УДК 519.862.6
DOI 10.18413/2411-3808-2019-46-1-117-129

ПРОГНОЗИРОВАНИЕ ГРУЗООБОРОТА ЖЕЛЕЗНОДОРОЖНОГО ТРАНСПОРТА ПО РЕГРЕССИОННЫМ МОДЕЛЯМ С ДЕТЕРМИНИРОВАННЫМИ И СТОХАСТИЧЕСКИМИ ОБЪЯСНЯЮЩИМИ ПЕРЕМЕННЫМИ

PREDICTION OF FREIGHT TURNOVER OF RAILWAY TRANSPORT USING REGRESSION MODELS WITH DETERMINISTIC AND STOCHASTIC EXPLANATORY VARIABLES

М.П. Базилевский
M.P. Bazilevskiy

Иркутский государственный университет путей сообщения,
Россия, 664074, г. Иркутск, ул. Чернышевского, 15

Irkutsk State Transport University,
15 Chernyshevsky St., Irkutsk, 664074, Russia

E-mail: mik2178@yandex.ru

Аннотация
Работа посвящена проблеме прогнозирования по регрессионным моделям с детерминированными и стохастическими переменными. Последние, при наличии только одной объясняющей переменной, принято называть регрессиями Деминга. Кратко рассмотрен метод оценивания таких моделей и известный, основанный не на вероятностной природе, способ интервального прогнозирования по ним. Впервые проведено тестирование последнего на примере интервального прогнозирования грузооборота железнодорожного транспорта. Полученные интервалы оказались такими же надежными, как и доверительные интервалы для классической регрессии. Предложен новый способ получения точечных прогнозов по регрессии Деминга, предполагающий решение задачи выбора такого соотношения дисперсий ошибок переменных, которое минимизирует среднюю абсолютную ошибку прогноза на экзаменующей выборке. Предложенный способ применен для точечного прогнозирования грузооборота железнодорожного транспорта. Найденные прогнозы оказались существенно лучше, чем прогнозы, полученные с помощью классической регрессии с детерминированными переменными.

Abstract
The paper is devoted to the problem of prediction using regression models with deterministic and stochastic variables. The latter, if there is only one explanatory variable, are called Deming regressions. A method for estimating such models and a well-known, not based on probabilistic nature, interval prediction method based on them are briefly considered. For the first time, the latter was tested using the example of interval forecasting of rail freight turnover. The intervals obtained were as reliable as the confidence intervals for the classical regression. A new method for obtaining point forecasts for the Deming regression is proposed, which involves solving the problem of choosing such a ratio of error variances of variables that minimizes the mean absolute error for the examining sample. The proposed method is applied for point forecasting of freight turnover of railway transport. The predictions found turned out to be significantly better than the predictions obtained using classical regression with deterministic variables.

Ключевые слова: регрессионная модель, регрессия Деминга, тренд, прогнозирование, средняя абсолютная ошибка прогноза, грузооборот железнодорожного транспорта.
Keywords: regression model, Deming regression, trend, forecasting, mean absolute error, rail freight turnover.

Введение

В начале XIX века практически одновременно двумя учеными-математиками А.М. Лежандром и К.Ф. Гауссом был изобретен метод наименьших квадратов (МНК), который вызвал к жизни регрессионный анализ [Draper, Smith, 1998]. При этом оценивание неизвестных параметров регрессионной модели проводилось в обязательном предположении, что объясняющие переменные являются детерминированными, т. е. неслучайными. Впервые о том, что объясняющие переменные в регрессии могут содержать случайные ошибки, т. е. носить стохастический характер, было упомянуто в работе Р. Эдкока [Adcock, 1878]. В дальнейшем полученные им результаты стали называть ортогональной регрессией, в которой дисперсии ошибок переменных одинаковы. Год спустя К. Куммель [Kummel, 1879] расширил работу Р. Эдкока, рассмотрев задачу с разными дисперсиями ошибок переменных. Регрессия К. Куммеля стала обобщением как ортогональной регрессии Р. Эдкока, так и классической регрессии без ошибок в объясняющих переменных.

Стоит признать, что в настоящее время регрессионные модели с ошибками в объясняющих переменных не находят такого широкого практического применения, как их детерминированные аналоги, представленные в многочисленной литературе [Аверин и др., 2018; Мельникова, 2018; Московкин, Лю Явэй, 2017; Муноз и др., 2017; Сизьунго Муненге, 2016]. Главным назначением последних является возможность интерпретации их оценок и получение прогнозов. В качестве исключения можно привести работу Э. Деминга [Deming, 1943]. Его книга стала столь популярной в клинической химии, что метод в этих областях получил название «регрессия Деминга». Суть регрессии Деминга в том, что в клинических лабораториях она служит прекрасным инструментом для численного сопоставления новых измерительных методов с существующими. Множество статей по данной тематике можно найти на сайте журнала Clinical Chemistry [Clinical Chemistry, 2019], например, [Ahmad et al., 2018; Taylor et al., 2017; Lewis et al., 2017; Kvisvik et al., 2017]. В работе [Jensen, Kjelgaard-Hansen, 2006] регрессия Деминга применена для исследования нового метода измерения аланинаминотрансферазы в крови собак, в [Wu, Yu, 2018] – для исследования поглощения света от массовой концентрации углерода, в [Dhanoa et al., 2016] – для исследования методов измерения арсенат-иона в природной речной воде.

Целью данной работы является тестирование старых и разработка новых способов точечного и интервального прогнозирования по регрессии Деминга, а также сопоставление прогнозов по моделям с детерминированными и стохастическими объясняющими переменными на примере моделирования грузооборота железнодорожного транспорта России.

