*УДК 519.862.6*



***М. П. Базилевский*** [1]

[1] *Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

# СИНТЕЗ МОДЕЛИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ И РЕГРЕССИИ ДЕМИНГА: ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ ОЦЕНОК ПАРАМЕТРОВ И КРИТЕРИЕВ АДЕКВАТНОСТИ ОТ СООТНОШЕНИЯ ДИСПЕРСИЙ ОШИБОК ПЕРЕМЕННЫХ

**Аннотация**. *В работе рассмотрен синтез модели множественной линейной регрессии и регрессии Деминга, который дает бесчисленное множество различных и не изучавшихся ранее оценок неизвестных параметров в рамках метода наименьших квадратов. На примере моделирования грузооборота железнодорожного транспорта для синтезированной регрессии впервые получены зависимости оценок параметров и критериев адекватности от соотношения дисперсий ошибок переменных. Показано, что с использованием рассмотренного синтеза можно снижать аппроксимационные качества классической модели множественной линейной регрессии ради повышения некоторых других её важных характеристик. Также синтезированная модель может выступать как инструмент для решения задачи отбора наиболее информативных регрессоров.*

**Ключевые слова**: *модель множественной линейной регрессии; регрессия Деминга; метод наименьших квадратов; коэффициент детерминации; критерий Дарбина-Уотсона; критерий согласованности поведения; грузооборот железнодорожного транспорта.*

***M.P. Bazilevskiy*** [1]

[1] *Irkutsk State Transport University, Irkutsk, Russia*

# SYNTHESIS OF MULTIPLE LINEAR REGRESSION AND DEMING REGRESSION MODEL'S: INVESTIGATION THE DEPENDENCES OF PARAMETER ESTIMATES AND ADEQUACY CRITERIA ON THE RATIO OF VARIANCE ERROR VARIABLES

**Abstract**. *The paper discusses a synthesis of multiple linear regression and Deming regression model's, which gives countless different and previously unknown estimates of unknown parameters in framework of ordinary least squares. Using example of railway freight turnover modeling for synthesized regression, for the first time, dependences of parameter estimates and adequacy criteria on a ratio of error variances of variables are obtained. It is shown that using the considered synthesis it is possible to reduce the approximation qualities of the classical multiple linear regression model for the sake of improving some of its other important characteristics. Also, the synthesized model can act as a tool for solving the problem of subset selection in regression.*

**Keywords**: *multiple linear regression model; Deming regression; ordinary least square; coefficient of determination; Durbin-Watson test; behavior consistency criterion; rail freight traffic.*

## Введение

Как известно, самым простым методом оценивания неизвестных параметров регрессионных моделей является метод наименьших квадратов (МНК) [1, 2], одной из предпосылок которого является то, что значения независимых переменных являются детерминированными. Однако на практике независимые переменные могут иметь и стохастический характер, что происходит, например, когда допущены ошибки при регистрации их значений [3]. В англоязычных источниках регрессионные модели с ошибками в независимых переменных известны как «Errors-in-Variables models», а их простейшим представителем является регрессия Деминга [4–6], содержащая только одну независимую переменную. Для оценки таких моделей уже нельзя руководствоваться классическим МНК, а следует применять другие известные методы.

К сожалению, регрессионные модели с ошибками в независимых переменных в настоящее время не находят такого широкого практического применения, как регрессии без ошибок. Главными назначениями последних являются прогнозирование значений зависимой переменной и интерпретация, например, коэффициентов регрессии. Однако в работе [7] впервые показано, что регрессия Деминга в сочетании с классической моделью множественной линейной регрессии формирует такую спецификацию, которая не лишена возможностей интерпретации и прогнозирования. При этом предложенный в работе [7] синтез дает бесчисленное множество различных и не изучавшихся ранее оценок в рамках МНК, а, следовательно, каждая такая оцененная модель характеризуется своим множеством критериев адекватности [8, 9]. Целью данной работы является исследование зависимостей этих оценок и критериев адекватности от соотношения дисперсий ошибок переменных на примере моделирования грузооборота железнодорожного транспорта.

**Синтез модели множественной линейной регрессии и регрессии Деминга**

Подробно этот синтез рассмотрен в работе [7]. Приведем его краткое описание.

