УДК 519.862.6

К ВОПРОСУ ОБ ОЦЕНКЕ КАЧЕСТВА ЭКОНОМЕТРИЧЕСКИХ МОДЕЛЕЙ

Орлова И.В.

ФГОБУ ВО «Финансовый университет при Правительстве РФ», Москва, e-mail: IVOrlova@fa.ru

В работе рассматриваются вопросы оценки качества моделей, выбора оптимальных моделей. Анализируются подходы к проверке адекватности регрессионных моделей, предназначенных для прогнозирования. Рассматриваются возможности Excel для проверки спецификации модели с помощью метода Салкевера при моделировании зависимости количества безработных в России от заявленной потребности в работниках по данным за период с 2001 по 2020 г. Для оценки адекватности модели и выбора лучшей модели используется перекрестная проверка с последовательным исключением одного наблюдения. При этом исследуются два способа реализации метода перекрестной проверки. В первом случае критерий перекрестной проверки CV может быть механически вычислен путем выполнения n регрессий, в которых каждый раз пропускается одно наблюдение, а все остальные используются для прогнозирования его значения. Другой способ, менее трудоемкий, связан с использованием так называемой матрицы шляп для вычисления критерия перекрестной проверки CV. Этот метод включен в свободно распространяемую программу Gretl. Применение метода перекрёстной проверки продемонстрировано на примере моделирования зависимости рождаемости (число родившихся на 1000 чел.) от индекса потребительских цен на товары и услуги 2020 г. на данных 16 регионов РФ за 2020 г. В заключении приведены выводы относительно применения инструментария для оценки адекватности моделей.

Ключевые слова: регрессия, адекватность модели, перекрестная проверка, программа Gretl

TO THE QUESTION OF ASSESSING THE QUALITY OF ECONOMETRIC MODELS

Orlova I.V.

Financial University under the Government of the Russian Federation, Moscow, e-mail: ivorlova@fa.ru

The paper deals with the issues of assessing the quality of models, the choice of optimal models. Approaches to checking the adequacy of regression models intended for forecasting are analyzed. The possibilities of Excel for checking the specification of the model using the Salkever method when modeling the dependence of the number of unemployed in Russia on the declared need for workers according to data for the period from 2001 to 2020 are considered. To assess the adequacy of the model and select the best model, cross-validation is used with the successive exclusion of one observation. In this case, two ways of implementing the cross-validation method are investigated. In the first case, the CV cross-validation criterion can be mechanically computed by running n regressions in which one observation is skipped each time and all the others are used to predict its value. Another way, less laborious, involves using the so-called hat matrix to calculate the CV cross-validation criterion. This method is included in the free Gretl program. The application of the cross-validation method is demonstrated by modeling the dependence of the birth rate (number of births per 1000 people) on the consumer price index for goods and services in 2020 using data from 16 regions of the Russian Federation for 2020. In conclusion, conclusions are given regarding the use of tools for assessing the adequacy of models.

Keywords: regression, model adequacy, cross-validation, Gretl program

При эконометрическом моделировании весьма важными являются вопросы оценки качества построенных моделей, выбора оптимальных моделей. Существуют различные подходы к решению этих вопросов. Будем рассматривать проблемы, связанные только с оценкой качества линейных регрессионных моделей. Пусть спецификация регрессионной модели имеет вид

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k + \varepsilon, \quad (1)$$

где Y – *эндогенная* (зависимая) переменная, k – количество регрессоров, ε – случайная составляющая эндогенной переменной (*случайное возмущение*), которая не может быть объяснена значениями объясняющих переменных $X_1, X_2, ..., X_k$. Количество параметров модели равно m, $m = k+1$.

Обычно считают, что «модель считается хорошей со статистической точки зрения, если она адекватна и достаточно точна» [1, с. 310]. Если вопросы оценки точности модели, как правило, не вызывают разночтений, то по оценке адекватности не существует единого мнения. Существует распространённое мнение, что проверка адекватности модели означает проверку гипотезы о равенстве нулю всех коэффициентов регрессии (H_0: $b_1 = b_2 = \ldots b_k = 0$), т.е. проверяется значимость модели регрессии в целом [2, 3]. Если основная гипотеза H_0 принимается, то модель считается неадекватной. Если же основная гипотеза отклоняется, то модель можно считать адекватной только после проверки выполнения предпосылок МНК относительно остатков:

равенство нулю математического ожидания, гомоскедастичность, случайность и независимость, соответствие нормальному закону распределения. Если эти предпосылки не выполняются, то модель признается неадекватной.

