УДК 656:004 (476)

*Н. Ю. ТРИФОНОВ [1], В. А. ЛИВИНСКАЯ [2], В. В. КОРЖУКОВ [2]*

# РЕГРЕССИОННАЯ МОДЕЛЬ ОЦЕНКИ АВТОМОБИЛЕЙ НА ОСНОВЕ ПАРСИНГА ИНТЕРНЕТ-ДАННЫХ

*[1] Белорусский государственный экономический университет*
*[2] Белорусско-Российский университет*

*Описываются результаты эконометрического моделирования рыночной стоимости на примере автомобилей класса В на основе сбора с помощью программы-парсера информации с популярного интернет-сайта. Полученная модель в виде двух статистически значимых регрессионных уравнений предназначена для использования в практике оценочной деятельности.*

***Ключевые слова:*** *легковые автомобили; оценочная деятельность; парсер сайтов; рыночная стоимость; уравнение регрессии; эконометрическая модель*

## Введение

Оценка стоимости автомобилей, находящихся на вторичном рынке, в настоящее время становится всё более востребованной. Это происходит, в первую очередь, из-за стремительного расширения этого рынка. Заказчиками в оценке выступают как физические лица (например, конкретный покупатель или продавец подержанного автомобиля) так и организации, нуждающиеся в оценке транспорта для различных целей. Информация об остаточной стоимости дорожного транспортного средства (автомобиля) востребована при его купле-продаже, постановке на учёт, оценке или переоценке основных средств предприятия, при передаче в залог, при оценке ущерба в результате дорожно-транспортного происшествия, при разводе супругов и иных имущественных спорах, в том числе для нужд судебной экспертизы [1–2]. Набирающая популярность программа покупки старых автомобилей по системе «trade-in» тоже предполагает владение актуальной рыночной информацией о его цене.

Развитость рынка подержанных автомобилей в странах ЕАЭС позволяет использовать для оценки статистические методы сравнительного подхода к оценке стоимости [2–3]. Выборки в несколько десятков объектов сравнения позволяли получать достаточно надёжные результаты. Тем не менее, с развитием информационных технологий стали создаваться более объёмные базы данных на интернет-порталах и сайтах, а также появились инструменты (т. н. парсеры) формирования на основе интернет-данных выборок с заданными характеристиками. Современные технологии позволяют собирать информацию с сайтов-агрегаторов объявлений для её дальнейшего использования. Это позволило поставить задачу эконометрического моделирования рыночной стоимости подержанного автомобиля, обладающего конкретными характеристиками. Для этого необходимо рассмотреть представительные выборки автомобилей различных классов, поскольку ранее [3] было показано, что параметры обесценивания со временем существенно зависямт от класса исследуемого автомобиля.

Данная статья посвящена описанию методики и результатам эконометрического моделирования средней цены на вторичном рынке на примере автомобилей, относящихся к одному из наиболее популярных классов – В (т. н. бизнес-класс).

## Исходные данные

Сбор первичной информации (row data) занимает обычно до 70 % всего времени, потраченного на моделирование. В данном исследовании он осуществлялся с помощью парсера Selenium WebDriver – инструмента для сбора информации с сайта AUTO.ru, содержащего на момент сбора около 560 000 объявлений о продаже.

В результате анализа выборки из 17742 объявлений (рассматривался город Москва, как наиболее интересный для белорусов сегмент российского рынка) было обнаружено, что 97,7% автомобилей имеют возраст до 36 лет (с 1983 по 2019 годы). Дальнейший анализ проводился по объявлениям для автомобилей с 1983 по 2019 годы выпуска. Группировка по классам и годам этих автомобилей представлен в таблице 1.

Распределение в выборке по возрасту следующее. Больше всего представлено автомобилей, возраст которых не более 7 лет (53%). Автомобилей, возраст которых от 7 до 17 лет, на рынке около 22%, от 17 до 27 лет – 18% и от 27 до 37 лет всего 7%.

Распределение в выборке по классам следующее. Больше всего присутствуют автомобили классов J (43%) и B (24%), причем 55% автомобилей класса B имеют срок эксплуатации от 17 до 27 лет. Следующий по представительности – класс C, 54% автомобилей этого класса имеют возраст от 7 до 17 лет, 41% автомобилей моложе 5 лет. Автомобили остальных классов в основном (80% и выше) эксплуатировались до 7 лет.

## Регрессионный анализ

В настоящей статье описано моделирование рыночной стоимости автомобиля на основе объявлений о продажи автомобилей класса B. В качестве инструмента анализа используется программа Statistica-7.

