis include Koopmans (1951), and Charnes, Cooper and Rhodes (1978). Koopmans (1951) used Data Envelopment Analysis to evaluate the efficiency of public services of the West African Economic Monetary Union. It has been shown that the level of wages and the degree of corruption control have a positive effect on the efficiency of public services. Charnes, Cooper and Rhodes (1978) fully developed the conical model of Data Envelopment Analysis with an algorithm for estimating production possibility frontier and cost function using linear programming. More recent works on Data Envelopment Analysis derive statistical properties of the production possibility frontier estimators. Specifically, Simar and Wilson (2000) proposed a bootstrap method of computing bias (and reducing it) and variance of efficiency score estimates, and also constructing confidence intervals for efficiency scores. These methods were used to evaluate the efficiency of an experimental program introduced in schools in the United States, and after correcting for bias, it was shown that the program did not improve school efficiency significantly (in contrast to the previous studies). Stochastic Frontier Analysis was developed by Aigner, Lovell and Schmidt (1977), in which the error term in the regression model is decomposed into two components: a random symmetric component and another asymmetric component, which corresponds to the technical efficiency of the economic agent. The authors derive the asymptotic behavior of the maximum likelihood estimates and accompany them by simulations using the Monte Carlo method. Also in the 2000s, quantile regression, which was first developed by Koenker and Bassett Jr. (1978), began to be used for estimating production possibility frontiers and investigating the efficiency of economic agents. The paper proposes a method for estimating the conditional quantile of order $\tau \in (0, 1)$ of a dependent variable by minimizing the sum of weighted absolute values of the residuals, derives the asymptotic distribution of coefficient estimators, and shows that for heavy-tailed errors, the quantile regression estimator is more efficient (have lower variance) than the least squares estimates. The application of quantile regression for panel data with fixed effects is complicated by the fact that conditional quantile, unlike conditional expectation, is not linear, so individual effects cannot be excluded, and accordingly, statistical inference can only be obtained for the so-called long panels (for which n/T is low enough). Dhaene and Jochmans (2015) propose a general method for reducing the bias when estimating nonlinear models for panel data, in which estimator is constructed as a linear combination of the estimators over the entire panel $t = 1, \dots, T$ and over its two halves $t = 1, \dots, T/2$, $t = T/2 + 1, \dots, T$. This method cannot be applied to ordinary quantile regression, since the expressions for the bias of coefficient estimators in panel data cases have not yet been obtained; accordingly, it is not known whether quantile regression satisfies the assumptions of Dhaene and Jochmans (2015). Galvao and Kato (2016) proposed a method for estimating quantile regression coefficients by minimizing a smoothed loss function, which made it possible to calculate the bias and reduce it using the method from Dhaene and Jochmans (2015) for short panels. Several papers develop quantile regression to estimate production possibility frontier and technical efficiency: Liu, Laporte and Ferguson (2008) and Jradi and Ruggiero (2019). Liu, Laporte and Ferguson (2008) introduce efficiency estimates for the production function obtained using quantile regression and also compare efficiency estimates obtained using different methods (quantile regression, stochastic frontier, data envelopment analysis) for simulated data. Jradi and Ruggiero (2019) consider the formal problem of stochastic data envelopment analysis and use quantile regression to estimate the production possibility frontier for this problem. There are also some applied works using this technique. Behr (2010) compares the efficiency of German banks estimated using quantile regression with the efficiency estimated using stochastic frontier analysis. The results are different, which indicates that the stochastic frontier analysis is not applicable in this case (quantile regression does not impose restrictions on the distribution of errors in the regression). Chidmi, Solís and Cabrera (2011) use similar methods to assess the technical efficiency of dairy farms in Wisconsin, USA. Besstremyannaya (2017) applied quantile regression to estimate the cost function of Japanese banks, as well as to identify the impact of the financial crisis of 2007–2009. and the 2011 earthquake on bank performance.

To this date there are several papers on application of econometric methods to study efficiency of Russian banks and firms. These are the monograph of Peresetsky (2012), which summarizes application of econometric methods to the banking industry, in particular, data envelopment analysis and stochastic frontier analysis. Afanasiev and Vasilieva (2006), and Afanasiev (2006) compare estimates of the efficiency of Moscow enterprises specializing in the production and sale of household goods, constructed using different models estimated by stochastic frontier analysis. They introduce the concept of controllable factors of efficiency, as well as the costs of managing these factors. Aivazian and Afanasiev (2007) introduced the concept of the achievable production potential of an enterprise (production potential taking into account factors of inefficiency) and assessed technical efficiency indicators for Moscow enterprises taking this concept into account. Aleskerov et al. (2008) and Aleskerov et al. (2010) apply stochastic frontier analysis to Russian banks. Aleskerov et al. (2008) study behavior of Russian banks over time. Aleskerov et al. (2010) study their efficiency depending on the ownership structure. They show that the more concentrated the ownership the more efficient the bank.

Objectives of the research

The goals of this dissertation are first the development of econometric methods for modeling the efficiency of economic agents,and then the application of the models to assess the efficiency of Russian banks from various points of view.

During the study, the following **problems** were solved:

1. Models for constructing a stochastic production possibility frontier for Russian banks are developed. Their efficiency in terms of the volume of loans issued and deposits attracted is estimated, and factors influencing the computed efficiency are identified (Golovan, 2006).

2. Models for constructing a stochastic frontier for estimating the cost function of Russian banks are developed. The efficiency of the banks is estimated from the point of view of minimizing costs (Golovan, Karminsky and Peresetsky, 2008).

3. Methods of data envelopment analysis (*DEA*) were used to estimate the production possibility frontier of Russian banks. The efficiency of the banks in terms of generating interest and non-interest income is assessed. The robustness of the conclusions is confirmed by comparing efficiency indicators estimated by the methods of data envelopment analysis and stochastic frontier analysis (Golovan, Nazin and Peresetsky, 2010).