Другое представление об адекватности модели заключается в проверке качества прогнозов, получаемых на базе обучающей выборки путем сравнения этих прогнозов с реальными значениями из контролирующей выборки [4–6]. При этом следует иметь в виду, что такая проверка осуществляется только при выполнении предпосылок МНК.

Подход, когда модель обучается на одном образце данных («обучающем наборе») и оценивается вне выборки на так называемом «тестовом наборе», известен как перекрестная проверка (cross-validation, сокращенно CV).

Целью работы является анализ разных подходов к оценке адекватности и выбору линейных регрессионных моделей, предназначенных для прогнозирования, и исследование инструментария для проведения перекрестной проверки моделей, построенных на пространственных наблюдениях.

Материалы и методы исследования

Использование интервальных прогнозов для проверки спецификации модели. Если значения эндогенной переменной из контролирующей выборки попадают в прогнозные интервалы, то спецификация модели подтверждается.

Исследуем зависимость количества безработных в среднем в млн чел в России от заявленной потребности в работниках (в тыс. чел.) [7].

По данным (табл. 1) за период с 2001 по 2019 г. (обучающая выборка) построена регрессионная модель зависимости количества безработных в среднем (Y) от заявленной потребности в работниках – X: $\hat{Y} = 8,481 - 0,0028 \cdot X$. В качестве контролирующей выборки используются данные за 2020 г.

Для оценки прогноза $\hat{Y}_p$ на 2020 г. по модели $\hat{Y} = 8,481 - 0,0028 \cdot X$ используем значения X за 2020 г. и получим точечный прогноз $\hat{Y}_p$:

$$\hat{Y}_p = 8,481 - 0,0028 \cdot 1578,9 = 4,044 \text{ млн чел.}$$

Ошибка прогноза s_p, необходимая для вычисления доверительного интервала $\hat{Y}_p$, вычисляется по формуле:

$$s_p = s_e \sqrt{1 + X_p^T (X^T X)^{-1} X_p},$$

где X_p^T – строка матрицы X, относящаяся к 2020 г., $X_p^T = (1;1578,9)$, s_e – стандартная ошибка модели. Вычисление ошибки прогноза s_p в Excel с использованием матричных функций МУМНОЖ, ТРАНСП, МОБР является несложной, но затратной по времени процедурой. Последовательность вычислений приведена на рис. 1. Как видим, s_p = 0,461. Значение t-статистики $t_{kp}(0,05;17)$ равно 2,11,

$$НГ_p = \hat{Y}_p - t_{kp} \cdot s_p = 4,044 - 2,11 \cdot 0,461 = 3,07$$
($НГ_p$ – нижняя граница)

$$ВГ_p = \hat{Y}_p + t_{kp} \cdot s_p = 4,044 + 2,11 \cdot 0,461 = 5,02$$
($ВГ_p$ – верхняя граница)

Так как значение эндогенной переменной из контролирующей выборки Y(2020), равное 4,3, попадает в 95%-ый доверительный интервал (3,07 5,02), то модель признаётся адекватной.