Первый шаг в анализе данных – визуализация. Очевидно, цена авто на вторичном рынке сильно зависит от возраста автомобиля. На рис. 1 представлено корреляционное поле для предиктора «цена» (в российских рублях) и одного из количественных регрессоров – «возраст» (автомобиля в годах) для класса B. В этом графике учтено, что из выборки были предварительно удалены результаты некоторых аномальных наблюдений для автомобилей возрастом до 10 лет.

Для отбора категориальных факторов строились частотные таблицы. Выяснилось: 53,3% автомобилей в выборке имеют задний привод и 46,53% – передний привод; 99,5% автомобилей в выборке имеет механическую коробку передач (присутствовали также автоматическая и роботизированная коробки передач); 96,93% автомобилей в выборке имеют бензиновый двигатель.

Таблица 1 – Результат выборочного наблюдения объявлений о продаже автомобилей на вторичном рынке (единиц)

| | Класс | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Год выпуска | A | B | C | D | E | F | J | M | S | Всего |
| 1983–1991 | 7 | 528 | 301 | 207 | 79 | 10 | 37 | 11 | 2 | 1182 |
| 1992–2001 | 98 | 2138 | 411 | 427 | 48 | 18 | 56 | 16 | 9 | 3221 |
| 2002–2011 | 114 | 1465 | 784 | 151 | 89 | 95 | 960 | 78 | 57 | 3794 |
| 2012–2019 | 11 | 273 | 249 | 612 | 948 | 500 | 6013 | 340 | 184 | 9137 |
| Всего | 223 | 3876 | 1444 | 1190 | 1085 | 613 | 7029 | 434 | 250 | 16152 |

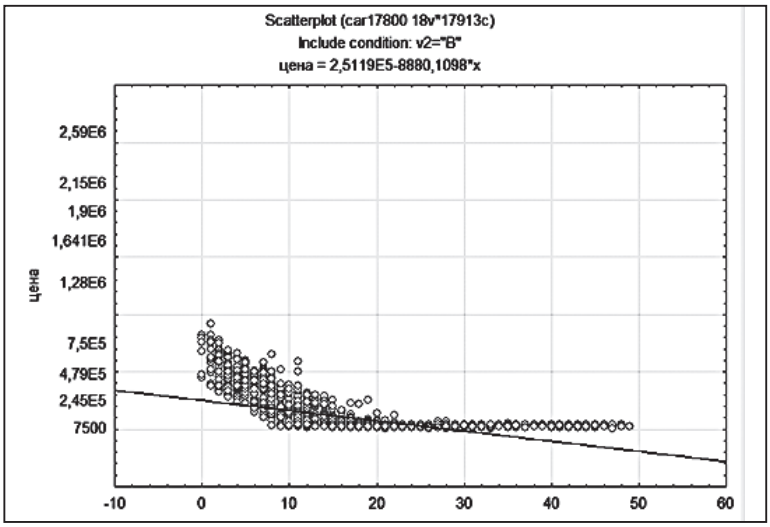


Рис. 1 – Корреляционное поле «цена»-«возраст» для анализируемой выборки

В качестве инструмента моделирования применялся множественный регрессионный анализ, реализованный в программе Statistica-7. На первом этапе был произведен отбор наиболее значимых признаков (feature selection), в модуле Data Mining. Он позволил сделать выводы о включения в модель тех же факторов, что и частотный анализ.

Для использования информации, содержащейся в категориальных факторах, применялся аппарат фиктивных (или, как их ещё называют, скалярных) переменных [4–5]. Для доказательства статистических различий у объектов с различными значениями категориальных признаков был использован программный модуль Noparametrics.

Так, для разделения всей совокупности объявлений на два класса по возрасту (до 10 лет и более) была введена бинарная переменная «возраст». На рис. 2 приведен результат сравнения цены в двух выборках с помощью непараметрического критерия Колмогорова. Нулевая гипотеза об отличии различий в средней остаточной стоимости автомобилей обеих групп отвергается на уровне значимости 0,05.

Аналогичные исследования были проделаны и для других признаков (фиктивных переменных): «производитель» (бинарная переменная «страна», равная 1 при производстве автомобиля в России, 0 – в противном случае), «коробка передач» (бинарная переменная FKOR(мех), равная 1 для автомобиля с механической коробкой передач, 0 – в противном случае), «привод» (бинарная переменная FPRIV(задний), равная 1 для автомобиля с задним приводом, 0 – с передним), «тип топлива» (бинарная переменная, равная 1 для автомобиля с бензиновым двигателем, 0 – для автомобиля с дизельным).