4. Counterexamples to the statements concerning asymptotic distribution of the quantile regression estimator, which was proposed in the paper by Canay (2011), for panel data (Besstremyannaya and Golovan, 2019) are constructed.

5. An estimate of the covariance matrix of the estimator of quantile regression coefficients with instrumental variables for the case of clustered observations (Besstremyannaya and Golovan, 2023) has been developed.

6. Resampling methods for obtaining statistical inference on quantile regression coefficients with instrumental variables for the case of clustered observations (Besstremyannaya and Golovan, 2023) have been developed.

The contribution of the dissertation is as follows:

1. Stochastic frontier methods have been applied for the first time to study the efficiency of Russian banks in attracting deposits (Golovan, 2006).

2. Stochastic frontier methods have been applied for the first time to study the efficiency of Russian banks in terms of cost minimization (Golovan, Karminsky and Peresetsky, 2008).

3. For the first time, methods of data envelopment analysis have been applied to study the efficiency of Russian banks in terms of maximizing interest and non-interest income (Golovan, Nazin and Peresetsky, 2010).

4. For the first time, the asymptotic properties of the estimator of quantile regression coefficients with instrumental variables for clustered data have been developed (Besstremyannaya and Golovan, 2023).

5. For the first time, the applicability of cluster resampling methods for approximating the distribution of the Kolmogorov–Smirnov and Cramér–von Mises statistics for statistical testing of hypotheses regarding estimates of quantile regression coefficients with instrumental variables for clustered data has been proven (Besstremyannaya and Golovan, 2023).

The research methodology

The object of the research is various economic agents with the examples of Russian banks.

The subject of the research is the unobservable technical efficiency of the agents (Russian banks) from various points of view: efficiency in attracting deposits and issuing loans (the main banking activities) and cost efficiency. The definition of efficiency means that there is a production possibility frontier, and, accordingly, its estimation is also included in the subject of study.

The informational base of the research is the statistical data on individual banks of the Russian banking system, as well as macroeconomic data on the Russian economy. In particular, the study employs balance sheet data and profit/loss statements of banks provided by the Central Bank of the Russian Federation, additional bank indicators calculated by the Mobile Information Agency, and data from Rosstat on macroeconomic indicators.

Main findings

1. Using the stochastic frontier approach, the production functions of Russian banks were estimated in terms of the volume of loans issued and in terms of the volume of deposits attracted. The model of lending takes the volume of deposits and loans from other banks, as well as administrative expenses, as inputs. The model of deposits attracted takes net assets and administrative expenses as inputs. The stability of production functions over time is shown. For each of the two models, estimates of the technical efficiency of banks were constructed. The following factors were revealed to impact the efficiency: the ratio of equity capital to net assets, overdue debt to loans, reserves for possible losses to net assets, and the of the bank in Moscow, in St. Petersburg, or in a regional city (Golovan, 2006).

2. Using the stochastic frontier approach, the cost function of Russian banks is estimated, and the technical efficiency of banks is computed as the excess of bank costs over the lowest possible cost value found. The cost function depends on the following factors: bank products (loans to individuals, firms, and other banks), prices of the resources (interest rates on deposits, loans, unit costs for securities), and fixed resources (equity capital). In addition, the cost function model has been improved by including risk factors and the quality of bank assets (overdue debt, other non-performing assets, liquidity ratio, and share of loan reserves). It has been shown that the introduction of risk factors leads to a significant improvement in the quality of the model (Golovan, Karminsky and Peresetsky, 2008).

3. Using nonparametric methods of data envelopment analysis, the production possibility frontier of Russian banks was constructed in a model with two outputs (net interest income and net non-interest income). The personnel expenses, reserves for possible losses and other expenses were considered as inputs. The technical efficiency of Russian banks was calculated as the distance to the production possibility frontier. It is shown that banks with foreign capital are on average more efficient than Russian banks, and regional banks are also more efficient on average than Moscow ones. The estimation results are stable over time; the use of stochastic boundary methods gives comparable results (Golovan, Nazin and Peresetsky, 2010).

4. Methods for estimating quantile regression for panel and clustered data have been developed. For quantile regression with panel data, several of methodological errors were indicated when deriving the asymptotic distribution of coefficient estimates in the paper by Canay (2011). For quantile regression with instrumental variables in the clustered case, the asymptotic distribution of coefficient estimates was derived and the possibility of using cluster resampling to get statistical inference regarding the conditional quantile process as a whole was proven (Besstremyannaya and Golovan, 2019; Besstremyannaya and Golovan, 2023).

Paper structure

This dissertation is a collection of research papers written by the author over the period from 2006 to 2023 and includes the application of the following econometric methods to study the efficiency of Russian banks:

1. Stochastic frontier analysis models.
2. Envelopment data analysis models.
3. Quantile regression models with panel data.
4. Quantile regression models with cluster data.

Stochastic frontier analysis models for studying banks' efficiency

The paper by Golovan (2006) was the first to examine the efficiency of Russian banks both in terms of issuing loans and attracting deposits. Golovan, Karminsky and Peresetsky (2008) were the first to evaluate the efficiency of Russian banks in terms of costs.

Stochastic frontier analysis models were developed to estimate the set of production possibilities of an economic agent. As a result, it is possible to use them to estimate the distance to the frontier of this set, which is interpreted as the technical efficiency of the agent (the larger the distance to the frontier, the lower the efficiency). Stochastic frontier analysis is widely used to evaluate the performance of banks around the world (Akhigbe and McNulty, 2003; Casu, Girardone and Molyneux, 2004; Casu and Molyneux, 2003; Hasan and Marton, 2003). However, at the time of writing Golovan (2006), only two papers were using this technique that analyzed the Russian banking system (Caner and Kontorovich, 2004; Styrin, 2005).