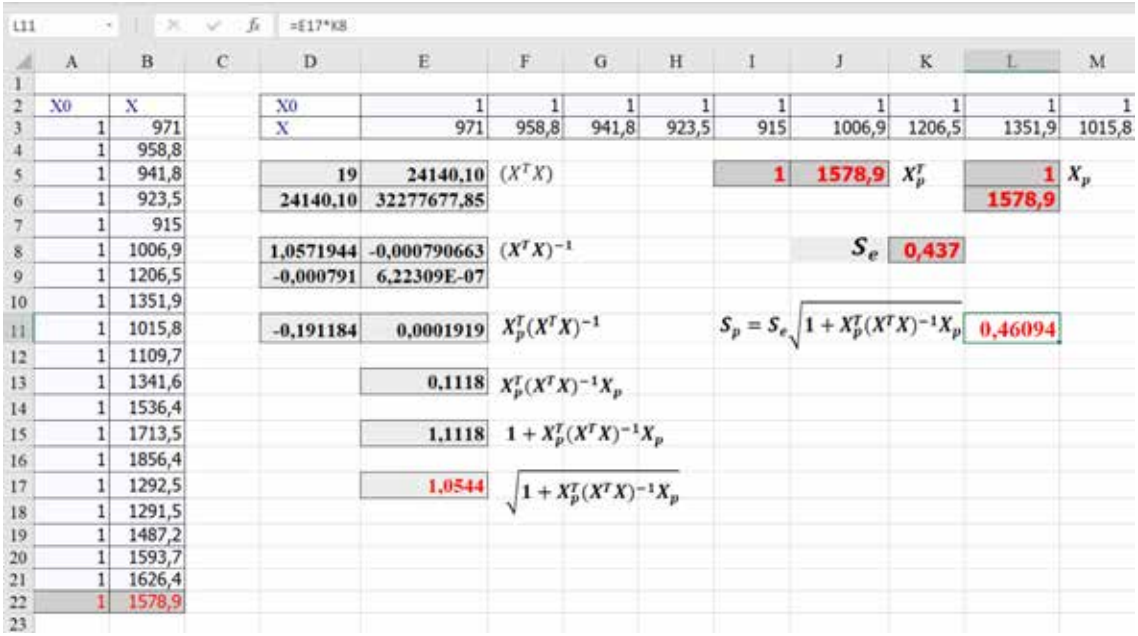
Таблица 1

Исходные данные

T	2001	2002	2003	2004	2005	2006	2007
X	971	958,8	941,8	923,5	915	1006,9	1206,5
Y	6,4	5,7	6,1	6	5,6	5,3	4,6

T	2008	2009	2010	2011	2012	2013	2014
X	1351,9	1015,8	1109,7	1341,6	1536,4	1713,5	1856,4
Y	4,8	6,3	5,6	5	4,2	4,1	3,9

T	2015	2016	2017	2018	2019	2020
X	1292,5	1291,5	1487,2	1593,7	1626,4	**1578,9**
Y	4,2	4,3	4	3,7	3,5	4,3



	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2	X0	X		X0	1	1	1	1	1	1	1	1	1
3	1	971		X	971	958,8	941,8	923,5	915	1006,9	1206,5	1351,9	1015,8
4	1	958,8											
5	1	941,8		19	24140,10	(X^TX)			1	1578,9	X_p^T	1	X_p
6	1	923,5		24140,10	32277677,85							1578,9	
7	1	915											
8	1	1006,9		1,0571944	-0,000790663	$(X^TX)^{-1}$					S_e	0,437	
9	1	1206,5		-0,000791	6,22309E-07								
10	1	1351,9											
11	1	1015,8		-0,191184	0,0001919	$X_p^T(X^TX)^{-1}$				$S_p = S_e\sqrt{1 + X_p^T(X^TX)^{-1}X_p}$		0,46094	
12	1	1109,7											
13	1	1341,6			0,1118	$X_p^T(X^TX)^{-1}X_p$							
14	1	1536,4											
15	1	1713,5			1,1118	$1 + X_p^T(X^TX)^{-1}X_p$							
16	1	1856,4											
17	1	1292,5			1,0544	$\sqrt{1 + X_p^T(X^TX)^{-1}X_p}$							
18	1	1291,5											
19	1	1487,2											
20	1	1593,7											
21	1	1626,4											
22	1	1578,9											
23													