Регрессия Деминга

Пусть y_i, x_i, $i=\overline{1,n}$ – наблюдаемые значения объясняемой переменной y и объясняющей переменной x, а y_i^*, x_i^*, $i=\overline{1,n}$ – их «истинные» значения, которые мы не можем наблюдать. Предположим, что «истинные» значения переменных y и x связаны линейной функциональной зависимостью:

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

где α, β – неизвестные параметры.

Наблюдаемые значения переменных y и x связаны с «истинными» значениями и случайными отклонениями следующими уравнениями:

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

$$x_i = x_i^* + \varepsilon_i^{(x)}, \qquad i=\overline{1,n}, \tag{3}$$

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

Совокупность уравнений (1)–(3) называется регрессией Деминга. Для оценивания её неизвестных параметров α и β требуется минимизировать функционал:

$$S = \frac{1}{\lambda}\sum_{i=1}^{n}\left(x_i - x_i^*\right)^2 + \sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i^*\right)^2 \to \min, \tag{4}$$

где $\lambda = \sigma^2_{\varepsilon^{(x)}} / \sigma^2_{\varepsilon^{(y)}}$ – соотношение дисперсий ошибок $\varepsilon^{(y)}$ и $\varepsilon^{(x)}$.

Оценка параметра b находится из решения квадратного уравнения:

$$K_{xy}\beta^2 - \left(D_y - \frac{D_x}{\lambda}\right)\beta - \frac{K_{xy}}{\lambda} = 0, \tag{5}$$

где D_x, D_y – дисперсии переменных x и y, K_{xy} – ковариация.

Условию задачи (4) удовлетворяет только один из корней уравнения (5):

$$\beta^* = \frac{\left(D_y - \frac{D_x}{\lambda}\right) + \sqrt{\left(D_y - \frac{D_x}{\lambda}\right)^2 + 4\frac{K_{xy}^2}{\lambda}}}{2K_{xy}}. \tag{6}$$

Оценка параметра α находится по формуле:

$$\alpha^* = \bar{y} - \beta^*\bar{x}, \tag{7}$$

где $\bar{y}$ и $\bar{x}$ – выборочные средние.

«Истинные» значения переменной x вычисляются по формулам:

$$x_i^* = -\frac{\alpha^*\beta^*}{\frac{1}{\lambda} + \left(\beta^*\right)^2} + \frac{\beta^*}{\frac{1}{\lambda} + \left(\beta^*\right)^2} y_i + \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \left(\beta^*\right)^2} x_i, \qquad i = \overline{1,n}. \tag{8}$$

Тогда оцененная модель будет иметь вид:

$$y^* = \alpha^* + \beta^* x^*. \tag{9}$$

В зависимости от соотношения дисперсий ошибок исследуемых признаков λ выделяют следующие частные случаи регрессии Деминга:

- прямая регрессия, если $\lambda \to 0$;
- ортогональная регрессия, если $\lambda = 1$;
- диагональная регрессия, если $\lambda = \frac{D_x}{D_y}$;
- обратная регрессия, если $\lambda \to \infty$.

В работе [Базилевский, 2016б] предложен простой и основанный не на вероятностной природе способ интервального прогнозирования по моделям множественной линейной регрессии со стохастическими переменными. В работе [Базилевский, 2017] этот способ рассмотрен для регрессии Деминга (1)–(3). Приведем его краткое описание.

Пусть по оцененному уравнению (9) требуется получить «истинное» значение переменной y, если прогнозное значение зашумленной переменной x равно x_0. Понятно, что простая подстановка значения x_0 в соотношение (9) возможна, но не совсем корректна. Для этого необходимо знать «истинное» значение объясняющей переменной x_0^*. Но зависимость «истинных» значений x_i^* от наблюдаемых x_i нам неизвестна, а может быть,

она и вовсе отсутствует. Тогда возможно получение интервального прогноза согласно следующей процедуре.

1. Для объясняющей переменной x определяется минимальное и максимальное значение ошибки аппроксимации: $\varepsilon_{\min}^{(x)} = \min(x - x^*)$, $\varepsilon_{\max}^{(x)} = \max(x - x^*)$.

2. Определяется интервал расчетных значений объясняющей переменной x: $x^* \in \left[\underline{x}; \overline{x}\right]$, где $\underline{x} = x_0 - \varepsilon_{\max}^{(x)}$, $\overline{x} = x_0 - \varepsilon_{\min}^{(x)}$.

3. Находится нижняя граница интервального прогноза $\underline{y}$. Для этого в уравнение (9) необходимо вместо переменной x^* подставить найденное на предыдущем шаге значение $\underline{x}$ или $\overline{x}$ по такому правилу: если коэффициент уравнения $\beta^* > 0$, то вместо значения x^* подставляется $\underline{x}$, а если $\beta^* < 0$, то $\overline{x}$.

4. По аналогии с предыдущим шагом, находится верхняя граница интервального прогноза $\overline{y}$. При этом подстановка осуществляется по такому правилу: если коэффициент уравнения $\beta^* > 0$, то вместо значения x^* подставляется $\overline{x}$, а если $\beta^* < 0$, то $\underline{x}$.

Следует отметить, что для получения более надежного прогноза на первом шаге вместо минимального и максимального значения ошибки аппроксимации можно определять максимальное из абсолютных значений ошибок, т. е. $\varepsilon_{abs}^{(x)} = \max\left|x - x^*\right|$. Тогда на втором шаге интервал расчетных значений объясняющей переменной $x^* \in \left[\underline{x}; \overline{x}\right]$, где $\underline{x} = x_0 - \varepsilon_{abs}^{(x)}$, $\overline{x} = x_0 + \varepsilon_{abs}^{(x)}$. Третий и четвертый шаги остаются прежними. При этом прогнозный интервал станет шире.