The stochastic frontier analysis model is formulated in Golovan (2006) as follows: if we assume that the bank's production function depends on some factors $x_1, \dots, x_k$ and has the form

$$y = F(x_1, \dots, x_k)$$

(here, the product y is considered either the volume of loans issued by the bank to non-financial organizations, or the volume of deposits attracted by the bank), then a real bank can produce less with the same set of factors:

$$y = F(x_1, \dots, x_k) \exp(-u) \leqslant F(x_1, \dots, x_k), \quad \text{where } u \geqslant 0.$$

The value of $\exp(-u)$ is called technical efficiency.

To estimate this value (together with other parameters of the bank), the Cobb–Douglas production function and the following formalization are assumed:

$$\begin{aligned} \ln y_t &= \beta_0 + \beta_1 \ln x_{1t} + \dots + \beta_k \ln x_{kt} + v_t - u_t, \\ v_t &\sim \mathcal{N}(0, \sigma_v^2), \\ u_t &\sim \mathcal{N}^+(0, \sigma_u^2) \quad \text{(half-normal distribution).} \end{aligned}$$

As one can see, the error splits into two components: random deviation v_t and asymmetric deviation u_t, the latter of which is interpreted as a source of inefficiency. Since u_t is an error component, it is impossible to estimate it accurately. Therefore, its expected value is used as an estimator of technical efficiency

$$\hat{E}_t = E(\exp(-u_t) \mid v_t - u_t = \hat{e}_t),$$

where $\hat{e}_t = \ln y_t - (\hat{\beta}_0 + \hat{\beta}_1 \ln x_{1t} + \dots + \hat{\beta}_k \ln x_{kt})$ are the regression residuals. The parameters of the model are estimated by the maximum likelihood method because of this composite structure of the error term.

In the second stage, for the estimates of the technical efficiency of individual banks, ordinary linear regression models are built, in which the influence of various factors on the efficiency of banks is studied.

Data for the paper are provided by the information agency “Mobile.” The database includes the banks’ quarterly balance sheets and profit/loss statements. The data for the period from the 1st quarter of 2003 to the 3rd quarter of 2005 were used to estimate the models. Vnesheconombank and Sberbank were excluded from the sample, as they operate under the conditions that are significantly different from conditions for other commercial banks.

The paper by Golovan (2006) evaluates the efficiency of the banks from two points of view: the efficiency of the banks in issuing loans and the efficiency of the banks in attracting deposits. At the same time, in each of the two cases, the production function of a bank is considered, i. e., the bank is considered a firm that uses resources to produce a product (loans or attracted deposits, respectively).

Model for issuing loans: A bank uses financial resources (loans from other banks, deposits of individuals and firms), labor resources (personnel), and physical capital (this variable was not included in the final model). The use of labor resources was introduced into the model through administrative expenses, while total deposits of individuals and firms and loans from other banks were presented directly.

Model for attracting deposits: In this model, as in the previous model, labor resources were introduced through administrative expenses, and net assets were used as the second factor reflecting the size of the bank.

In both models, estimates of the production function coefficients are stable with respect to changing the sample, and the distribution of technical efficiency is such that most of the banks fall into the range of 0.3–0.6 for the model for issuing loans (also, there are almost no banks with an efficiency close to one). On the other hand, the distribution of technical efficiency is close to being uniform at $[0, 1]$ for the model of attracting deposits.

For the second stage regression (influence of various factors on efficiency), the following factors were selected: log of equity, equity/net assets, arrears/loans, reserves/net assets, and binary variables for Moscow and St. Petersburg. It turns out that, other things being equal, in Moscow and St. Petersburg, banks issue fewer loans and attract fewer deposits than in other regions, which can be explained by the higher competition in these cities. The signs of the coefficients with other factors do not change when the sample is changed (we consider samples starting and ending by different quarters) and have meaningful economic explanations.

The second paper in which the stochastic frontier model is used is the paper by Golovan, Karminsky and Peresetsky (2008). It examines the efficiency of Russian banks in terms of minimizing costs, taking into account various risk factors. In this case, an efficient bank is a bank whose costs, other things being equal, are the lowest, so the stochastic frontier model is formulated a bit differently:

$$y = C(x_1, \dots, x_k)\exp(u) \geqslant C(x_1, \dots, x_k),$$

where $u \geqslant 0$ represents the bank's inefficiency. The regression equation is written as follows (in this paper, the panel data structure is taken into account):

$$\ln C_{it} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{jit} + v_{it} + u_i,$$

where C_{it} is the cost of the bank i in the period t; x_{it} is the vector of variables $(w_{it}, y_{it}, z_{it}, q_{it})$; x_{jit}, $j = 1, \dots, k$ are the components of the resource vector x_{it}; and β_j are the model coefficients. In this model, the costs are partly explained using resource prices, output, and volume of fixed resources, and are partly explained by technical inefficiency. The model accounts for panel data structure by including individual effects. The parameters of the model can be estimated by different methods, one of which is the maximum likelihood method. In this case, it is necessary to make assumptions about

Table 1: Indicators of banks used in the models of cost efficiency

Group of indicators	Indicators
Costs	Bank Operating Expenses
Fixed Resources	Equity
Resources	Deposits of individuals, interest expenses on deposits of individuals, deposits of firms, interest expenses on deposits of firms, loans of other banks, interest expenses on loans of banks, securities issued, expenses on securities
Estimated prices of resources	Interest rates on deposits of individuals, interest rates on deposits of firms, interest rates on loans from other banks, unit costs for servicing securities
Products	Loans to individuals, loans to firms, loans to other banks
Factors of risk and quality	Current liquidity ratio, non-performing loans (proxied by arrears or other non-performing assets), share of provisions for possible losses on loans in bank loans

the distributions of v_{it} and u_i: 1) $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$; 2) $u_i \sim \mathcal{N}^+(\mu, \sigma_u^2)$ (truncated normal distribution); and 3) v_{it} and u_i are independent of each other and the explanatory variables.

