УДК 378
DOI 10.25688/2072-9014.2019.48.2.07

В.Б. Яковлев

Линейное и нелинейное оценивание параметров регрессии в Microsoft Excel

В статье рассматривается сравнительная оценка линейного и нелинейного оценивания параметров регрессии в Microsoft Excel. Для нелинейного оценивания предлагается применение нелинейного метода обобщенного понижающего градиента, реализованного в надстройке *Поиск решения*.

Ключевые слова: регрессионный анализ; оценивание параметров регрессии; метод обобщенного понижающего градиента.

При оценивании параметров линейной регрессии применяют различные методы (см., например, [1–5]). Из них наиболее часто используемым является метод наименьших квадратов, с помощью которого находят значения параметров уравнения регрессии, минимизируя суммы квадратов отклонений фактических данных от расчетных $\hat{y}_x$:

$$S = \sum \left(y - \hat{y}_x \right)^2 \to \min .$$

Для нахождения минимума функции S решают систему нормальных уравнений, полученную путем вычисления частных производных по каждому из параметров уравнения регрессии, приравненных к нулю.

Рассмотрим, как формируется система нормальных уравнений для уравнения линейной регрессии:

$$\hat{y}_x = a + bx,$$

где $\hat{y}_x$ — результативный признак; x — факторный признак; a — свободный член; b — коэффициент регрессии.

С учетом вида уравнения линейной регрессии величина S является функцией неизвестных параметров a и b:

$$S = \sum \left(y - a - bx \right)^2 .$$

Для данной функции найдем частные производные по каждому из параметров a и b и приравниваем их к нулю:

$$\frac{\partial S}{\partial a} = -2\sum(y - a - bx) = -2\sum y + 2na + 2b\sum x = 0;$$

$$\frac{\partial S}{\partial b} = -2b\sum(y - a - bx) = -2\sum xy + 2a\sum x + 2b\sum x^2 = 0.$$

После небольших преобразований получим следующую систему нормальных уравнений:

$$\begin{cases} \sum y = na + b\sum x, \\ \sum xy = a\sum x + b\sum x^2. \end{cases}$$

Соответствующую систему нормальных уравнений можно получить для любого уравнения линейной и нелинейной регрессии. При этом для нелинейной зависимости необходимо произвести линеаризацию переменных.

Нелинейные регрессии бывают трех типов.

1. Регрессии линейные по оцениваемым параметрам:

- равносторонняя гипербола — $y = a + \frac{b}{x}$;
- логарифмическая гипербола — $y = a + b\ln x$;
- полиномы различных степеней — $y = a + b_1 x + b_2 x^2$ (квадратичная степень), $y = a + b_1 x + b_2 x^2 + b_3 x^3$ (кубическая степень) и т. п.

2. Регрессии нелинейные по оцениваемым параметрам, но внутренне линейные:

- степенная — $y = ax^b$;
- экспоненциальная — $y = ae^{bx}$;
- показательная — $y = ab^x$.

3. Регрессии нелинейные по оцениваемым параметрам, но внутренне нелинейные:

$$y = a + b_1 x^{b_2};\ y = a\left(1 - \frac{1}{1 - x^b}\right).$$

Линеаризацию переменных можно произвести только для первых двух типов нелинейных уравнений регрессии. Первые приводят к линейному виду простой заменой переменных, вторые — с помощью логарифмирования (см. табл. 1). Внутренне нелинейные к линейному виду не приводятся.

Оценка параметров регрессий нелинейных по оцениваемым параметрам, но внутренне линейных, основывается, как правило, на минимизации суммы квадратов отклонений в логарифмах. В результате оценки параметров для линеаризуемых уравнений оказываются несколько смещенными, то есть заниженными.

Таблица 1

Линеаризация нелинейных регрессий

Регрессия	Исходное уравнение	Преобразованное уравнение
Линейные по оцениваемым параметрам		
Равносторонняя гипербола	$y = a + \frac{b}{x}$	$y = a + bx_1 \ \left(x_1 = \frac{1}{x}\right)$
Логарифмическая	$y = a + b\ln x$	$y = a + bx_1 \ \left(x_1 = \ln x\right)$
Квадратичная (полином 2-й степени)	$y = a + b_1 x + b_2 x^2$	$y = a + b_1 x_1 + b_2 x_2$ $\left(x_1 = x,\ x_2 = x^2\right)$
Кубическая (полином 3-й степени)	$y = a + b_1 x + b_2 x^2 + b_3 x^3$	$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$ $\left(x_1 = x,\ x_2 = x^2, x_3 = x^3\right)$
Нелинейные по оцениваемым параметрам, но внутренне линейные		
Степенная	$y = ax^b$	$\ln y = \ln a + b\ln x$
Экспоненциальная	$y = ae^{bx}$	$\ln y = \ln b + bx$
Показательная	$y = ab^x$	$\ln y = \ln a + x\ln b$

Это будет наглядно видно из дальнейшего анализа. Поэтому для таких регрессий предпочтительнее использовать нелинейные методы наименьших квадратов. В них, как и в классическом методе наименьших квадратов, находят значения параметров уравнения регрессии, при которых сумма квадратов отклонений S фактических данных y от расчетных $\hat{y}_x$ является минимальной.