Таким образом, ширина интервального прогноза для переменной y зависит от дисперсии ошибок $\varepsilon^{(x)}$. Чем больше эта дисперсия, тем шире интервал, и наоборот. Если $\sigma^2_{\varepsilon^{(x)}} \to 0$, то $\lambda \to 0$, поэтому интервальный прогноз регрессии Деминга в данном случае станет точеным прогнозом для классической прямой регрессии.

К сожалению, рассмотренный способ позволяет получать лишь интервальные прогнозы, что вряд ли может полностью удовлетворить исследователя, который занимается прогнозированием конкретного социально-экономического процесса или явления. В этом случае можно с помощью вспомогательных переменных попытаться смоделировать зависимость между переменными x и x^*. Заметим, что эти переменные не просто тесно коррелируют друг с другом, но и, в силу формулы (8), связаны между собой линейной функциональной зависимостью через переменную y. Т. е. в регрессии Деминга точно определить истинные значения переменных x^* и y^* можно только через известные значения обеих переменных x и y. Из этого следует, что регрессия Деминга отлично подходит, например, для решения задач определения истинных координат объекта (морского судна, летательного аппарата) по имеющимся зашумленным координатам системы навигации GPS. Но для прогнозирования переменной y использовать уравнение (8) не представляется возможным.

Дальнейшее описание будем проводить в предположении, что y_t и x_t, $t = \overline{1,n}$ – временные ряды. Пусть при заданном соотношении дисперсии ошибок λ по этим временным рядам найдены оценки регрессии Деминга α^* и β^*, а также «истинные» значения x_t^*. Если с помощью МНК попробовать оценить параметры модели парной линейной регрессии $x_t = a + bx_t^* + \varepsilon_t$, то её оценки, согласно соотношениям (3), окажутся $a^* = 0$, $b^* = 1$. Такая регрессия не несет никакой новой информации о соотношении между переменными x_t и x_t^*. Поэтому для получения прогнозных значений переменной y_t будем

моделировать зависимость между временными рядами x_t и x_t^* с помощью трендовых моделей вида

$$x_t = bx_t^* + f(t) + \varepsilon_t, \qquad t = \overline{1,n}, \tag{10}$$

где $f(t) = \sum_{i=0}^{m} c_i t^i$.

Полученная с помощью МНК оценка b^* модели (10) при переменной x_t может быть отлична от 1, что противоречит соотношениям (3). Поэтому справедливее моделировать не саму переменную x_t, а её ошибку $\varepsilon_t^{(x)} = x_t - x_t^*$:

$$x_t - x_t^* = f(t) + \varepsilon_t, \qquad t = \overline{1,n}. \tag{11}$$

Опустив ошибку регрессии в (11), выразим переменную x_t^* и подставим её в (9). Тогда пригодная для прогнозирования регрессия Деминга примет вид:

$$\begin{cases} x_t^* = x_t - f(t), \\ y_t^* = \alpha^* + \beta^* x_t - \beta^* f(t). \end{cases} \tag{12}$$

При построении модели (12) для прогнозирования следует придерживаться следующих рекомендаций. Во-первых, качество самой регрессии Деминга (1)–(3) для заданного соотношения λ по возможности должно быть достаточно высоким. Оценивать это качество можно, например, с помощью коэффициентов детерминации [Базилевский, 2016а]. Во-вторых, высоким качеством аппроксимации должна обладать трендовая модель (11), т. е. необходимо очень точно «разгадать» закономерность изменения ошибок переменной x от времени t. Понятно, что шансов на успех для этого будет больше, если критерий Дарбина – Уотсона регрессии Деминга (1)–(3) будет близок к нулю, т. е. при наличии положительной автокорреляции остатков, которая проявляется в чередовании зон положительных и отрицательных остатков. Если критерий Дарбина – Уотсона регрессии Деминга (1)–(3) будет равен двум, то автокорреляция в остатках отсутствует, а значит, остатки чередуются случайным образом, поэтому вряд ли удастся добиться высокого качества регрессии (11). В-третьих, поскольку прогнозирование по нелинейным трендовым моделям чрезвычайно опасно ввиду существования высокой вероятности совершения ошибки, то при построении модели (12) следует проверять её прогностические способности с помощью соответствующего критерия. В качестве такого критерия может выступать, например, средняя абсолютная ошибка прогноза MAE, рассчитанная по экзаменующей выборке по формуле

$$MAE = \sum_{t=\tau+1}^{n} \left| y_t - \alpha^* - \beta^* x_t + \beta^* f(t) \right|, \tag{13}$$

где $1, \ldots, \tau$ – номера обучающей выборки, а $\tau+1, \ldots, n$ – номера экзаменующей выборки.

В приведенной выше схеме считалось, что соотношение дисперсий ошибок λ известно, поэтому оценки регрессии Деминга (1)–(3) и трендовой модели (11) так же, как и их критерии адекватности, определялись единственным образом. Варьирование этого соотношения позволяет получить бесчисленное множество других оценок тех же моделей и их критериев, в котором классическая прямая регрессия является лишь одним из его элементов при $\lambda \to 0$. Поскольку для выбора соотношения λ не существует каких-либо универсальных правил, то возникает задача определения такого значения λ, которое доставляет минимум или максимум заданному критерию адекватности. Применительно к задаче построения модели (12) возникает вопрос: при каком значении параметра λ значение средней абсолютной ошибки прогноза MAE будет минимальным? Для ответа на

этот вопрос достаточно разбить интервал $\lambda \in (0,\infty)$ заданным количеством точек. Затем для каждой точки по формулам (6)–(8) найти оценки регрессии Деминга, оценить с помощью МНК трендовую модель (11) и среднюю абсолютную ошибку прогноза по формуле (13). Затем из этих точек выбрать такую, для которой значение *MAE* минимально. Данный подход призван повысить надежность прогнозных расчетов по модели (12).