Факторы были проранжированы по степени влияния на результативный признак (цена) с помощью F-критерия. Наибольшее влияние оказывает фактор «производитель», далее по степени влияния – «возраст», «коробка передач», «мощность», «привод».

В целом анализ значимости различных факторов приведен на рис. 3.

Поскольку возраст автомобиля – это единственный параметр, изменение которого вызывает изменение цены конкретной модели на вторичном рынке, для спецификации регрессионной модели было, помимо исследования связи «цена»-«возраст», дополнительно построено корреляционное поле для предиктора «логарифм цены» и регрессора «возраст». Окончательный выбор был сделан в пользу полулогарифмической модели. Результат оценки выборки представлен ниже на рис. 4. Полученное уравнение является статистически значимым ($p < 0,05$). Величина коэффициента детерминации $R2 = 0,86$ показывает, что около 86% вариации средней цены предложения автомобилей класса В определяется именно вариацией выбранных факторов.

Для окончательного вывода о возможности использовать модель для предсказания средней цены на рынке проведен анализ остатков на нормальное распределение (в соответствии с предпосылками МНК Гаусса-Маркова). Гистограмма остатков (рис. 5) позволяет не отвергать гипотезу о нормальности остатков.

Также была проверена гипотеза об отсутствии гетероскедастичности с помощью графического анализа остатков. Зависимость между возрастом и вектором остатков не наблюдалась, что позволило не отклонять эту гипотезу.

Как итог, для предсказания стоимости автомобилей класса В, согласно сложившейся конъюнктуре цен предложения, может быть использована следующая модель:

$$Ln(цена) = 12,61 - 0,9 \times страна - 0,9 \times возраст - 0,094M - 0,3FKOR - 0.2 \times FPRIV.$$

При этом

$$страна = \begin{cases} 1, \text{ если автомобиль производства РФ,} \\ 0, \text{ в противном случае;} \end{cases}$$

$M$ – мощность двигателя в лошадиных силах;

*возраст* – в годах (разность между 2019 и годом выпуска авто);

$$FKOR = \begin{cases} 1, \text{ если у автомобиля механическая коробка передач,} \\ 0, \text{ в противном случае;} \end{cases}$$

$$FPRIV = \begin{cases} 1, \text{ если у автомобиля задний привод,} \\ 0, \text{ в противном случае.} \end{cases}$$

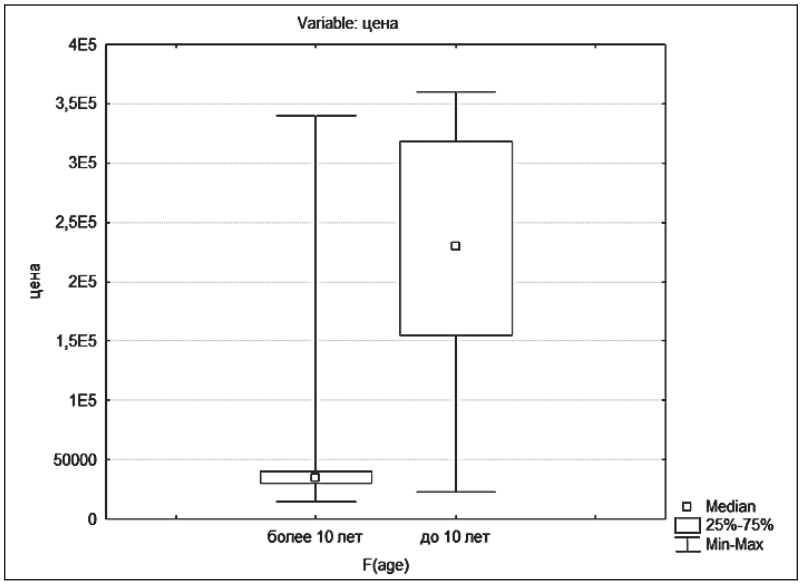


Рис. 2 – Сравнение статистической значимости различия средней цены автомобилей в выборках с возрастом до и после 10 лет