Для решения этой задачи применяют два основных метода:

1) прямую минимизацию функции S методами нелинейной оптимизации, которые позволяют находить экстремумы выпуклых линий (сюда можно отнести различные методы наискорейшего спуска (градиентные методы), например метод обобщенного понижающего градиента, используемый в табличном процессоре Microsoft Excel, и др.);

2) решение системы нелинейных уравнений, полученной из необходимого условия экстремума функции — равенства нулю частных производных по каждому из параметров (эта система решается итерационными методами, например методом Гаусса-Ньютона, который используется в статистическом пакете Statistica, и др.).

Существуют также методы оценивания параметров нелинейной регрессии, которые сочетают в себе два вышеизложенных метода. Сюда можно отнести метод Левенберга-Марквардта, являющийся сочетанием направления Ньютона-Гаусса и метода наискорейшего спуска. Данный метод используется во многих статистических пакетах (Statistica, IBM SPSS Statistics и др.).

Рассмотрим методику линейного и нелинейного оценивания параметров регрессии для зависимости численности обучающихся от количества организаций высшего образования по Центральному федеральному округу (табл. 2).

Таблица 2

Количество государственных организаций высшего образования и численность обучающихся в них в Центральном федеральном округе в 2016 г.

№	Регион[1]	Численность обучающихся, тыс. чел. (*y*)	Государственные организации высшего образования, ед. (*x*)
1	Белгородская область	45,4	7
2	Брянская область	26,3	4
3	Владимирская область	27,3	2
4	Воронежская область	82,0	10
5	Ивановская область	27,0	6
6	Калужская область	15,5	1
7	Костромская область	11,3	2
8	Курская область	36,2	5
9	Липецкая область	19,8	4
10	Московская область	69,4	12
11	Орловская область	30,5	4
12	Рязанская область	28,8	4
13	Смоленская область	18,9	5
14	Тамбовская область	26,5	4
15	Тверская область	22,6	4
16	Тульская область	25,9	2
17	Ярославская область	28,7	7

1

Линейное оценивание параметров регрессии

В табличном процессоре Microsoft Excel линейное оценивание параметров регрессии может проводиться с помощью надстройки *Анализ данных* и с помощью добавления выбранных регрессий (линий тренда) на диаграмму зависимости результативного признака от факторного признака.

На основе данных диаграммы имеется возможность получать пять типов линейно оцененных регрессий или линий тренда, таких как линейная, полиномиальная различных степеней, логарифмическая, степенная и экспоненциальная.

При подборе линии тренда автоматически рассчитывается коэффициент детерминации R^2, который характеризует достоверность аппроксимации: чем ближе значение R^2 к единице, тем надежнее линия тренда аппроксимирует исследуемый процесс.

[1] Без г. Москвы — столицы Российской Федерации (город федерального значения).

Осуществим линейное оценивание параметров указанных выше регрессий с помощью точечной диаграммы, на которую добавим соответствующие линии трендов. Для этого подготовим данные в таблице MS Excel в соответствии с рисунком 1.

	A	B	C	D
1	№ п/п	Регион	Государственные организации высшего образования, ед. (x)	Численность обучающихся, тыс. чел. (y)
2	1	Белгородская область	7	45,4
3	2	Брянская область	4	26,3
18	17	Ярославская область	7	28,7

Рис. 1. Данные в MS Excel

Построим точечную диаграмму данной зависимости. Для этого выделим ячейки C1:D18 и выполним команду *Вставка, Рекомендуемые диаграммы*. В открывшемся диалоговом окне *Вставка диаграммы* установим параметры в соответствии с рисунком 2.