Рис. 1. Вычисление ошибки прогноза s_p в Excel

	A	B	C	D	E	F	G	H	I	J
1	T	Y	X	Z		ВЫВОД ИТОГОВ				
2	2001	6,4	971	0						
3	2002	5,7	958,8	0		*Регрессионная статистика*				
4	2003	6,1	941,8	0		Множественный R	0,89			
5	2004	6	923,5	0		R-квадрат	0,80			
6	2005	5,6	915	0		Нормированный R-квад	0,78			
7	2006	5,3	1006,9	0		Стандартная ошибка	0,44			
8	2007	4,6	1206,5	0		Наблюдения	20			
9	2008	4,8	1351,9	0						
10	2009	6,3	1015,8	0		Дисперсионный анализ				
11	2010	5,6	1109,7	0			*df*	*SS*	*MS*	*F*
12	2011	5	1341,6	0		Регрессия	2	13,04325515	6,521627576	34,12631
13	2012	4,2	1536,4	0		Остаток	17	3,248744849	0,191102638	
14	2013	4,1	1713,5	0		Итого	19	16,292		
15	2014	3,9	1856,4	0						
16	2015	4,2	1292,5	0			*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*
17	2016	4,3	1291,5	0		Y-пересечение	8,481	0,449	18,868	0,000
18	2017	4	1487,2	0		X	-0,003	0,000	-8,149	0,000
19	2018	3,7	1593,7	0		Z	0,256	0,461	0,555	0,586
20	2019	3,5	1626,4	0						
21	2020	4,3	1578,9	1						

Рис. 2. Оценка параметров модели с фиктивной переменной: $\widehat{Yz} = 8,481 - 0,003 \cdot X + 0,256 \cdot Z$
(0,449) (0,000) ***(0,461)***

Для вычисления стандартных ошибок прогнозов воспользуемся методом Салкевера [8, 9]. В этом методе для оценки стандартной ошибки прогноза на момент $t = n + 1$ в матрицу регрессоров добавляется строка X_{n+1} и столбец фиктивных переменных Z, содержащий нули для всех наблюдений, кроме $(n + 1)$-го, в котором фиктивная переменная равна 1. На рис. 2 приведена таблица исходных данных и результаты оценки параметров модели с фиктивной переменной. Стандартная ошибка оценки параметра при фиктивной переменной равна стандартной ошибке прогноза. В нашем случае она равна $s_p = 0,461$.

Использование перекрестной проверки для оценки адекватности модели и выбора лучшей модели

Рассмотрим применение одного из самых простых методов перекрестной про-

верки LOOCV (Leave One Out Cross Validation) – перекрестная проверка с исключением одного наблюдения. В LOOCV каждое наблюдение рассматривается как контролирующий набор, а остальные (n–1) наблюдений – как обучающий набор. Подгонка модели и прогнозирование повторяется n раз. Критерий перекрестной проверки CV может быть вычислен путем выполнения n регрессий, в которых каждый раз пропускается одно наблюдение, а все остальные используются для прогнозирования его значения. Сумма n квадратов ошибок прогноза – это и есть CV – та статистика, которая используется для оценки качества модели. Однако в вычислении n регрессий нет необходимости. Эту статистику можно найти иначе, воспользовавшись так называемой матрицей шляп H от английского слова hat. Матрица H равна

$$H = X(X^T X)^{-1} X^T. \qquad (2)$$

Покажем, как ошибку прогноза i-го наблюдения зависимой переменной по уравнению регрессии без i-го наблюдения можно вычислить, зная лишь ошибку прогноза этого наблюдения по уравнению регрессии с полным набором наблюдений и диагональные элементы матрицы *H*.

Пусть *X*, *Y* – матрица регрессоров и вектор значений зависимой переменной, $X_{[i]}$, $Y_{[i]}$ получены из *X*, *Y* после удаления из них i-го наблюдения, X_i^T – i-я строка *X* и пусть $\hat{b}_{[i]} = \left(X_{[i]}^T X_{[i]}\right)^{-1} X_{[i]}^T Y_{[i]}$ – оценка вектора коэффициентов регрессии *b* без i-го наблюдения.