The paper uses the indicators of banks listed in Table 1. The risk factors are of particular interest because they had not been included in the cost function model previously. It should also be noted that estimated interest rates were considered resource prices, since true interest rates on deposits and loans are not published in the financial statements of the banks. Following the Laeven and Majnoni (2005) approach, the ratio of interest payments to the volume of deposits, taking into account inflation, was taken as interest rates on attracted funds, namely:

$$i_t^d = \frac{I_t/((P_{t-1} + P_t)/2)}{(D_{t-1}/P_{t-1} + D_t/P_t)/2},$$

where D_t is the deposits at the corresponding points in time, I_t is the interest expense on deposits, and P_t is the consumer price index in the corresponding period.

In Golovan, Karminsky and Peresetsky (2008), models with and without risk factors were estimated and compared to each other. They showed that the inclusion of risk factors and/or asset quality in the model improves the quality of the model and makes a more adequate assessment of the performance of the banks.

Additionally, another group of models was considered to explain inefficiency estimates similar to Golovan (2006). In these models, the estimated efficiency scores were treated as a dependent variable, and they were regressed on various parameters of the banks. The results were as follows. First, the negative effect of the size of the bank on its efficiency was confirmed. Second, it was shown that Moscow banks are more efficient than regional ones. Third, efficiency does not depend on whether the bank belongs to the deposit insurance system. For foreign banks, two trends can be noted: 1) higher quality of management associated with greater experience and availability of better technologies, and 2) use of expansionary tactics to conquer a new market. These two trends impact bank costs oppositely. As a result, it was shown that foreign banks are not more efficient than Russian ones. Moreover, when compared with large Russian banks, foreign banks were estimated to be slightly less efficient, since the second trend prevails. In addition, banks with foreign ownership, are more likely engaged in operations that are not related to traditional banking than banks with Russian ownership. Among the largest banks, newer banks are more efficient in terms of traditional banking, perhaps because they provide a smaller range of services. In addition, among the largest banks, growth in size is positively associated with efficiency.

Golovan (2006) and Golovan, Karminsky and Peresetsky (2008) apply a two-step procedure to estimate the impact of various factors on the output and cost efficiency of Russian banks, respectively. However, Wang and Schmidt (2002), and Schmidt (2011) show that for the stochastic frontier analysis models, the two-step procedure can lead to inconsistent estimates of the coefficients both at the first and the second steps. When selecting the two-step procedure, we followed existing literature where it has been used often in research on technical efficiency: Reinhard, Lovell and Thijssen (2000) used it to evaluate efficiency of dairy farms in the Netherlands, Hasan and Marton (2003) used it to research technical inefficiency of Hungarian banks, Greene (2004) used it to assess efficiency of provision of medical care using data of the World Health Organization, Afanasiev and Vasilieva (2006) used it to model production potential of Russian firms. Chidmi, Solís and Cabrera (2011) employed a two-step procedure to estimate the efficiency of dairy farms. It should also be noted that the results of estimation using a one-step procedure using the same data as in Golovan (2006) and Golovan, Karminsky and Peresetsky (2008) provide estimates of the coefficients of the production function and the cost function, which do not differ significantly from the estimates of the coefficients of the first step obtained by the two-step procedure.

Data envelopment analysis models for evaluating the efficiency of Russian banks

The paper by Golovan, Nazin and Peresetsky (2010) for the first time develops non-parametric models for evaluating the technical efficiency of Russian banks.

Data envelopment analysis models (*DEA*) are non-parametric models that construct production possibility sets as convex or convex conical hulls of individual data points whose coordinates represent the factors and outputs of the entities (enterprises/banks). Unlike parametric methods based on regression analysis (which include the stochastic frontier method discussed above), the DEA method allows one to naturally build a set of possibilities for production with several outputs, and it is not limited to a fixed parametrization of production or cost functions. The DEA method was introduced by Farrell (1957), and the two most popular specifications for the computable DEA models were developed by Charnes, Cooper and Rhodes (1978) and Banker, Charnes and Cooper (1984).

The CCR (Charnes, Cooper and Rhodes, 1978) model is a classic input-oriented linear programming problem, so the technical efficiency estimation problem is formulated as follows:

$$\begin{aligned} \max_{\lambda, t^-, t^+} \quad & \theta \\ \text{subject to } x_0/\theta &= X\lambda + t^-, \\ y_0 &= Y\lambda - t^+, \\ \lambda &\geqslant 0, \\ t^- &\geqslant 0, \\ t^+ &\geqslant 0. \end{aligned}$$

Here X is a $r \times n$-matrix consisting of the resource vectors of each bank in the sample; Y is a $s \times n$-matrix of outputs; x_0 and y_0 are the resource and output vectors of the bank whose technical efficiency is being estimated, which have dimensions $r \times 1$ and $s \times 1$, respectively; $X\lambda$, $Y\lambda$ are the resource and output vectors of some "artificial" bank which belongs to the cone hull of all banks in the sample constructed in the space of resources and outputs; λ is a $n \times 1$ vector of weights for all banks in the sample; t^- is a $r \times 1$ vector of resource slacks, i. e., the amount of resources used that could be discarded; and t^+ is a $s \times 1$ vector of potential additional outputs (output slacks), i. e., additional amount of products that the "artificial" bank is capable of producing. The value $\theta \geqslant 1$ is a measure of technical inefficiency ($\theta = 1$ for an efficient bank), and with values lying between 0 and 1, $1/\theta$ is an estimate of efficiency showing which share of the resources used by the bank was really necessary to produce the same amount of output. The weights λ in the model are non-negative, i. e., "artificial" banks are built in the conical hull in the space of resources and output. This means that the production function is considered to be a homogeneous function of degree 1, i. e., by multiplying all resource factors by 100, the bank is supposed to be able to increase all outputs by a factor of 100. This assumption is not always plausible. To make the assumption more strict, Banker, Charnes and Cooper (1984) proposed to use non-negative weights λ whose sum is equal to one, i. e., replace the conical hull with the convex one. In such a situation, each bank is compared with banks that are close to it in size and other indicators. This model is called BCC.