Пусть $y_i$, $x_{i1}$, $i=\overline{1,n}$ – фактические значения объясняемой и объясняющей переменной $y$ и $x_1$, а $y_i^*$, $x_{i1}^*$, $i=\overline{1,n}$ – их «истинные» значения. Предположим, что переменные $y^*$ и $x_1^*$ связаны линейной функциональной зависимостью:

$$y_i^* = \alpha + \beta x_{i1}^*, \qquad i=\overline{1,n}, \tag{1}$$

где $\alpha$, $\beta$ – неизвестные параметры.

Фактические значения переменных $y$ и $x_1$ связаны с их «истинными» значениями следующими уравнениями:

$$x_{i1} = x_{i1}^* + \varepsilon_i^{(x_1)}, \qquad y_i = y_i^* + \varepsilon_i^{(y)}, \qquad i=\overline{1,n}, \tag{2}$$

где $\varepsilon_i^{(y)}$, $\varepsilon_i^{(x_1)}$, $i=\overline{1,n}$ – случайные ошибки переменных $y$ и $x_1$.

Совокупность уравнений (1) – (2) называется регрессией Деминга. Оценки её параметров находятся следующим образом [5, 6].

Оценка параметра $\beta$ по формуле:

$$\beta^* = \frac{\left(D_y - \frac{D_{x_1}}{\lambda}\right) + \sqrt{\left(D_y - \frac{D_{x_1}}{\lambda}\right)^2 + 4\frac{K_{x_1y}^2}{\lambda}}}{2K_{x_1y}}, \tag{3}$$

где $\lambda = \sigma^2_{\varepsilon^{(x_1)}} / \sigma^2_{\varepsilon^{(y)}}$ – заданное соотношение дисперсий ошибок $\varepsilon^{(y)}$ и $\varepsilon^{(x_1)}$; $D_{x_1}$, $D_y$ – дисперсии переменных $x_1$ и $y$; $K_{x_1y}$ – ковариация.

Оценка параметра $\alpha$ находится по формуле:

$$\alpha^* = \overline{y} - \beta^*\overline{x_1}, \tag{4}$$

где $\overline{y}$ и $\overline{x_1}$ – выборочные средние.

«Истинные» значения переменной $x_1$ вычисляются по формулам:

$$x_{i1}^* = -\frac{\alpha^*\beta^*}{\frac{1}{\lambda} + \left(\beta^*\right)^2} + \frac{\beta^*}{\frac{1}{\lambda} + \left(\beta^*\right)^2} y_i + \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \left(\beta^*\right)^2} x_{i1}, \qquad i=\overline{1,n}. \tag{5}$$

При $\lambda \to 0$ регрессия Деминга принимает вид прямой регрессии $y^* = \alpha^* + \beta^* x_1$, а при $\lambda \to \infty$ – обратной регрессии $x_1^* = -\frac{\alpha^*}{\beta^*} + \frac{1}{\beta^*} y$.

Рассмотрим синтез множественной регрессии и регрессии Деминга, предполагающий двухэтапное оценивание [7]. Для этого введем совокупность детерминированных объясняющих переменных $x_2$, $x_3$, …, $x_m$, которые так же, как и переменная $x_1$, оказывают влияние на $y^*$ и $x_1^*$. Заметим, что для оценивания синтезированной модели предварительно необходимо для заданного соотношения $\lambda$ по формулам (1), (3), (4), (5) найти оценки $\alpha^*$, $\beta^*$, $x_i^*$, $y_i^*$, $i=\overline{1,n}$ регрессии Деминга.

*Этап 1*. С помощью МНК оценивается модель

$$y_i^* = d_0 + d_1 x_{i1} + d_2 x_{i2} + \ldots + d_m x_{im} + \varepsilon_i^{(y^*)}, \qquad i=\overline{1,n}, \tag{6}$$

где $\varepsilon_i^{(y^*)}$, $i=\overline{1,n}$ – случайные ошибки переменных $y^*$; $d_0$, $d_1$, …, $d_m$ – неизвестные параметры.