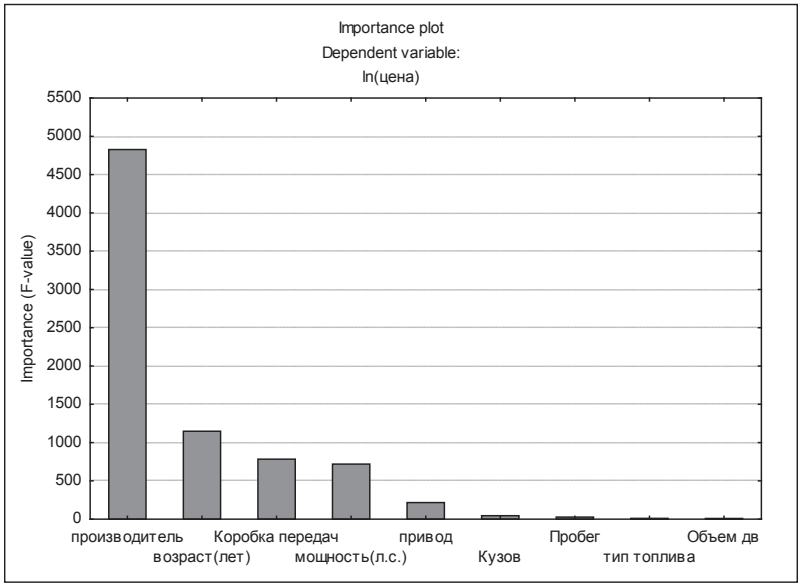


Рис. 3 – Сравнение факторов по силе влияния на результат по F-критерию

Regression Summary for Dependent Variable: ln(цена) (dCAR.sta)
R= ,92784345 R?= ,86089346 Adjusted R?= ,86062450
F(5,2586)=3200,8 p<0,0000 Std.Error of estimate: ,34615
Include condition: v6<=20

| N=2592 | Beta | Std.Err. of Beta | B | Std.Err. of B | t(2586) | p-level |
|---|---|---|---|---|---|---|
| Intercept | | | 12,61863 | 0,085479 | 147,6226 | 0,000000 |
| страна | -0,329865 | 0,010979 | -0,90468 | 0,030112 | -30,0438 | 0,000000 |
| **возраст(лет)** | -0,468925 | 0,010473 | -0,09420 | 0,002104 | -44,7760 | 0,000000 |
| M(л.с.) | 0,164393 | 0,010881 | 0,00951 | 0,000630 | 15,1080 | 0,000000 |
| FKOR(мех) | -0,076684 | 0,009865 | -0,30444 | 0,039166 | -7,7733 | 0,000000 |
| F(PRIV(задний)) | -0,106280 | 0,007907 | -0,20363 | 0,015150 | -13,4415 | 0,000000 |

Рис. 4 – Регрессионное уравнение для логарифма средней цены предложения автомобилей бизнес-класса на вторичном рынке

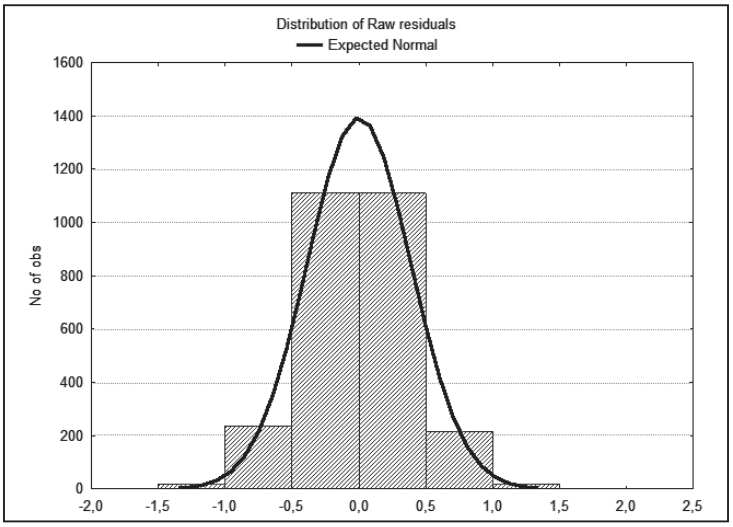


Рис. 5 – Демонстрация соответствия остатков нормальному распределению

## Заключение

Таким образом, с помощью современных информационных технологий для автомобилей класса В была получена эконометрическая модель рыночной стоимости в зависимости от возраста и других ценообразующих характеристик в виде регрессионного уравнения, использование которой в практике оценочной деятельности позволяет существенно повысить достоверность расчётов и уменьшить их трудоёмкость.