Since efficiency estimates based on DEA models are biased, a bootstrap procedure proposed by Simar and Wilson (2000) is used to reduce this bias.

Quarterly data on the balance sheets of Russian banks for the period from October 2002 to October 2006 were used to estimate the technical efficiency of the banks. The data were provided by the Mobile Informational Agency. For each reporting period, the sample consists of banks with general licenses from the Central Bank of the Russian Federation, for which the indicators listed in Table 2 are available.

Table 2: Indicators from DEA models

Group of indicators	Indicator
Resources	Maintenance costs Provisions for possible losses Other expenses
Outputs	Net interest income Net non-interest income

Following Drake, Hall and Simper (2006), to analyze the efficiency of the banks, they were divided into groups, and the dynamics of the average efficiency of these groups over time were evaluated. Banks were divided into groups according to the following sets of criteria:

- whether the bank is registered in Moscow or in another region (VTB, registered in St. Petersburg, was assigned to the group of Moscow banks);
- whether the bank is a foreign bank, i. e., having more than 50% of equity shares in the bank held by non-residents.

When analyzing differences between the Moscow and regional banks, the results of the CCR model demonstrate that in more than half of the periods, the median efficiency of the Moscow banks was significantly lower than that of the regional banks. The regional analysis in the BCC model shows that compared with the CCR model, the efficiency of the regional banks was higher than that of the Moscow ones in a smaller number of periods. Both models show a gradual increase in the efficiency of the regional banks since the beginning of 2005.

When analyzing differences between foreign and domestically owned banks in general, according to both models, it can be concluded that until 2004 it was not possible to find statistically significant differences in the efficiency of these two groups. However, since the autumn of 2004, there has been some increase in the relative efficiency of foreign banks.

In addition, Golovan, Nazin and Peresetsky (2010) compare estimates of technical efficiency obtained using DEA with the estimates obtained using stochastic frontier analysis. They show that the behavior of efficiency scores calculated by these two methods is similar. Ranking by efficiency scores from both models, the Spearman rank correlation coefficients between banks are high (0.6–0.9), i. e., both methods give similar results.

Quantile regression models with panel data for estimating efficiency

The paper by Besstremyannaya and Golovan (2019) is devoted to the development of quantile regression methods with panel data.

Quantile regression models were introduced in the 1970s (Koenker and Bassett Jr., 1978) and have been used, in particular, to estimate production possibility sets and the efficiency of enterprises since then. The estimation of technical efficiency is closely related to the estimation of a production possibility frontier. Quantile regression allows one to estimate a production possibility frontier as a sufficiently high quantile ($\tau = 0.9$–0.95) of the distribution of output conditional on fixed factors of production (Jradi and Ruggiero, 2019; Liu, Laporte and Ferguson, 2008). Residuals of such a quantile regression can be regarded as non-normalized performance indicators. Since microeconomic data often have a panel structure, in which the same economic agents are observed over several periods, models that take into account individual effects are necessary when working with this type of data. Such models for quantile regressions began to be developed at the beginning of the 21st century and continue to be developed to this day (Chetverikov, Larsen and Palmer, 2016; Galvao and Kato, 2016; Harding and Lamarche, 2014; 2016; Koenker, 2004; Machado and Santos Silva, 2019). At the same time, unlike conventional linear regression with panel data, quantile regression, due to its non-linearity, does not allow for a reduction in the number of estimated parameters (similar to within-group transformation). Therefore, to obtain good quantile regression estimates with individual effects, it is necessary either to put significant restrictions on the behavior of these effects (for example, make them parametric or require that they be random) or to use long panels, i. e., panels with a large number of periods.

In Besstremyannaya and Golovan (2019), theoretical issues with the statements and proofs from Canay (2011) are analyzed rigorously. It is shown that there are some errors in Canay (2011). The first error is that in the statement of the main theorem describing properties of the estimator it is claimed that the asymptotic behavior of the estimator is valid under the condition that n/T^a tends to zero for some $a > 0$ (that is, the estimators are valid for short panels if $a \in (0, 1)$). This statement is incorrect, and Besstremyannaya and Golovan (2019) give an example of a data-generating process for which the bias of the estimator for $n/T^a \to 0$ is too large relative to its standard error. This leads to a huge bias of z-statistics and misleading inferences. Accordingly, we show that without n/T tending to zero, standard statistical tests are invalid. The second error in Canay (2011) is that the statement of the theorem about properties of the estimators claims that the behavior of the estimator $\hat{\beta}_0(\tau)$ is asymptotically the same as the behavior of the other parameter estimators. Unfortunately, this is not true for any panel regression. Since information about the constant does not become more and more accurate as the number of periods T increases, the order in which $\hat{\beta}_0(\tau)$ converges to $\beta_0(\tau)$ is $1/\sqrt{n}$, and not $1/\sqrt{nT}$, as for other coefficients. Finally, the third error in Canay (2011) (from which the first two follow) is that the author uses an incorrect statement from the theory of ordinary linear regressions with panel data. In Chen and Huo (2021), the authors change the specification of the model, which makes it possible to rigorously prove asymptotic properties for the modified model estimator.