При $\lambda \to 0$ МНК-оценки модели (6) имеют вид $d_0^* = \alpha^*$, $d_1^* = \beta^*$, $d_2^* = \ldots = d_m^* = 0$, т.е. они совпадают с МНК-оценками парной линейной регрессии, а при $\lambda \to \infty$ МНК-оценки модели (6) совпадают с МНК-оценками множественной линейной регрессии [7].

Пусть для выбранного значения $\lambda$ оцененная с помощью МНК модель (6) имеет вид:

$$y^{**} = d_0^* + d_1^* x_1 + d_2^* x_2 + \ldots + d_m^* x_m, \tag{7}$$

где $y_i^{**}$ – расчетные значения переменной $y^*$.

*Этап 2*. Будем моделировать наблюдаемую переменную $y$ моделью парной линейной регрессии

$$y_i = a + b y_i^{**} + \varepsilon_i^{(y)}, \qquad i=\overline{1,n}, \tag{8}$$

или, с учетом (7), моделью

$$y_i = a + b\left(d_0^* + d_1^* x_{i1} + d_2^* x_{i2} + \ldots + d_m^* x_{im}\right) + \varepsilon_i^{(y)}, \qquad i=\overline{1,n}. \tag{9}$$

Как отмечено в [7], при $\lambda \to 0$ МНК-оценки регрессии (9) составляют $a^* = 0$, $b^* = 1$, $d_0^* = \alpha^*$, $d_1^* = \beta^*$, $d_2^* = \ldots = d_m^* = 0$, где $\alpha^*$, $\beta^*$ – МНК-оценки парной регрессии. А при $\lambda \to \infty$ – $a^* = 0$, $b^* = 1$, $d_0^*$, …, $d_m^*$, где $d_0^*$, …, $d_m^*$ – МНК-оценки множественной регрессии. Таким образом, с ростом соотношения дисперсий ошибок $\lambda$ от 0 до $\infty$ модель (9) будет плавно «трансформироваться» из модели парной во множественную линейную регрессию. Поскольку при этом параметр $\lambda$ принимает на интервале $(0,\infty)$ бесчисленное множество значений, то рассмотренная двухэтапная процедура оценивания синтеза множественной регрессии и регрессии Деминга на практике дает бесчисленное множество различных и не изучавшихся ранее оценок модели (9) в рамках МНК. Таких оценок станет еще больше, если в предложенном синтезе вместо переменной $x_1$ рассмотреть поочередно переменные $x_2$, $x_3$, …, $x_m$. Естественным образом возникает интерес в выборе из этого богатого скрытого множества таких оценок, которые были бы наиболее приемлемы с точки зрения заданных критериев адекватности. Помимо этого, из-за эффекта плавного «перехода» от парной к множественной регрессии, интерес вызывает то, как этот синтез соотносится с известной проблемой отбора информативных регрессоров [9].

**Пример**

В качестве примера рассматривалась задача моделирования грузооборота железнодорожного транспорта на Красноярской железной дороге по данным [9] за 2000 – 2014 гг. Переменными выступали: $y$ – грузооборот, млн. т. км; $x_1$ – динамическая нагрузка, т. км / км; $x_2$ – среднесуточный пробег локомотива, км; $x_3$ – техническая скорость, км / час; $x_4$ – провозная способность железнодорожной линии, млн. т. км.

Сначала с помощью МНК была оценена модель парной линейной регрессии:

$$y^* = -251796 + 6058{,}77 x_1, \tag{10}$$

для которой коэффициент детерминации $R^2 = 0{,}6456$, критерий Дарбина-Уотсона $DW = 0{,}8057$, критерий согласованности поведения $SP = 0$, средняя относительная ошибка аппроксимации $E = 10{,}5971\%$.

Затем была построена модель множественной линейной регрессии:

$$y^{*} = -389084 + 4356{,}14x_1 - 61{,}868x_2 + 1389{,}04x_3 + 228779x_4, \qquad (11)$$

для которой $R^2 = 0{,}9018$, $DW = 1{,}3692$, $SP = 6$, $E = 4{,}4297\,\%$.