## ЛИТЕРАТУРА

1. Оценка стоимости машин, оборудования и транспортных средств: учебник / А. П. Ковалев [и др.]. – Москва: Интерреклама, 2003. – 488 с.
2. **Трифонов, Н. Ю.** Теория оценки стоимости: учебное пособие / Н. Ю. Трифонов. – Минск: Вышэйшая школа, 2017. – 208 с.
3. **Трифонов, Н. Ю.** Характеристика накопленного износа автомобилей методами финансовой математики / Н. Ю. Трифонов, С. В. Скрыган // Белорусский экономический журнал. – 2014. – № 3. – С. 133–143.
4. **Трифонов, Н. Ю.** Применение регрессионного анализа в сравнительном подходе оценки недвижимости в условиях переходной экономики / Н. Ю. Трифонов, С. А. Шимановский // Бухгалтерский учёт и анализ. – 2002. – № 4. – С. 44–52.
5. **Ливинская В. А.** Использование фиктивных переменных в прикладном регрессионном анализе вторичного рынка автомобилей / Ливинская В. А., Чегерова Т. И. // Государство и право: актуальные проблемы формирования правового сознания: сборник статей II Международной научно-практической конференции, 30 ноября 2018 г., г. Могилев: МГУ имени А. А. Кулешова, 2019. – С. 36–39.

## REFERENCES

1. Machinery, equipment and vehicles valuation: manual / A. P. Kovalev [a. o.]. – Moscow: Interreklama, 2003. – 488 p.
2. **Trifonov, N. Yu.** Theory of valuation: tutorial / N. Yu. Trifonov. – Minsk: Vysheyshaya shkola, 2017. – 208 p.
3. **Trifonov, N.** Description of vehicles' accumulated depreciation by financial mathematics techniques / N. Trifonov, S. Skrygan // Belarusian Economic Journal. – 2014. – № 3. – P. 133–143.
4. **Trifonov, N. Yu.** The use of regression analysis in a comparative approach to real estate valuation in transition economy / N. Yu. Trifonov, S. A. Shimanovsky // Accounting and Analysis. – 2002. – № 4. – P. 44–52.
5. **Livinskaya V. A.** Use of fictitious variables in applied regression analysis of the secondary car market / Livinskaya V. A., Chegerova T. I. // State and law: actual problems of legal consciousness formation: collection of articles of the II International scientific and practical conference, November 30, 2018. – Mogilev: Kuleshov Mogilev State University, 2019. – P. 36–39.

*Поступила* *10.04.2020*
*После доработки* *10.05.2020*
*Принята к печати* *01.06.2020*

*TRIFONOV N. Yu.[1], LIVINSKAYA V. A. [2], KORZHUKOV V. V. [2]*

# REGRESSION MODEL FOR CAR VALUATION BASED ON INTERNET DATA PARSING

*[1] Belarus State Economic University*
*[2] Belarusian-Russian University*

*The results of econometric modeling of market value are described on the example of class B cars based on the collection of information from a popular website using a parser program. The resulting model in the form of two statistically significant regression equations is intended for use in the valuation practice.*

***Keywords:*** *cars; econometric modeling, market value, site parser, statistical regression, valuation.*

**Трифонов Николай Юрьевич** – кандидат физико-математических наук, доцент, почётный оценщик Республики Казахстан, доцент БГЭУ, председатель Белорусского общества оценщиков. Научные интересы – теория оценки стоимости, экономика предприятий, методы статистического моделирования.

**Trifonov Nikolai Yurievich** – PhD (Phys.-Math.), Docent, Honorary Appraiser of the Republic of Kazakhstan, Associate Professor of Belarus State Economic University, Chairman of the Belarusian Society of Valuers. His research interests focus on valuation theory, business microeconomics, statistical modeling methods.

**Ливинская Виктория Александровна**– кандидат физико-математических наук, доцент. Научные интересы – прикладная статистика и эконометрика, машинное обучение, анализ медицинских данных, наука о данных.

**Livinskaya Victoriya Alexandrovna** – PhD (Phys.-Math.), Docent. His research interests focus on applied statistics and econometrics, machine learning, medical data analysis, data science.

**Коржуков Владимир Витальевич** – младший инженер по автоматизации тестирования программного обеспечения в компании EPAM Systems, магистрант по специальности «Системный анализ и управление обработка информации». Научные интересы – теория оценки стоимости, методы статистического моделирования.

**Korzhukov Vladimir Vitalievich** – Junior Software Test Automation Engineer at EPAM Systems, undergraduate in «System Analysis and Information Processing Management». His research interests focus on valuation theory, statistical modeling methods.

Работа выполнена рамках работы над научной темой государственного бюджетного финансирования Республики Беларусь ГБ 17–200 «Теоретические, методические и практические вопросы оценки стоимости в условиях современного рынка».