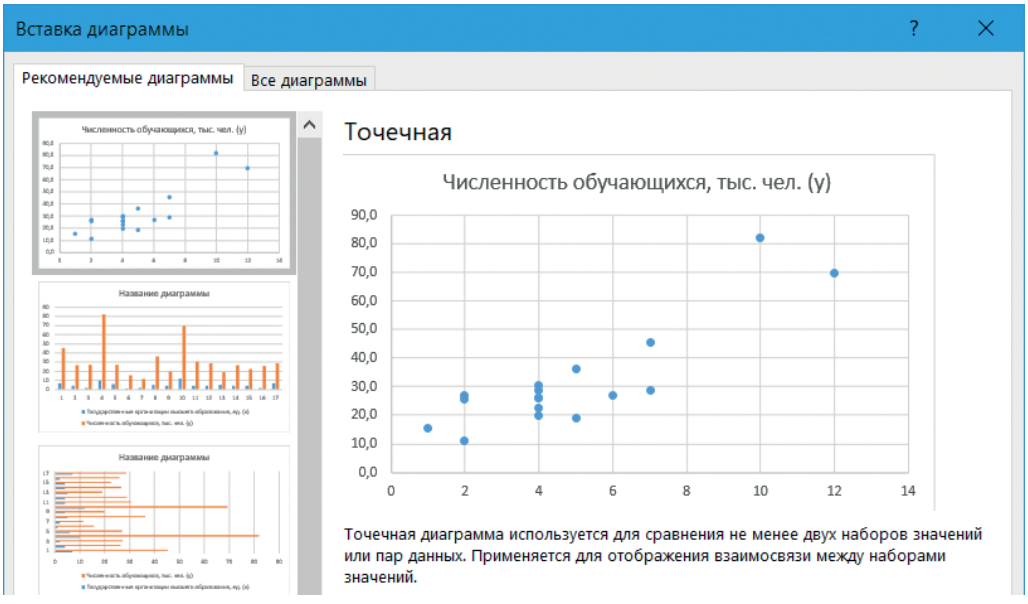

Рис. 2. Диалоговое окно *Вставка диаграммы*

Точечная диаграмма выводится в следующем отформатированном виде (рис. 3).

Вначале проведем оценивание параметров линейного уравнения регрессии. Для этого установим курсор на любую точку диаграммы и щелкнем правой кнопкой мыши. В появившемся контекстном меню нажмем на кнопку *Добавить линию тренда*.

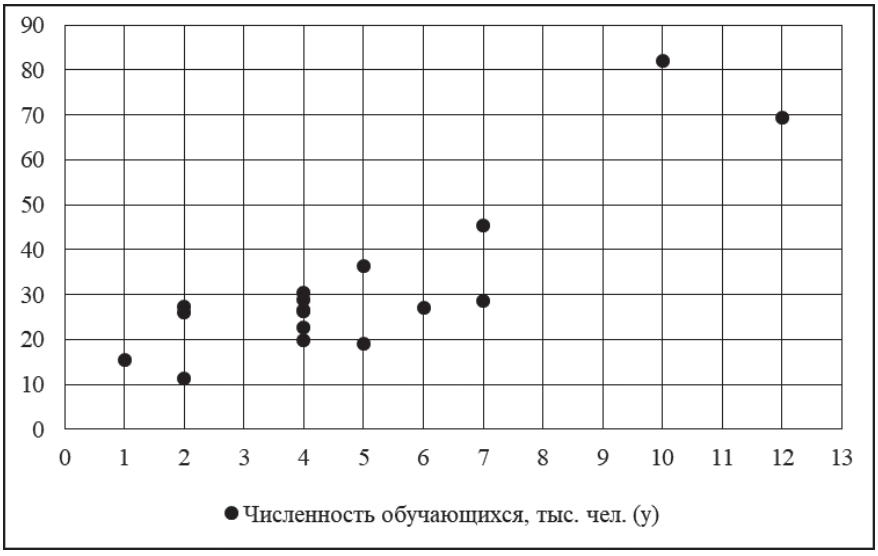

Рис. 3. Точечная диаграмма

В диалоговом окне *Формат линии тренда* установим параметры в соответствии с рисунком 4.