После чего с использованием эконометрического пакета Gretl двухэтапная процедура оценивания модели (9) для заданного интервала изменения $\lambda$ была реализована в виде скрипта. Вопрос выбора такого интервала изменения $\lambda$, который бы обеспечивал полную наглядность результатов моделирования, пока остается открытым. В нашем случае этот интервал задавался вручную в виде $\left[10^{-9}; 1{,}71\cdot 10^{-7}\right]$. Затем в скрипте этот промежуток был разбит 169 точками на равные отрезки, и в каждой такой точке были найдены оценки модели (9) и её критерии адекватности. Зависимости оценок параметров модели (9) $\alpha_j = b\cdot d_j^{*}$, $j=\overline{1,4}$ от параметра $\lambda$ представлены на рис. 1 (а) – (г).

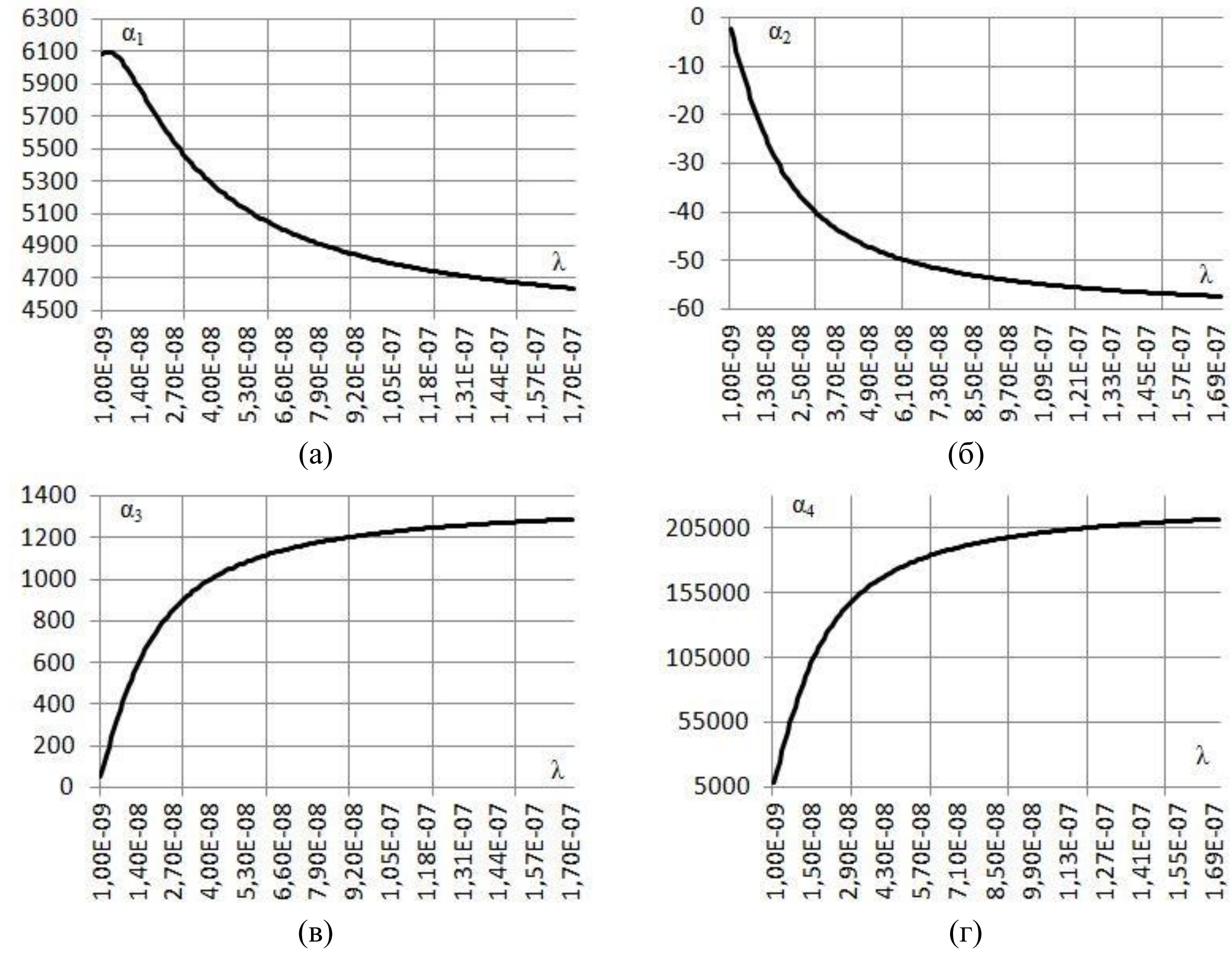


**Рис. 1.** Зависимости оценок модели (9) от параметра $\lambda$

Как видно по рис. 1 (а) при $\lambda = 10^{-9}$ коэффициент $\alpha_1$ модели (9) при переменной $x_1$ практически равен угловому коэффициенту парной линейной регрессии (10). С увеличением параметра $\lambda$ коэффициент $\alpha_1$ даже достигает некоторой точки максимума при $\lambda = 3\cdot 10^{-9}$, а затем происходит его уменьшение, и при $\lambda = 1{,}71\cdot 10^{-7}$ он уже практически не отличается от оценки множественной регрессии (11) при переменной $x_1$. С зависимостями на рис. 1 (б) – (г) ситуация выглядит несколько иначе. При $\lambda = 10^{-9}$ коэффициенты $\alpha_2$, $\alpha_3$ и $\alpha_4$ практически не отличаются от нуля, как и положено для случая однофакторной зависимости $y$ от $x_1$ (10). А с ростом параметра $\lambda$ коэффициенты $\alpha_2$, $\alpha_3$ и $\alpha_4$ плавно начинают стремится к соответствующим оценкам множественной регрессии (11).

Зависимости критериев адекватности модели (9) от параметра $\lambda$ представлены на рис. 2 (а) – (г).