Стоит отметить, что для моделирования ошибок вместо трендовых моделей (11) можно использовать, например, авторегрессионные модели. А еще лучше строить зависимости ошибок от других детерминированных переменных, организуя при этом выбор наилучшей спецификации модели.

Моделирование грузооборота железнодорожного транспорта

В таблице 1 приведена динамика двух показателей за 1990–2018 годы:
y – грузооборот железнодорожного транспорта России (млрд т км) [Хусаинов, 2017; Погрузка и грузооборот по годам: с 1988 по 2018];
x – ВВП России (млрд долл., в ценах 1990 г.) [ВВП России в 2019 году].

Таблица 1
Table 1

Динамика ВВП и грузооборота железнодорожного транспорта России
Dynamics of GDP and rail freight traffic in Russia

Год	y	x	Год	y	x
1990	2523	570,4	2005	1858	516,2
1991	2325,9	541,9	2006	1951	558,3
1992	1967,1	463,3	2007	2090,3	606
1993	1607,7	423	2008	2116,2	637,8
1994	1195,5	369,3	2009	1865,3	587,9
1995	1213,7	354,1	2010	2011,3	614,4
1996	1131,3	341,3	2011	2127,8	640,6
1997	1100,3	346,1	2012	2222,4	662,6
1998	1019,5	327,6	2013	2196,2	671,3
1999	1204,5	348,4	2014	2298,6	675,3
2000	1373,2	383,4	2015	2304,8	649,6
2001	1433,6	402,9	2016	2342,6	656,6
2002	1508,8	422	2017	2491,4	666,5
2003	1669	452,8	2018	2596,4	679,2
2004	1802	485,3			

Отметим, что моделирование грузооборота по данным из таблицы 1 за период с 1990 по 2013 г. уже проводилось в работах [Базилевский, Гефан, 2016а; Базилевский, Гефан, 2016б]. В них было установлено, что коэффициент корреляции между переменными y и x составляет 0,91, что говорит о высокой положительной корреляции между данными величинами, а соответствующее уравнение парной линейной регрессии имеет вид

$$y^* = 47,279 + 3,443x. \quad (14)$$

Коэффициент детерминации модели (14) $R^2 = 0,828$.

Оценив те же характеристика за период с 1990 по 2018 г., было установлено, что коэффициент корреляции между переменными y и x составляет 0,933, а уравнение регрессии имеет вид:

$$y^* = -1,462 + 3,5598x, \quad (15)$$

с коэффициентом детерминации $R^2 = 0,87$. Таким образом, с течением времени степень линейной зависимости между грузооборотом y и ВВП x лишь усилилась, что подтверждает корректность выбора формы связи между данными величинами.

Затем были сформулированы следующие задачи.

Задача 1. Сравнить точечные и интервальные прогнозы, полученные по линейной модели (14) и регрессии Деминга.

Задача 2. Для модели (12) подобрать такое значение параметра λ, при котором средняя абсолютная ошибка прогноза MAE будет минимальна.

Для решения поставленных задач исходная выборка была разделена на обучающую (с 1990 по 2013 г.) и экзаменующую (с 2014 по 2018 г.). Все представленные далее модели построены по обучающей выборке, а экзаменующая выборка использована для оценки их прогностических способностей.

Задача 1. Используя эконометрический пакет Gretl, по линейной регрессии (14) были получены точечные прогнозы, а также интервальные для среднего (таблица 2) и наблюдаемого (таблица 3) значения переменной y. При этом доверительный интервал для наблюдаемого значения переменной y шире, чем для среднего. В таблицах 2 и 3 в столбце «Факт» приведены фактические значения переменной y, «Расчет» – точечный прогноз переменной y, «Левая» и «Правая» – левая и правая граница интервального прогноза.

Таблица 2
Table 2

Прогнозы по линейной регрессии для среднего значения переменной y
Linear regression predictions for average variable y

Год	Факт	Расчет	Левая	Правая
2014	2298,6	2372,5	2220,7	2524,3
2015	2304,8	2284	2147,1	2420,9
2016	2342,6	2308,1	2167,3	2449,0
2017	2491,4	2342,2	2195,6	2488,8
2018	2596,4	2385,9	2231,9	2540,0

Таблица 3
Table 3

Прогнозы по линейной регрессии для наблюдаемого значения переменной y
Linear regression predictions for observed variable y

Год	Факт	Расчет	Левая	Правая
2014	2298,6	2372,5	1956	2789
2015	2304,8	2284	1872,7	2695,3
2016	2342,6	2308,1	1895,5	2720,8
2017	2491,4	2342,2	1927,6	2756,8
2018	2596,4	2385,9	1968,6	2803,3

Средняя абсолютная ошибка прогноза MAE для линейной регрессии (14) составила 97,758. По таблице 2 видно, что фактическое значение переменной y не попало в доверительные интервалы только для 2017 и 2018 г., а в таблице 3 фактическое значение попадает в них всегда.

Затем была построена регрессия Деминга диагонального вида, т. е. при $\lambda = \frac{D_x}{D_y} = 0,0698$. Её уравнение имеет вид:

$$y^* = -119,118 + 3,7838x^*. \quad (16)$$

Коэффициент детерминации модели (16) как по переменной y, так и по переменной x составляет 0,955. По отношению к линейной регрессии (14) такой результат можно интерпретировать следующим образом: ухудшив качество линейной регрессии по переменной x на 4,5%, её качество по переменной y улучшится на 12,7%.

Как уже отмечалось выше, получение точечных прогнозов по модели (16) не совсем корректно, поскольку неизвестно «истинное» значение переменной x^*. Это значит, что, подставив в уравнение (16) фактические значения ВВП x за 2014–2018 гг., мы должны получить существенные расхождения между фактическими и расчетными значениями для грузооборота y. Однако точечные прогнозы по регрессии (16) оказались весьма неплохими (таблица 4).