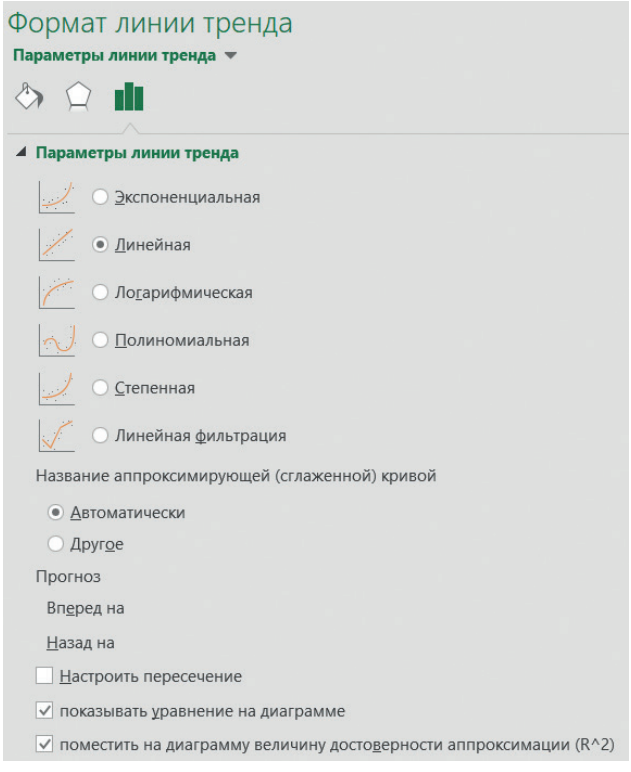

Рис. 4. Формат линии тренда

Диаграмма выводится в следующем отформатированном виде (рис. 5).

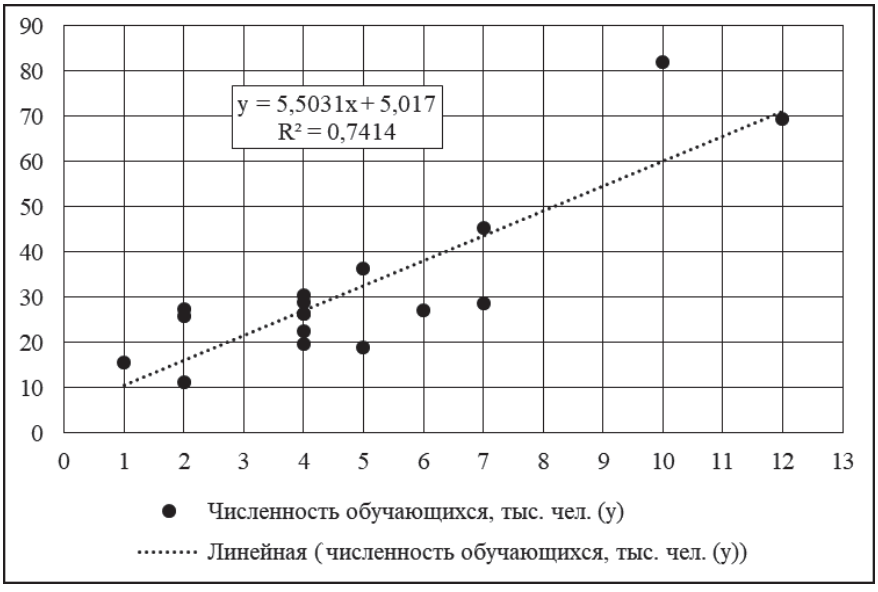

Рис. 5. Диаграмма

Проведем выравнивание по остальным уравнениям тренда аналогично выравниванию по линейному уравнению тренда (рис. 6).