Besstremyannaya and Golovan (2019) also propose a method for modifying the Canay (2011) estimate, which reduces bias.

Quantile regression models with clustered standard errors for estimating efficiency

The paper by Besstremyannaya and Golovan (2023) is devoted to the development of corrections to the standard errors of quantile regression coefficients with instrumental variables for clustered data.

Clustered data, i. e., data in which observations are combined into clusters, within which there is interdependence of individual observations, are quite common in applied econometrics. Applying ordinary standard errors to regressions estimated on clustered data leads to a significant overestimation of the accuracy of estimates and often to incorrect statistical inference (Abadie et al., 2023; Cameron and Miller, 2015). Accordingly, for the case of clustered data, corrections to the standard errors of ordinary regression (Cameron and Miller, 2015; Wooldridge, 2003), quantile regression (Parente and Santos Silva, 2016), and instrumental variables (Cameron and Miller, 2015) have been developed.

Besstremyannaya and Golovan (2023) construct consistent standard errors for coefficients of quantile regression with instrumental variables for clustered data. The following econometric model introduced in Chernozhukov and Hansen (2005) is considered:

$$
\begin{gathered}
Y_{ik} = D'_{ik}\alpha(U_{ik}) + X'_{ik}\beta(U_{ik}), \\
D'_{ik} = \delta(X_{ik}, Z_{ik}, \nu_{ik}), \\
U_{ik} \sim U[0,1] \text{ does not depend on } X_{ik}, Z_{ik}, \\
\tau \mapsto D'_{ik}\alpha(\tau) + X'_{ik}\beta(\tau) \text{ is strictly increasing,}
\end{gathered}
$$

for which data are clustered. The vectors

$$
\begin{aligned}
\{(Y_{i1}, \dots, Y_{iK}), (D_{i1}, \dots, D_{iK}), (X_{i1}, \dots, X_{iK}), (Z_{i1}, \dots, Z_{iK}), i = 1, \dots, N\} \\
\sim \{(Y_1, \dots, Y_K), (D_1, \dots, D_K), (X_1, \dots, X_K), (Z_1, \dots, Z_K)\}
\end{aligned}
$$

are treated as independent identically distributed (i. i. d.) vectors (for simplicity, cluster size is assumed to be fixed). Next, the paper considers the parameter estimator $\theta = (\alpha, \beta)$ proposed in Chernozhukov and Hansen (2006). For this estimator its consistency is proved, and the asymptotic distribution of the quantile regression process is derived: denote $\varepsilon_k(\tau) = Y_k - D'_k\alpha(\tau) - X'_k\beta(\tau)$, $l_k(\tau, \theta(\tau)) = \tau - I(\varepsilon_k(\tau) < 0)$ and $\Psi_k(\tau) = [\Phi_k(\tau, Z_k, X_k)', X'_k]'$, then

$$
\sqrt{N}(\hat{\theta}(\cdot) - \theta(\cdot)) \Rightarrow b(\cdot)
$$

for $N \to \infty$, where $b(\cdot)$ is a zero-mean Gaussian process on $(0,1)$ which has covariance function $E(b(\tau)b(\tau')') = J(\tau)^{-1}S(\tau,\tau')J(\tau')^{-1}$, where

$$J(\tau) = E\left(\sum_{k=1}^{K} f_{\varepsilon_k(\tau)}(0\,|\,X_k, D_k, Z_k)\Psi_k(\tau)[D_k', X_k']\right),$$
$$S(\tau,\tau') = E\left(\sum_{k=1}^{K}\sum_{s=1}^{K} l_k(\tau,\theta(\tau))l_s(\tau',\theta(\tau'))\Psi_k(\tau)\Psi_s(\tau')'\right).$$

Here $f_{\varepsilon_k(\tau)}(0\,|\,X_k, D_k, Z_k)$ is the value of the regression error density at point zero. Next, consistent estimates of the components of the covariance function of the coefficient process are proposed as sample analogs of the expressions for $J(\tau)$ and $S(\tau,\tau')$ above:

$$\hat{J}(\tau) = \frac{1}{2Nh_N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} I(|\hat{\varepsilon}_{ik}(\tau)| \le h_N)\hat{\Psi}_{ik}(\tau)[D_{ik}', X_{ik}']\right),$$
$$\hat{S}(\tau,\tau') = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K}\sum_{s=1}^{K} l_{ik}(\tau,\hat{\theta}(\tau))l_{is}(\tau',\hat{\theta}(\tau'))\hat{\Psi}_{ik}(\tau)\hat{\Psi}_{is}(\tau')'\right),$$

where $\hat{\varepsilon}_{ik}(\tau) = Y_{ik} - D_{ik}'\hat{\alpha}(\tau) - X_{ik}'\hat{\beta}(\tau)$ and h_N are such that $h_N \to 0$ and $Nh_N^2 \to \infty$ (Parente and Santos Silva, 2016). This estimator of standard errors makes it possible to build test statistics for testing statistical hypotheses both for coefficients within one quantile index $\tau \in (0,1)$, and for coefficients for a finite set of different quantile indices $\{\tau_j \in (0,1), j = 1,\dots,J\}$.