Таблица 4
Table 4

Точечные прогнозы по регрессии Деминга
Deming regression point predictions

Год	Факт	Расчет
2014	2298,6	2436,1
2015	2304,8	2338,8
2016	2342,6	2365,3
2017	2491,4	2402,8
2018	2596,4	2450,8

Средняя абсолютная ошибка прогноза MAE для регрессии Деминга (16) составила 85,686. Это означает, что в равных условиях прогностические способности регрессии Деминга (16) оказались лучше, чем линейной регрессии (14).

Для получения интервальных прогнозов по регрессии Деминга (16) было найдено минимальное и максимальное значение ошибки аппроксимации $\varepsilon_{\min}^{(x)} = -63,934$, $\varepsilon_{\max}^{(x)} = 31,725$. Найденные интервальные прогнозы представлены в таблице 5. В ней в столбце «Факт» приведены фактические значения переменной y, «x левая» и «x правая» – левая и правая граница переменной x, «y левая» и «y правая» – левая и правая граница переменной y.

Таблица 5
Table 5

Интервальные прогнозы по регрессии Деминга
Interval forecasts for Deming regression

Год	Факт	x левая	x правая	y левая	y правая
2014	2298,6	643,6	739,2	2316,0	2678,0
2015	2304,8	617,9	713,5	2218,8	2580,8
2016	2342,6	624,9	720,5	2245,3	2607,2
2017	2491,4	634,8	730,4	2282,8	2644,7
2018	2596,4	647,5	743,1	2330,8	2692,8

Из таблицы 5 следует, что фактическое значение переменной y не попало в прогнозный интервал только для 2014 г.

После этого для получения более надежного прогноза было определено максимальное из абсолютных значений ошибок $\varepsilon_{abs}^{(x)} = 63,934$. Найденные интервальные прогнозы представлены в таблице 6.

Таблица 6
Table 6

Интервальные прогнозы повышенной надежности по регрессии Деминга
Interval forecasts of increased reliability in Deming regression

Год	Факт	x левая	x правая	y левая	y правая
2014	2298,6	611,4	739,2	2194,2	2678,0
2015	2304,8	585,7	713,5	2096,9	2580,8
2016	2342,6	592,7	720,5	2123,4	2607,2
2017	2491,4	602,6	730,4	2160,9	2644,7
2018	2596,4	615,3	743,1	2208,9	2692,8

По таблице 6 видно, что все фактические значения переменной y попали в прогнозные интервалы. Таким образом, предложенный в работе [Базилевский, 2016б] способ интервального прогнозирования ничем не хуже на практике, чем способ доверительных интервалов для классических моделей без ошибок в объясняющей переменной.

Задача 2. В пакете Gretl был разработан скрипт, который оценивает множество регрессий Деминга для заданного интервала соотношений дисперсий ошибок λ. Для моделирования грузооборота по исходным данным был выбран интервал $(0,1)$. Затем этот интервал разбивался точками с шагом 0,001, и в каждой точке оценивались параметры регрессии Деминга и определялись её критерии адекватности. После чего строились графики зависимостей этих критериев от величины λ и идентифицировалось оптимальное значение.

Сначала в регрессии Деминга ошибка $\varepsilon_t^{(x)} = x_t - x_t^*$ вообще не моделировалась с помощью тренда (11), т. е. расчетные по модели значения переменной y на экзаменующей выборке вычислялись подстановкой фактических значений переменной x в уравнение (9). Полученная в результате зависимость средней абсолютной ошибки прогноза MAE от величины λ представлена на рис. 1. По нему видно, что критерий MAE для прямой регрессии (14) при $\lambda \to 0$, как и ранее, составляет 97,758. При $\lambda = 0,032$ критерий MAE достигает своего наименьшего значения 81,712, что меньше соответствующего значения 85,686 для диагональной регрессии (16). Оцененная при $\lambda = 0,032$ регрессия Деминга имеет вид:

$$y^* = -52,858 + 3,648x^*. \qquad (17)$$

Этот пример демонстрирует, что использовать уравнение (9) для точечного прогнозирования по указанной методике всё же можно. Если регрессия Деминга (1)–(3) оценена при $\lambda \to 0$, то допустимо незначительное снижение качества модели (3), выражаемое коэффициентом детерминации R_x^2, ради повышения прогностических способностей, выражаемых критерием MAE. В нашем случае при $\lambda = 0,032$ коэффициент детерминации $R_x^2 = 0,983$. Таким образом, снизив качество линейной регрессии по переменной x на 1,7%, удалось уменьшить значение критерия MAE с 97,758 до 81,712, т. е. повысить прогностические способности модели.

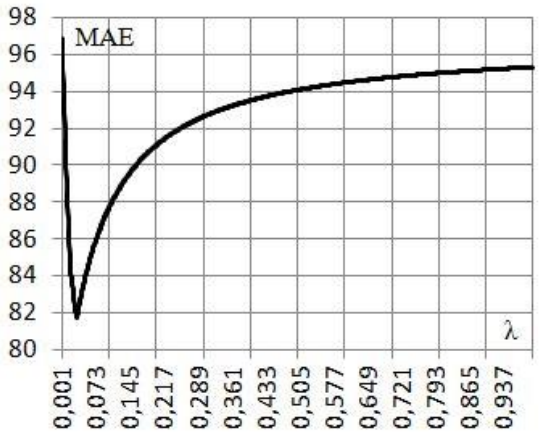


Рис. 1. Зависимость MAE от параметра λ для регрессии Деминга (без моделирования ошибки)
Fig. 1. Dependence MAE on parameter λ for Deming regression (without error modeling)