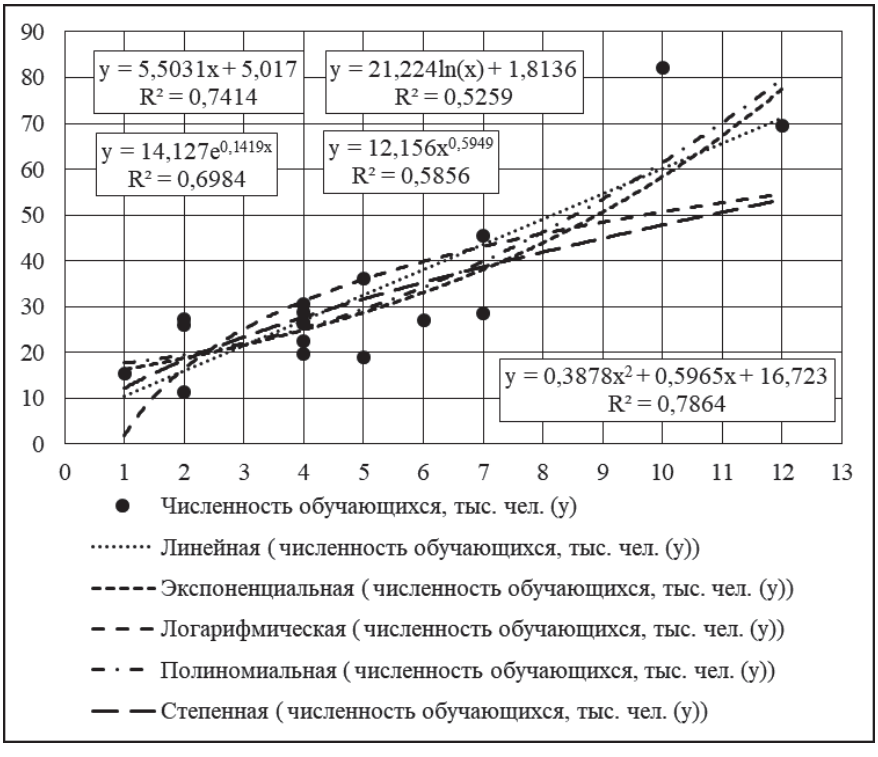

Рис. 6. Выравнивание по остальным уравнениям тренда

Более высокий индекс детерминации R^2 получается у полиномиальной (0,786) и линейной (0,741) регрессий. Поэтому можно сделать вывод, что данные регрессии в лучшей степени отражают зависимость численности обучающихся от количества организаций высшего образования.

Нелинейное оценивание параметров регрессии

Осуществим расчет параметров линейной, полиномиальной, логарифмической, степенной и экспоненциальной регрессий в Microsoft Excel с помощью нелинейного оценивания. Для этого используем надстройку *Поиск решения*, в которой реализован поиск решения нелинейных задач методом обобщенного понижающего градиента (ОПГ).

Подготовим данные в MS Excel в соответствии с рисунком 7.

	A	B	C	D	E	F
1	№ п/п	Регион	Государственные организации высшего образования, ед. (x)	Численность обучающихся, тыс. чел. (y)	Предсказанное y_x	Остатки $y-y_x$
2	1	Белгородская область	7	45,4		
3	2	Брянская область	4	26,3		
18	17	Ярославская область	7	28,7		
19						
20	**Расчет параметров уравнения регрессии**					
21	a					
22	b_1					
23	b_2					
24	$\Sigma(y-y_x)^2 \rightarrow min$					
25	**Оценка параметров уравнения регрессии**					
26	Общая дисперсия					
27	Остаточная дисперсия					
28	Индекс детерминации					

Рис. 7. Подготовка данных в MS Excel

Рассчитаем параметры уравнения линейной регрессии $\hat{y}_x = a + bx$. Для этого вначале введем в ячейки E2 и F2 соответственно формулы **=$C$21+$C$22*C2** и **=D2-E2** и скопируем их в ячейки E3:F18.

Затем найдем оптимальные значения параметров уравнения регрессии *a* и *b*, минимизируя сумму квадратов остатков (отклонений фактических уровней ряда от предсказанных значений ряда). Для этого введем в ячейку C24 функцию **=СУММПРОИЗВ(F2:F18;F2:F18)** и выполним команду *Данные, Поиск решения*. В диалоговом окне *Параметры поиска решения* установим параметры в соответствии с рисунком 8.