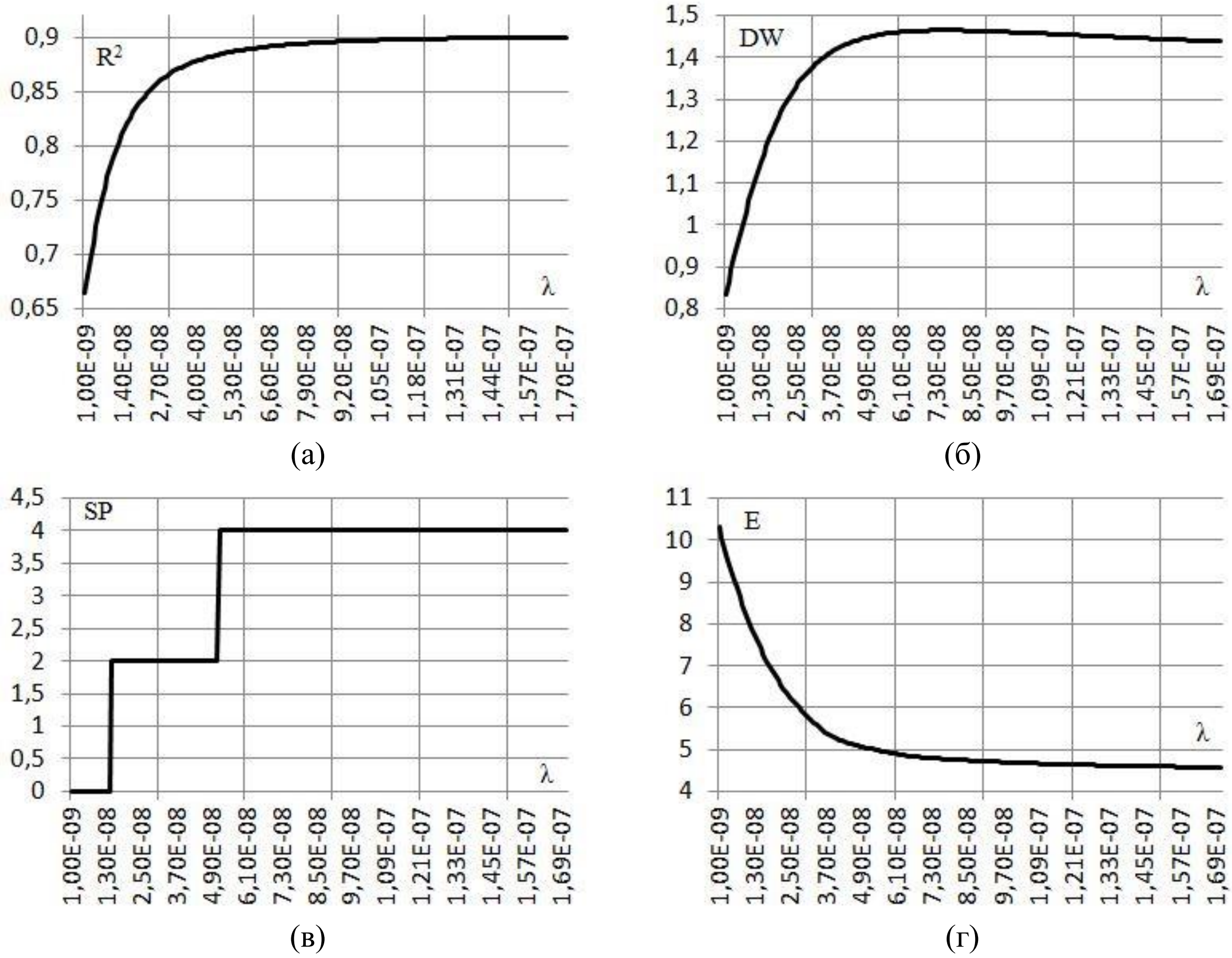


**Рис. 2.** Зависимости критериев адекватности модели (9) от параметра $\lambda$

На рис. 2 при $\lambda = 10^{-9}$ все критерии практически равны соответствующим критериям парной линейной регрессии (10), а при $\lambda = 1,71 \cdot 10^{-7}$ – критериям множественной линейной регрессии (11). Изображенная на рис. 2 (а) зависимость коэффициента детерминация $R^2$ от параметра $\lambda$ согласуется с известной теоремой, утверждающей, что с ростом числа факторов значение коэффициента $R^2$ не убывает.

Получившиеся монотонные зависимости на рис. 2 (а), (в) и (г) говорят о том, что наилучшей по каждому из этих критериев в отдельности является множественная регрессия (11). Чего нельзя сказать о зависимости критерия Дарбина-Уотсона $DW$ от параметра $\lambda$ на рис. 2 (б), имеющей максимум в точке $\lambda = 7,6 \cdot 10^{-8}$. В этой точке значения критериев адекватности составляют: $R^2 = 0,8937$, $DW = 1,4642$, $SP = 4$, $E = 4,7798\%$. Это означает, что существуют такие оценки модели (9), которые снизят значение коэффициента детерминации множественной регрессии (11) с 0,9018 до 0,8937, но при этом повысят значение критерия Дарбина-Уотсона с 1,3692 до 1,4642, т.е. снизится эффект автокорреляции остатков.

Аналогичные зависимости оценок и критериев адекватности были получены и для синтезов, в которых в регрессиях Деминга вместо переменной $x_1$ рассматривались поочередно переменные $x_2$, $x_3$ и $x_4$. Не будем рассматривать абсолютно все полученные результаты, а приведем два наиболее интересных из них.

При использовании в регрессии Деминга объясняющей переменной $x_2$ была получена зависимость коэффициента $\alpha_2$ модели (9) при переменной $x_2$ от параметра $\lambda$, отображенная на рис. 3. Интерес на этом графике вызывает то, что существует такое значение $\lambda$, приближенно равное $\lambda = 3{,}42 \cdot 10^{-5}$, для которого коэффициент $\alpha_2$ равен нулю, т.е. в этой точке четырехфакторная модель становится трехфакторной. Таким образом, предложенный синтез впервые продемонстрировал себя, как новый инструмент для решения задачи отбора наиболее информативных регрессоров, которой в регрессионном анализе уделяется немало внимания [9].