Затем в регрессии Деминга ошибка $\varepsilon_t^{(x)} = x_t - x_t^*$ моделировалась с помощью полиномиальных трендов: $f_1(t) = c_0 + c_1 t$, $f_2(t) = c_0 + c_1 t + c_2 t^2$, $f_3(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3$, $f_4(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3 + c_4 t^4$. Графики зависимостей коэффициентов детерминации трендов R^2 и средних абсолютных ошибок прогноза MAE от величины λ представлены на рис. 2(а)-(з). Зависимости на рис. 2(а), 2(в), 2(д), 2(ж) демонстрируют, что чем меньше величина λ, тем сложнее адекватно моделировать ошибки $\varepsilon_t^{(x)} = x_t - x_t^*$. Упрощает задачу повышение степени полинома. Графики на рис. 2(б) и 2(е) показывают, что зависимости критерия MAE от величины λ не всегда достигают экстремума. В этом случае для прогнозирования необходимо воспользоваться классическими линейными регрессиями. На рис. 2(г) и 2(з) зависимости имеют экстремумы. В этом случае для прогнозирования необходимо пользоваться соответствующими регрессиями Деминга.

Наименьшее значение 27,143 критерий MAE достигает при $\lambda = 0,028$ на рис. 2(з). Этому результату соответствует регрессия Деминга вида:

$$y^* = -43,2349 + 3,6285x^*, \tag{18}$$

с коэффициентами детерминации $R_x^2 = 0,986$ и $R_y^2 = 0,907$.

Трендовая модель для ошибки $\varepsilon_t^{(x)} = x_t - x_t^*$:

$$x - x^* = -67,716 + 26,6859t - 3,4063t^2 + 0,1729t^3 - 0,0029786t^4, \tag{19}$$

с коэффициентом детерминации $R^2 = 0,898834$.

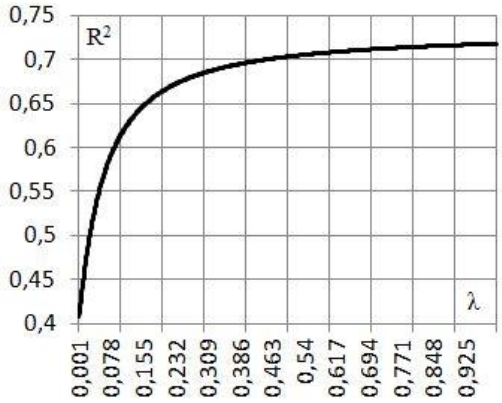
Тогда пригодная для прогнозирования регрессия Деминга имеет вид:

$$\begin{cases} x_t^* = x_t + 67,716 - 26,686t + 3,4063t^2 - 0,1729t^3 + 0,0029786t^4, \\ y_t^* = 3,6285x_t + 202,473 - 96,830t + 12,3598t^2 - 0,62737t^3 + 0,010808t^4. \end{cases} \tag{20}$$

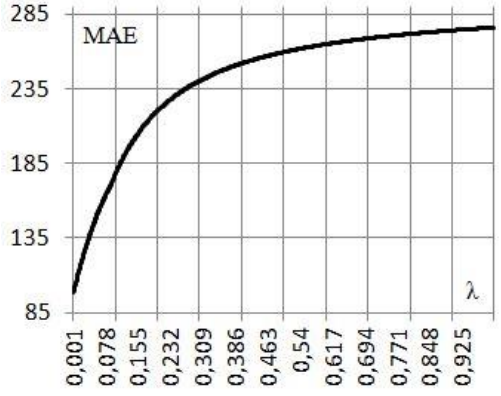
Для сравнения, оцененная с помощью МНК модель множественной линейной регрессии без ошибок в переменной x имеет вид

$$y^* = 108,818 + 4,735x - 247,479t + 36,042t^2 - 2,095t^3 + 0,039t^4, \tag{21}$$

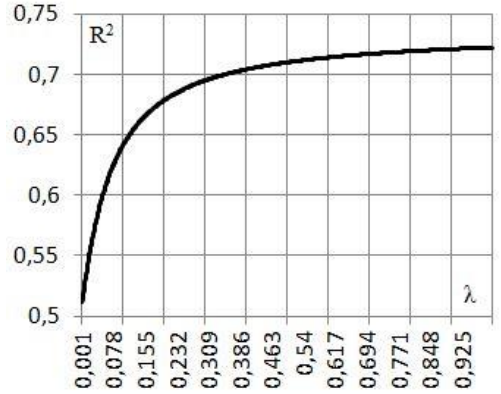
с коэффициентом детерминации $R^2 = 0,984$, что больше, чем для модели (19). Однако средняя абсолютная ошибка прогноза MAE для регрессии (21) составляет 257,28, т. е. по прогностическим способностям она значительно проигрывает модели (20), для которой $MAE = 27,143$.