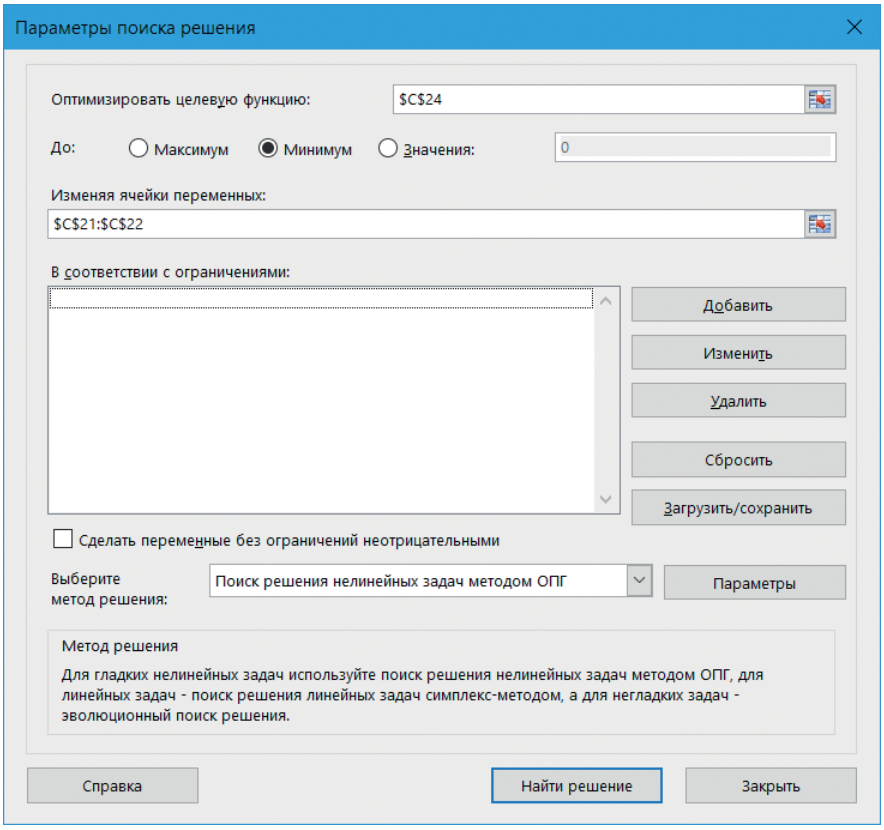

Рис. 8. Параметры поиска решения

В результате будут получены оптимальные значения параметров уравнения регрессии *a* и *b* (см. рис. 9).

	A	B	C	D	E	F
1	№ п/п	Регион	Государственные организации высшего образования, ед. (x)	Численность обучающихся, тыс. чел. (y)	Предсказанное y_x	Остатки $y-y_x$
2	1	Белгородская область	7	45,4	43,5	1,9
3	2	Брянская область	4	26,3	27,0	-0,7
18	17	Ярославская область	7	28,7	43,5	-14,9
19						
20	**Расчет параметров уравнения регрессии**					
21	a		5,017			
22	b_1		5,503			
23	b_2					
24	$\Sigma(y-y_x)^2 \rightarrow min$		1391,8			

Рис. 9. Оптимальные значения параметров уравнения регрессии *a* и *b*

Уравнение линейной регрессии имеет вид:

$$\hat{y}_x = 5,017 + 5,503x.$$

Насколько уравнение регрессии соответствует изучаемой совокупности, оценим с помощью индекса детерминации R^2, который рассчитывается по формуле:

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2},$$

где $\sigma_y^2 = \frac{\sum(y - \overline{y})^2}{n}$ — общая дисперсия результативного признака;

$\sigma_\varepsilon^2 = \frac{\sum(y - \hat{y}_x)^2}{n}$ — остаточная дисперсия результативного признака;

n – численность совокупности.

Для этого введем:

- в ячейку C26 функцию **=ДИСП.Г(D2:D18)** для расчета общей дисперсии;
- в ячейку C27 формулу **=C24/A18** для расчета остаточной дисперсии;
- в ячейку C28 формулу **=1-C27/C26** для расчета индекса детерминации.

Результаты выводятся в следующем виде (рис. 10).

	A	B	C
25	**Оценка параметров уравнения регрессии**		
26	Общая дисперсия		316,6
27	Остаточная дисперсия		81,9
28	Индекс детерминации		0,741

Рис. 10. Параметры уравнения регрессии

Как видим, получены те же результаты, что и при линейной оценке параметров линейного уравнения регрессии (см. рис. 5).

Аналогично проведем нелинейное оценивание параметров полиномиальной, логарифмической, степенной и экспоненциальной регрессий. Для этого для каждого уравнения регрессии создадим копии листа с расчетными данными по линейной регрессии и на скопированных листах проведем соответствующие расчеты: на каждом листе введем в ячейку E2 соответствующие формулы и скопируем их в ячейки E3:E15:

- **=$C$21+$C$22*C2+$C$23*C2^2** (полиномиальная регрессия);
- **=$C$21+$C$22*LN(C2)** (логарифмическая регрессия);
- **=$C$21*C2^$C$22** (степенная регрессия);
- **=$C$21*EXP($C$22*C2)** (экспоненциальная регрессия).

Получим следующие результаты нелинейного оценивания (см. табл. 3). Для сравнения в таблице приведены также результаты линейного оценивания параметров регрессий.

Таблица 3

Линейное и нелинейное оценивание параметров регрессии

Регрессия	Линейное оценивание	Нелинейное оценивание
Линейная	$\hat{y}_x = 5,017 + 5,503x,$ $R^2 = 0,741$	$\hat{y}_x = 5,017 + 5,503x,$ $R^2 = 0,741$
Нелинейные регрессии, линейные по оцениваемым параметрам		
Полиномиальная	$\hat{y}_x = 16,723 + 0,597x + 0,388x^2,$ $R^2 = 0,786$	$\hat{y}_x = 16,723 + 0,597x + 0,388x^2,$ $R^2 = 0,786$
Логарифмическая	$\hat{y}_x = 1,814 + 21,224 \ln x,$ $R^2 = 0,526$	$\hat{y}_x = 1,814 + 21,224 \ln x,$ $R^2 = 0,526$
Нелинейные регрессии, нелинейные по оцениваемым параметрам		
Степенная	$\hat{y}_x = 12,156x^{0,595},$ $R^2 = 0,586$	$\hat{y}_x = 7,17x^{0,933},$ $R^2 = 0,605$
Экспоненциальная	$\hat{y}_x = 14,127e^{0,142x},$ $R^2 = 0,698$	$\hat{y}_x = 14,736e^{0,141x},$ $R^2 = 0,777$