For completeness of the study, Besstremyannaya and Golovan (2023) also propose a method for constructing percentage points and P-values for tests that test null hypotheses of the form $H_0\colon g(\theta(\tau)) = 0$, $\tau \in \mathscr{T} \subset (0,1)$. In this case, the statistics $v(\tau)$ designed to test the null hypotheses $H_0\colon g(\theta(\tau)) = 0$ against the hypotheses $H_1\colon g(\theta(\tau)) \neq 0$ for separate values of τ aggregated according to Kolmogorov–Smirnov or according to Cramér–von Mises are considered as test statistics, respectively:

$$KS = \max_{\tau\in\mathscr{T}} v(\tau), \quad CM = \int_{\tau\in\mathscr{T}} v(\tau)d\tau.$$

Such statistics have non-standard distribution under the null hypothesis, so resampling or bootstrap is used to obtain P-values. For quantile regression on i. i. d. data sample the resampling method was proposed in Chernozhukov and Fernández-Val (2005), for quantile regression with instrumental variables and independent observations it was developed in Chernozhukov and Hansen (2006). In the paper by Besstremyannaya and Golovan (2023) it is shown that resampling by separate clusters helps to get consistent estimates of percentage points and P-values for the tests under consideration. It should also be noted that for quantile regression with instrumental variables, direct resampling is challenging because obtaining $\hat{\theta}(\tau)$ estimates is a computationally complex task. Therefore, to test the null hypothesis $H_0\colon R(\tau)(\theta(\tau) - r(\tau)) = 0$ for all $\tau \in \mathscr{T}$, resampling of individual terms in the expansions

$$\sqrt{N}(\hat{\theta}(\cdot) - \theta(\cdot)) = -J(\cdot)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{k=1}^{K} l_{ik}(\cdot,\theta(\cdot))\Psi_{ik}(\cdot) + o_p(1)$$

and

$$\sqrt{N}(\hat{r}(\cdot) - r(\cdot)) = -H(\cdot)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{k=1}^{K} d_{ik}(\cdot,r(\cdot))\Upsilon_{ik}(\cdot) + o_p(1)$$

in $\ell^{\infty}(\mathscr{T})$, where $J(\tau)$ and $H(\tau)$ are non-random non-singular matrices and vectors $(l_{i1}(\tau,\theta(\tau))\Psi_{i1}(\tau),\ldots,l_{iK}(\tau,\theta(\tau))\Psi_{iK}(\tau))$ and $(d_{i1}(\tau,r(\tau))\Upsilon_{i1}(\tau),\ldots,d_{iK}(\tau,r(\tau))\Upsilon_{iK}(\tau))$ are independent and identically distributed for all τ.

Approbation of the research results

The results of this research were published by the author of the dissertation in Russian and foreign scientific journals, and were reported by the author or one of the co-authors at international conferences of scientific communities:

- VII April HSE International Conference "Modernization of Economy and the State" (2006)
- VIII April HSE International Conference "Modernization of Economy and Public Development" (2007)
- IX April HSE International Conference "Modernisation of Economy and Globalisation" (2008)
- 3rd April Conference "Applied Econometrics" by the Department of Applied Economics of the Faculty of Economic Sciences NRU HSE (2021)
- 5th April Conference "Applied Econometrics" by the Department of Applied Economics of the Faculty of Economic Sciences NRU HSE (2023)

The reliability of the results of econometric modeling is ensured by using alternative methods (parametric and non-parametric models), models with different sets of factors of production, types of output, control variables, alternative specifications (inclusion and exclusion of explanatory variables), and comparison of methodology and results of other studies carried out for Russian data and data for other countries.

Publications

Besstremyannaya, G., Golovan, S. (2019). Reconsideration of a simple approach to quantile regression for panel data. *Econometrics Journal*, *22*, 292–308. [2 author lists, Besstremyannaya: 1 author list, sections 1, 3, 4, 5, S2, S3, Golovan: 1 author list, sections 2, S1, the HSE A list, Scopus Q1, Web of Science Q2].

Besstremyannaya, G., Golovan, S. (2023). Instrumental variable quantile regression for clustered data. *Econometrics and Statistics*, *in press*. https://doi.org/10.1016/j.ecosta.2023.06.005 [1.5 author list, Besstremyannaya: 0.75 author lists, sections 1, 2.1, 3.1, 5, 6, 7, B, Golovan: 0.75 author lists, sections 2.2, 3, 4, A, the HSE C list, Scopus Q2].

Golovan, S. (2006). Factors influencing the efficiency of the Russian banks performance. *Applied econometrics*, *2*, 3–17. [In Russian, 0.5 author list, the HSE B list, Scopus, no quartile].

Golovan, S., Karminsky, A., Peresetsky, A. (2008). Cost efficiency of Russian banks taking risk factors into account. *Ekonomika i Matematicheskie Metody*, *44*(4), 28–38. [In Russian, 0.75 author list, Golovan: 0.25 author list, sections 3, 4, 5, the HSE D list].

Golovan, S., Nazin, V., Peresetsky, A. (2010). Non-parametric estimations of Russian banks' efficiency. *Ekonomika i Matematicheskie Metody*, *46*(3), 43–57. [In Russian, 1 author list, Golovan: 0.35 author list, sections 3, 5, the HSE D list].

References

Abadie, A., Athey, S., Imbens, G. W., Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics*, *138*(1), 1–35.

Afanasiev, M. (2006). Model of production potential with controllable factors of inefficiency. *Applied Econometrics*, *2*, 74–89. [In Russian].

Afanasiev, M., Vasilieva, N. (2006). Moleling production potential of a firm, accounting for inefficiency and risk factors. *Economic Science of the Modern Russia*, *1*, 104–119. [In Russian].

Aigner, D., Lovell, C. A. K., Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, *6*, 21–37.

Aivazian, S., Afanasiev, M. (2007). Evaluations of measures which control the factors of inefficiency of production. *Applied Econometrics*, *4*, 27–41. [In Russian].

Akhigbe, A., McNulty, J. E. (2003). The profit efficiency of small US commercial banks. *Journal of Banking & Finance*, *27*, 307–325.