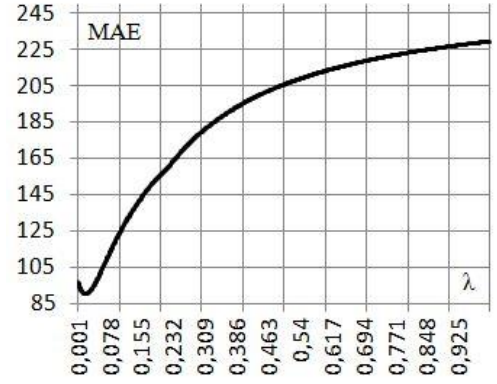
(а) зависимость R^2 от λ для $f_1(t)$



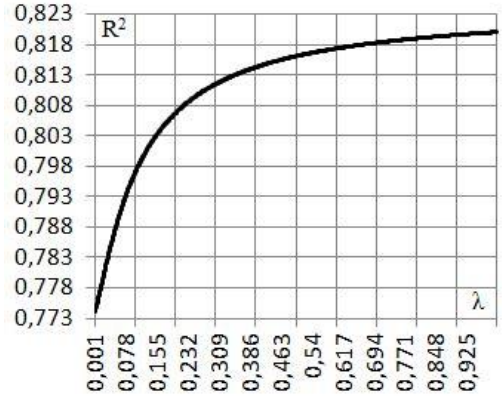
(б) зависимость MAE от λ для $f_1(t)$



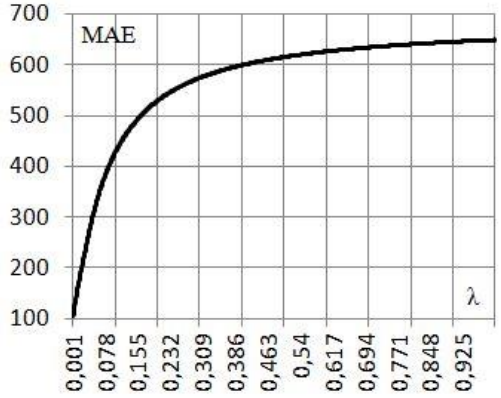
(в) зависимость R^2 от λ для $f_2(t)$



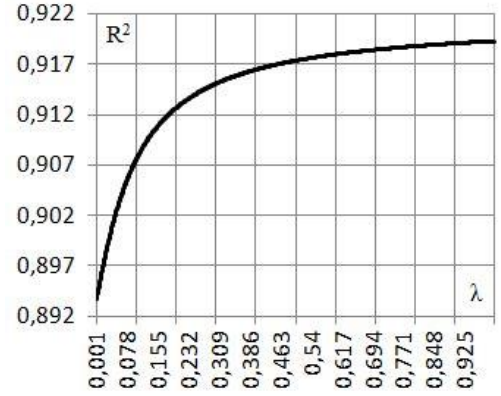
(г) зависимость MAE от λ для $f_2(t)$



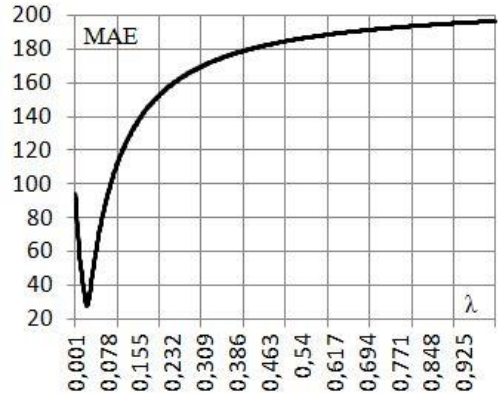
(д) зависимость R^2 от λ для $f_3(t)$



(е) зависимость MAE от λ для $f_3(t)$



(ж) зависимость R^2 от λ для $f_4(t)$



(з) зависимость MAE от λ для $f_4(t)$

Рис. 2. Зависимости MAE от параметра λ для регрессии Деминга (с моделированием ошибки)

Fig. 2. Dependencies MAE on parameter λ for Deming regression (with error modeling)

Заключение

В работе предложен новый способ получения точечных прогнозов по регрессии Деминга, предполагающий решение задачи выбора такого соотношения дисперсий ошибок переменных λ, которое минимизирует среднюю абсолютную ошибку прогноза на экзаменующей выборке. При этом прогностические способности регрессии Деминга можно повысить, если моделировать ошибки в объясняющей переменной X. Разработанный способ был применен для точечного прогнозирования грузооборота железнодорожного транспорта. Найденные прогнозы оказались существенно лучше, чем прогнозы, полученные с помощью классической регрессии без ошибок в переменных. Также впервые проведено тестирование основанного не на вероятностной природе способа получения интервальных прогнозов по регрессии Деминга. Полученные интервалы оказались такими же надежными, как и доверительные интервалы для классической регрессии. Таким образом, за счет возможности варьирования соотношения дисперсий ошибок переменных λ, регрессия Деминга представляет собой гораздо более гибкий и поддающийся контролю инструмент для прогнозирования, нежели классическая регрессия.

Список литературы
References

1. Аверин Г.В., Звягинцева А.В., Швецова А.А. 2018. О подходах к предсказательному моделированию сложных систем. Научные ведомости БелГУ. Сер. Экономика. Информатика, 45(1): 140–148.

Averin G.V., Zviagintseva A.V., Shvetsova A.A. 2018. On approaches to predictive modeling of complex system. Bulletin of BSU. Ser. Economics. Informatics, 45(1): 140–148 (in Russian).

2. Базилевский М.П. 2016. Аналитические зависимости между коэффициентами детерминации и соотношением дисперсий ошибок исследуемых признаков в модели регрессии Деминга. Математическое моделирование и численные методы, 10(2): 104–116.

Bazilevskiy M.P. 2016. Analytical dependence between the coefficients of determination and the ratio of error variances of investigated signs in Deming regression model. Mathematical modeling and numerical methods, 10(2): 104–116 (in Russian).

3. Базилевский М.П. 2016. Численный метод оценивания параметров линейной модели множественной регрессии со стохастическими переменными. Современные технологии. Системный анализ. Моделирование, 52(4): 121–126.

Bazilevskiy M.P. 2016. A numerical method for estimating the parameters of a linear model of multiple regression with stochastic variables. Modern technologies. System analysis. Modeling, 52(4): 121–126 (in Russian).

4. Базилевский М.П. 2017. Об одной методике прогнозирования по эконометрическим моделям со стохастическими переменными. Международный журнал экспериментального образования, 3-1: 53.

Bazilevskiy M.P. 2017. On one method of forecasting for econometric models with stochastic variables. International Journal of Experimental Education, 3-1: 53 (in Russian).

5. Базилевский М.П., Гефан Г.Д. 2016. Моделирование грузооборота железнодорожного транспорта в зависимости от валового внутреннего продукта России. Труды седьмой Международной научно-практической конференции «Транспортная инфраструктура Сибирского региона», 1: 305–309.