Нелинейное оценивание степенной и экспоненциальной регрессий показывает лучшие результаты по сравнению с линейным оцениванием (см. рис. 6), поскольку получен более высокий индекс детерминации, характеризующий достоверность аппроксимации. Это следует из того, что оценки параметров для линеаризуемых уравнений, как правило, оказываются несколько смещенными, то есть заниженными.

Нанесем на диаграммы результаты линейного и нелинейного оценивания степенной (рис. 11) и экспоненциальной регрессий (рис. 12). На рисунках наглядно видны различия в их оценке, особенно степенной регрессии.

Проверка итогов нелинейного оценивания регрессии в статистическом пакете Statistica 10 показала аналогичные результаты (см. рис. 13, 14).

Таким образом, нелинейное оценивание регрессий в Microsoft Excel, нелинейных по оцениваемым параметрам, приводит к лучшим результатам по сравнению с линейным оцениванием. Поэтому его рекомендуется применять на практике для такого рода регрессий, причем как внутренне линейных, так и внутренне нелинейных.

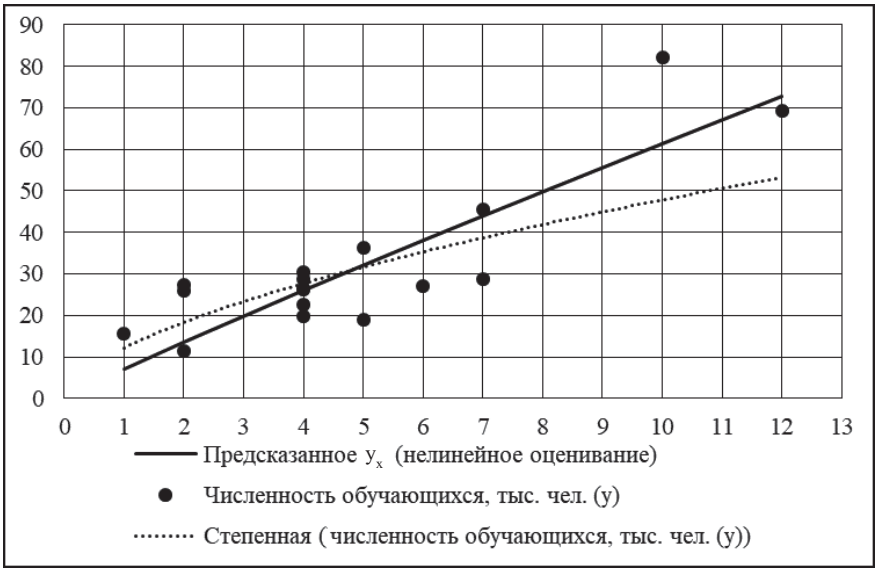

Рис. 11. Результаты линейного и нелинейного оценивания степенной регрессии

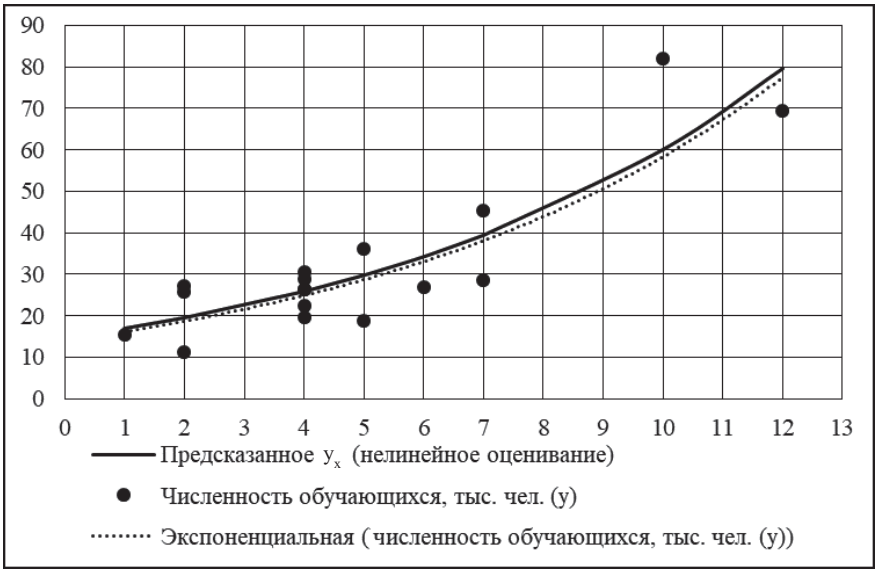

Рис. 12. Результаты линейного и нелинейного оценивания экспоненциальной регрессии