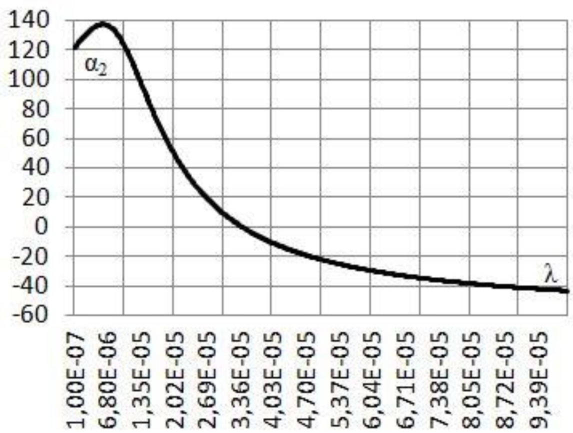


**Рис. 3.** Зависимость $\alpha_2$ от параметра $\lambda$

При использовании в регрессии Деминга объясняющей переменной $x_4$ была получена зависимость критерия согласованности поведения $SP$ модели (9) от параметра $\lambda$, изображенная на рис. 4. Интерес, в отличие от предыдущих случаев, вызывает непредсказуемость поведения этого критерия в зависимости от величины $\lambda$. При этом график указывает на то, что существует такое значение $\lambda$, приближенно равное $\lambda = 7{,}1 \cdot 10^{-12}$, при котором $SP = 10$. В этой точке $R^2 = 0{,}76$, т.е. существуют такие оценки модели (9), которые снизят значение коэффициента детерминации множественной регрессии (11) с 0,9018 до 0,76, но при этом повысят согласованность поведения с 6 до 10.

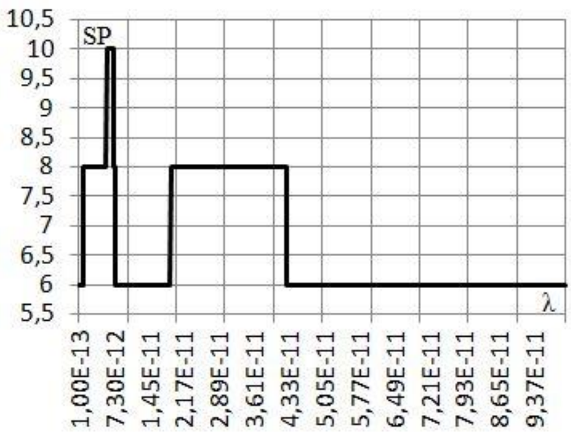


**Рис. 4.** Зависимость $SP$ от $\lambda$

**Заключение**

В работе рассмотрен синтез модели множественной линейной регрессии и регрессии Деминга. На примере моделирования грузооборота железнодорожного транспорта на Красноярской железной дороге для синтезированной регрессии получены зависимости её оценок параметров и критериев адекватности от соотношения дисперсий ошибок переменных. Все эти зависимости демонстрируют, что с ростом соотношения дисперсий ошибок $\lambda$ от 0 до $\infty$ синтезированная модель плавно «трансформируется» из классической модели парной во множественную линейную регрессию. При этом полученные функции могут вести себя непредсказуемым образом, поэтому требуют дальнейших исследований. Показано, что с использованием рассмотренного синтеза можно снижать аппроксимационные качества классической модели множественной линейной регрессии ради повышения некоторых других её важных характеристик. На примере показано, что синтезированная модель может выступать в роли инструмента для решения задачи отбора наиболее информативных регрессоров.

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. – М.: Финансы и статистика, 1985. – 487 с.

2. Draper N.R., Smith H. Applied regression Analysis, 3rd edition. – John Wiley & Sons, 1998. – 736 p.

3. Доугерти К. Введение в эконометрику. – М. : Инфра-М, 2009. – 465 с.

4. Deming W.E. Statistical adjustment of data. – New York: Wiley, 1943. – 273 p.

5. Базилевский М.П. Аналитические зависимости между коэффициентами детерминации и соотношением дисперсий ошибок исследуемых признаков в модели регрессии Деминга // Математическое моделирование и численные методы, 2016. – № 2 (10). – С. 104–116.

6. Базилевский М.П. Аналитические зависимости для некоторых критериев адекватности модели регрессии Деминга // Вестник ИрГТУ. – Иркутск, 2016. – Т. 20 – № 10. – С. 81–89.

7. Базилевский М.П. Синтез модели множественной линейной регрессии и регрессии Деминга // Информационные технологии в моделировании и управлении: подходы, методы, решения: материалы II Всероссийской научной конференции с международным участием. – Тольятти, 2019. – Часть 1. – С. 64–69.

8. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. – Иркутск : Облинформпечать, 1996. – 321 с.

9. Носков С.И., Базилевский М.П. Построение регрессионных моделей с использованием аппарат линейно-булевого программирования. – Иркутск : ИрГУПС, 2018. – 176 с.

## REFERENCES

1. Ajvazyan S.A., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika: Issledovanie zavisimostej* [Applied Statistics: Addiction Research]. Moscow, Finance and Statistics, 1985, 487 p.

2. Draper N.R., Smith H. Applied regression Analysis, 3rd edition. John Wiley & Sons, 1998, 736 p.

3. Dougerti K. *Vvedenie v ehkonometriku* [Introduction in Econometrics]. Moscow, Infra-M, 2009, 465 p.

4. Deming W.E. Statistical adjustment of data. New York: Wiley, 1943, 273 p.

5. Bazilevskij M.P. *Analiticheskie zavisimosti mezhdu koehfficientami determinacii i sootnosheniem dispersij oshibok issleduemyh priznakov v modeli regressii Deminga* [Analytical dependencies between the coefficients of determination and the ratio of the error dispersions of the studied features in the Deming regression model]. *Matematicheskoe modelirovanie i chislennye metody* [Mathematical modeling and numerical methods]. 2016, vol.10, no. 2, pp. 104–116.

6. Bazilevskij M.P. *Analiticheskie zavisimosti dlya nekotoryh kriteriev adekvatnosti modeli regressii Deminga* [Analytical dependencies for some criteria of adequacy of the Deming regression model]. *Vestnik IrGTU* [Bulletin of ISTU]. Irkutsk, 2016, vol. 20, no. 10, pp. 81–89.

7. Bazilevskij M.P. *Sintez modeli mnozhestvennoj linejnoj regressii i regressii Deminga* [Synthesis of the Multiple Linear Regression and Deming Regression Model]. *Informacionnye tekhnologii v modelirovanii i upravlenii: podhody, metody, resheniya: materialy II Vserossijskoj nauchnoj konferencii s mezhdunarodnym uchastiem* [Information technologies in modeling and management: approaches, methods, solutions: materials of the II All-Russian Scientific Conference with international participation]. Tolyatti, 2019, part 1, pp. 64–69.

8. Noskov S.I. *Tehnologija modelirovanija ob'ektov s nestabil'nym funkcionirovaniem i neopredelennost'ju v dannyh* [Modeling technology for objects with unstable operation and data uncertainty]. Irkutsk, RIC GP «Oblinformpechat'» Publ., 1996, 321 p.

9. Noskov S.I., Bazilevskij M.P. *Postroenie regressionnyh modelej s ispol'zovaniem apparat linejno-bulevogo programmirovaniya* [Construction of regression models using linear-boolean programming device]. Irkutsk, IrGUPS, 2018, 176 p.

**Информация об авторах**

*Михаил Павлович Базилевский* – к. т. н., доцент, доцент кафедры «Математика», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: mik2178@yandex.ru

**Authors**

*Mikhail Pavlovich Bazilevskiy* – Ph. D. in Engineering Science, Associate Professor, the Subdepartment of Mathematics, Irkutsk State Transport University, Irkutsk, e-mail: mik2178@yandex.ru

**Для цитирования**

Базилевский М.П. Синтез модели множественной линейной регрессии и регрессии Деминга: исследование зависимостей оценок параметров и критериев адекватности от соотношения дисперсий ошибок переменных // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2019. – №2. – С. 18-25 – Режим доступа: http://ismm-irgups.ru/toma/23-2019, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 19.06.2019)

**For citations**

Bazilevskiy M.P. *Sintez modeli mnozhestvennoy lineynoy regressii i regressii Deminga: issledovaniye zavisimostey otsenok parametrov i kriteriyev adekvatnosti ot sootnosheniya dispersiy oshibok peremennykh* [Synthesis of multiple linear regression and Deming regression model's: investigation the dependences of parameter estimates and adequacy criteria on the ratio of variance error variables] // *Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal* [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2019. No. 2. P. 18-25. Access mode: http://ismm-irgups.ru/toma/23-2019, free. – Title from the screen. – Language Russian, English. [Accessed 19/06/19]