Bazilevskiy M.P., Gefan G.D. 2016. Modeling of railway freight turnover depending on the gross domestic product of Russia. Proceedings of the Seventh International Scientific and Practical Conference «Transport Infrastructure of the Siberian Region», 1: 305–309 (in Russian).

6. Базилевский М.П., Гефан Г.Д. 2016. Проблема автокорреляции остатков регрессии на примере моделирования грузооборота железнодорожного транспорта по данным временных рядов. Современные технологии. Системный анализ. Моделирование, 49(1): 141–147.

Bazilevskiy M.P., Gefan G.D. 2016. The problem of autocorrelation of regression residues by the example of modeling the freight turnover of railway transport according to time series data. Modern technologies. System analysis. Modeling, 49(1): 141–147 (in Russian).

7. ВВП России в 2019 году. Available at: http://fincan.ru/articles/56_vvp-rossii-v-2019-godu/.

Russia's GDP in 2019 Available at: http://fincan.ru/articles/56_vvp-rossii-v-2019-godu/ (in Russian).

8. Мельникова О.А. 2018. Модель прогнозирования потребности в непродовольственных товарах: на примере лекарственных средств. Научные ведомости БелГУ. Сер. Экономика. Информатика, 45(1): 86–92.

Melnikova O.A. 2018. Model of forecasting the need for non-food products: on the example of drugs. Bulletin of BSU. Ser. Economics. Informatics, 45(1): 86–92 (in Russian).

9. Московкин В.М., Лю Явэй. 2017. Методология оценки региональной публикационной активности и цитируемости на примере университетов центрального федерального округа Российской Федерации. Научные ведомости БелГУ. Сер. Экономика. Информатика, 42(9): 42–52.

Moskovkin V.M., Liu Yawei. 2017. Methodology for assessing regional publication activity and citation: a case study of the central federal district universities of the Russian Federation. Bulletin of BSU. Ser. Economics. Informatics, 42(9): 42–52 (in Russian).

10. Муноз А.Л., Ваганова О.В., Флигинских Т.Н. 2017. Сравнительный анализ динамики иностранных инвестиций и их влияние на экономический рост региона (на примере субъектов ЦФО РФ). Научные ведомости БелГУ. Сер. Экономика. Информатика, 44(23): 5–15.

Munoz A.L., Vaganova O.V., Fliginskikh T.N. 2017. Comparative analysis of the dynamics of foreign investment and their impact on the economic growth of the region (on the example of the subjects of the central federal district of the Russian Federation). Bulletin of BSU. Ser. Economics. Informatics, 44(23): 5–15 (in Russian).

11. Погрузка и грузооборот по годам: с 1988 по 2018. Available at: https://f-husainov.livejournal.com/626128.html.

Loading and freight turnover by year: from 1988 to 2018. Available at: https://f-husainov.livejournal.com/626128.html (in Russian).

12. Сизьунго Муненге. 2016. Анализ регрессионной взаимосвязи между количеством объектов российской региональной инновационной и университетской инфраструктуры и региональными макроэкономическими показателями. Научные ведомости БелГУ. Сер. Экономика. Информатика, 40(23): 30–39.

Sizyoongo Munenge. 2016. Analysis of regression relationship between the number of objects of the russian regional innovation and the university infrastructure and regional macroeconomic indicators. Bulletin of BSU. Ser. Economics. Informatics, 40(23): 30–39 (in Russian).

13. Хусаинов Ф.И. 2017. Железнодорожная статистика: инструкции по применению. Available at: http://www.liberal.ru/upload/files/Husanov_30032017_JD_final.pdf.

Husainov F.I. 2017. Railway statistics: instructions for use. Available at: http://www.liberal.ru/upload/files/Husanov_30032017_JD_final.pdf (in Russian).

14. Adcock R.J. 1878. A problem in least squares. The Analyst, 5(2): 53–54.

15. Ahmad S., Mora S., Franks P.W., Orho-Melander M., Ridker P.M., Hu F.B., Chasman D.I. 2018. Adiposity and Genetic Factors in Relation to Triglycerides and Triglyceride-Rich Lipoproteins in the Women's Genome Health Study. Clinical Chemistry, 64(1): 231–241.

16. Clinical Chemistry. Available at: http://clinchem.aaccjnls.org/.

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

18. Dhanoa M.S., Sanderson R., López S., France J. 2016. Bivariate relationships incorporating method comparison: a review of linear regression methods. CAB Reviews, 11(028).

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

20. Jensen A.L., Kjelgaard-Hansen M. 2006. Method comparison in the clinical laboratory. Veterinary Clinical Pathology, 35 (3): 276–286.

21. Kummel C.H. 1879. Reduction of observed equations which contain more than one observed quantity. The Analyst, 6: 97–105.

22. Kvisvik B., Mørkrid L., Røsjø H., Cvancarova M., Rowe A.D., Eek C., Bendz B., Edvardsen T., Gravning J. 2017. High-Sensitivity Troponin T vs I in Acute Coronary Syndrome: Prediction of Significant Coronary Lesions and Long-term Prognosis. Clinical Chemistry, 63(2): 552–562.

23. Lewis L.K., Raudsepp S.D., Yandle T.G., Prickett T.C., Richards A.M. 2017. Development of a BNP1-32 Immunoassay That Does Not Cross-React with proBNP. Clinical Chemistry, 63(6): 1110–1117.

24. Taylor D.R., Ghataore L., Couchman L., Vincent R.P., Whitelaw B., Lewis D., Diaz-Cano S., Galata G., Schulte K-M, Aylwin S., Taylor N.F. 2017. A 13-Steroid Serum Panel Based on LC-MS/MS: Use in Detection of Adrenocortical Carcinoma. Clinical Chemistry, 63(12): 1836–1846.

25. Wu C., Yu J.Z. 2018. Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting. Atmospheric Measurement Techniques, 11: 1233–1250.