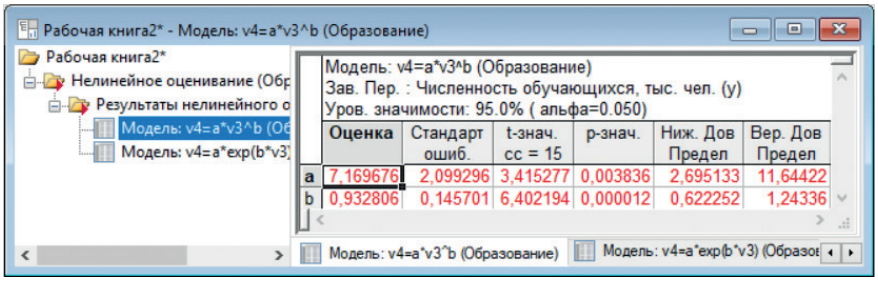

Рис. 13. Проверка итогов нелинейного оценивания регрессии в статистическом пакете Statistica 10

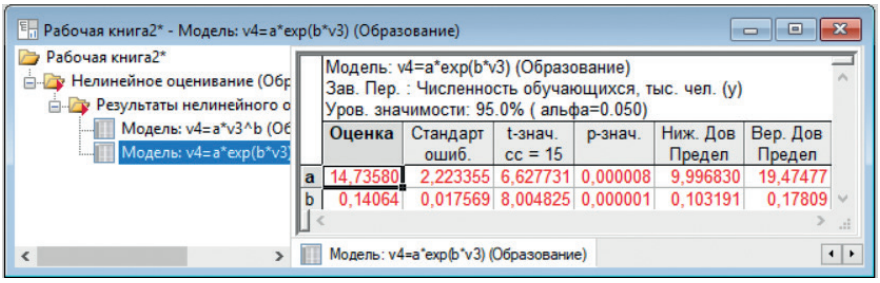

Рис. 14. Проверка итогов нелинейного оценивания регрессии в статистическом пакете Statistica 10

Литература

1. *Воскобойников Ю.Е.* Построение регрессионных эконометрических моделей (с примерами в Excel): учебное пособие. Новосибирск: НГАСУ (Сибстрин), 2014. 224 с.

2. *Конрад Карлберг.* Регрессионный анализ в Microsoft Excel. М.: Диалектика, 2017. 400 с.

3. *Яковлев В.Б.* Статистика. Расчеты в Microsoft Excel: учебное пособие. М.: Юрайт, 2017. 353 с.

4. *Яковлев В.Б.* Эконометрика в Excel и Statistica: учебное пособие. Ч. 1. Регрессионный анализ. М.: Эдитус, 2018. 168 с.

5. *Яковлев В.Б., Яковлева О.А.* Практикум по общей теории статистики: учебное пособие. М.: ИНФРА-М, 2016. 382 с.

Literatura

1. *Voskobojnikov Yu.E.* Postroenie regressionny'x e'konometricheskix modelej (s primerami v Excel): uchebnoe posobie. Novosibirsk: NGASU (Sibstrin), 2014. 224 s.

2. *Konrad Karlberg.* Regressionny'j analiz v Microsoft Excel. M.: Dialektika, 2017. 400 s.

3. *Yakovlev V.B.* Statistika. Raschety' v Microsoft Excel: uchebnoe posobie. M.: Yurajt, 2017. 353 s.

4. *Yakovlev V.B.* E'konometrika v Excel i Statistica: uchebnoe posobie. Ch. 1. Regressionny'j analiz. M.: E'ditus, 2018. 168 s.

5. *Yakovlev V.B., Yakovleva O.A.* Praktikum po obshhej teorii statistiki: uchebnoe posobie. M.: INFRA-M, 2016. 382 s.

V.B. Yakovlev

Linear and Nonlinear Estimation of Regression Parameters in Microsoft Excel

The article deals with the comparative evaluation of linear and nonlinear estimation of regression parameters in Microsoft Excel. For nonlinear estimation, it is proposed to apply the nonlinear method of the generalized gradient, reduction implemented in Solver add-in.

Keywords: regression analysis; estimation of regression parameters; generalized gradient reduction method.