Aleskerov, F., Belousova, V., Ivashkovskaya, I., Pogorelsky, K., Stepanova, A. (2010). Analysis of cost efficiency and influence distribution among bank shareholders. *Upravlenie v Kreditnoi Organizatsii*, *2*, 49–64. [In Russian].

Aleskerov, F., Belousova, V., Serduk, M., Solodkov, V. (2008). Behavioral stereotypes of the Russian banks. *Bankovskoe Delo*, *7*, 44–50. [In Russian].

Banker, R. D., Charnes, A., Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*, 1078–1092.

Behr, A. (2010). Quantile regression for robust bank efficiency score estimation. *European Journal of Operational Research*, *200*(2), 568–581.

Besstremyannaya, G. (2017). Heterogeneous effect of the global financial crisis and the great east japan earthquake on costs of japanese banks. *Journal of Empirical Finance*, *42*, 66–89.

Cameron, A. C., Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

Canay, I. (2011). A simple approach to quantile regression for panel data. *The Econometrics Journal*, *14*, 368–386.

Caner, S., Kontorovich, V. (2004). Efficiency of the banking sector in the Russian Federation with international comparison. *Economic Journal of the Higher School of Economics*, *8*, 357–375.

Casu, B., Girardone, C., Molyneux, Ph. (2004). Productivity change in European banking: A comparison of parametric and non-parametric approaches. *Journal of Banking & Finance*, *28*, 2521–2540.

Casu, B., Molyneux, Ph. (2003). A comparative study of efficiency in European banking. *Applied economics*, *35*, 1865–1876.

Charnes, A., Cooper, W. W., Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429–444.

Chen, L., Huo, Y. (2021). A simple estimator for quantile panel data models using smoothed quantile regressions. *The Econometrics Journal*, *24*(2), 247–263.

Chernozhukov, V., Fernández-Val, I. (2005). Subsampling inference on quantile regression processes. *Sankhyā: The Indian Journal of Statistics*, *67*, 253–276.

Chernozhukov, V., Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, *73*(1), 245–261.

Chernozhukov, V., Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, *132*(2), 491–525.

Chetverikov, D., Larsen, B., Palmer, C. (2016). IV quantile regression for group-level treatments, with an application to the distributional effects of trade. *Econometrica*, *84*(2), 809–833.

Chidmi, B., Solís, D., Cabrera, V. E. (2011). Analyzing the sources of technical efficiency among heterogeneous dairy farms: A quantile regression approach. *Journal of Development and Agricultural Economics*, *3*(7), 318–324.

Dhaene, G., Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, *82*(3), 991–1030.

Drake, L., Hall, M. J. B., Simper, R. (2006). The impact of macroeconomic and regulatory factors on bank efficiency: A non-parametric analysis of Hong Kong's banking system. *Journal of Banking & Finance*, *30*, 1443–1466.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society: Series A (General)*, *120*, 253–281.

Galvao, A. F., Kato, K. (2016). Smoothed quantile regression for panel data. *Journal of Econometrics*, *193*(1), 92–112.

Greene, W. (2004). Distinguishing between heterogeneity and inefficiency: Stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health economics*, *13*(10), 959–980.

Harding, M., Lamarche, C. (2014). Estimating and testing a quantile regression model with interactive effects. *Journal of Econometrics*, *178*, 101–113.

Harding, M., Lamarche, C. (2016). Penalized quantile regression with semiparametric correlated effects: An application with heterogeneous preferences. *Journal of Applied Econometrics*.

Hasan, I., Marton, K. (2003). Development and efficiency of the banking sector in a transitional economy: Hungarian experience. *Journal of Banking & Finance*, *27*, 2249–2271.

Jradi, S., Ruggiero, J. (2019). Stochastic data envelopment analysis: A quantile regression approach to estimate the production frontier. *European Journal of Operational Research*, *278*(2), 385–393.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, *91*(1), 74–89.

Koenker, R., Bassett Jr., G. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.

Koopmans, T. C. (1951). An analysis of production as an efficient combination of activities. In T. C. Koopmans (Ed.), *Activity Analysis of Production and Allocation* (Cowles Commission for Research in Economics, Monograph No. 13, pp. 33–97). Wiley.

Laeven, L., Majnoni, G. (2005). Does judicial efficiency lower the cost of credit? *Journal of Banking & Finance*, *29*, 1791–1812.

Liu, Ch., Laporte, A., Ferguson, B. S. (2008). The quantile regression approach to efficiency measurement: Insights from monte carlo simulations. *Health Economics*, *17*(9), 1073–1087.

Machado, J. A., Santos Silva, J. M. (2019). Quantiles via moments. *Journal of Econometrics*, *213*(1), 145–173.

Parente, P. M., Santos Silva, J. M. (2016). Quantile regression with clustered data. *Journal of Econometric Methods*, *5*(1), 1.

Peresetsky, A. (2012). *Econometric methods in remote analysis of Russian banks*. Publishing house of the Higher School od Economics. [In Russian].

Reinhard, S., Lovell, C. A. K., Thijssen, G. J. (2000). Environmental efficiency with multiple environmentally detrimental variables; estimated with SFA and DEA. *European Journal of Operational Research*, *121*(2), 287–303.

Schmidt, P. (2011). One-step and two-step estimation in SFA models. *Journal of Productivity Analysis*, *36*, 201–203.

Simar, L., Wilson, P. W. (2000). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, *27*, 779–802.

Styrin, K. (2005). *What explains differences in efficiency across Russian banks?* (Working Paper). EERC Research Network, Russia and CIS.

Wang, H.-J., Schmidt, P. (2002). One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *journal of Productivity Analysis*, *18*, 129–144.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, *93*(2), 133–138.