Тогда ошибка прогноза i-го наблюдения, вычисленная по регрессии без i-го наблюдения, равна $e_{[i]} = y_i - X_{[i]}^T \hat{b}_{[i]}$. Очевидно, произведение $X_{[i]}^T X_{[i]}$ можно представить в виде $X_{[i]}^T X_{[i]} = (X^T X - X_i^T X_i)$. По формуле Шермана – Моррисона – Вудбери [10]

$$\left(X^T X\right)^{-1} = \left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} X_i X_i^T \left(X^T X\right)^{-1}}{1 - X_i^T \left(X^T X\right)^{-1} X_i}.$$

Но произведение $X_i^T \left(X^T X\right)^{-1} X_i$ равно диагональному элементу h_i матрицы *H*, $X_i^T \left(X^T X\right)^{-1} X_i = h_i$. Тогда

$$\left(X_{[i]}^T X_{[i]}\right)^{-1} = \left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} X_i X_i^T \left(X^T X\right)^{-1}}{1 - h_i}.$$

Учитывая также, что $X_{[i]}^T Y_{[i]} = X^T Y - X_i Y_i$, получаем

$$\hat{b}_{[i]} = [\left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} X_i X_i^T \left(X^T X\right)^{-1}}{1 - h_i}](X^T Y - X_i Y_i) =$$

$$= \left(X^T X\right)^{-1} X^T Y - \frac{\left(X^T X\right)^{-1} X_i}{1 - h_i}((1 - h_i) Y_i - X_i^T \left(X^T X\right)^{-1} X^T Y + X_i^T \left(X^T X\right)^{-1} X_i Y_i) =$$

$$= \hat{b} - \frac{\left(X^T X\right)^{-1} X_i}{1 - h_i}(Y_i - h_i Y_i - X_i^T \hat{b} + h_i Y_i) = \hat{b} - \frac{\left(X^T X\right)^{-1} X_i}{1 - h_i}(Y_i - X_i^T \hat{b}) = \hat{b} - \frac{\left(X^T X\right)^{-1} X_i}{1 - h_i} e_i,$$

где $e_i = y_i - \widehat{y_i} = y_i - X_i^T \hat{b}$ – ошибка прогноза i-го наблюдения по уравнению регрессии с полным набором переменных. Тогда

$$e_{[i]} = y_i - X_i^T \hat{b}_{[i]} = y_i - X_i^T (\hat{b} - \frac{\left(X^T X\right)^{-1} X_i}{1 - h_i} e_i) =$$

$$= y_i - X_i^T \hat{b} + \frac{X_i^T \left(X^T X\right)^{-1} X_i}{1 - h_i} e_i = e_i + \frac{h_i e_i}{1 - h_i} = \frac{e_i}{1 - h_i}.$$

Таким образом, получили, что ошибка прогноза i-го наблюдения по регрессии без i-го наблюдения $e_{[i]}$ равна $e_{[i]} = e_i / (1 - h_i)$. Полученная формула существенно упрощает процедуру вычисления критерия перекрестной проверки CV,

$$CV = \sum_{i=1}^{n} \left[\frac{e_i}{(1 - h_i)}\right]^2, \qquad (3)$$

где h_i – диагональный элемент матрицы *H*.

Для пояснения смысла матрицы H запишем предсказываемые моделью значения эндогенной переменной Y в виде

$$\hat{Y} = X\hat{b} = \mathrm{X}\left(X^T X\right)^{-1} X^T Y = H \cdot Y$$

или, в координатной форме,

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n,\ i=1,\ldots n.$$

Диагональные элементы матрицы H изменяются от нуля до единицы и в сумме равны числу параметров модели m. Показатель h_{ii} (диагональный элемент h_{ii} матрицы H отражает расстояние между точкой с координатами X_i и центром данных. Если значение h_i близко к нулю, то это означает, что i-я точка X_i располагается недалеко от центра, если h_i близка к единице, то i-я точка является удаленной. Считается, что наблюдение оказывает существенное влияние на параметры модели, если $h_i > \frac{2\mathrm{m}}{n}$. Чем дальше от центра системы находится наблюдение, тем больше его влияние на оценку коэффициентов регрессии. Такие наблюдения называют точками разбалансировки (леверидж). Показатель h_i является удобным индикатором того, является ли i-е наблюдение точкой разбалансировки. Именно диагональные значения матрицы H используются при вычислении статистики CV. При выборе лучшей модели из нескольких выбирается та, у которой меньше значение статистики CV.

Результаты исследования и их обсуждение

Применим метод LOOCV для выбора лучшей модели при моделировании зависимости рождаемости (число родившихся на 1000 чел.) [11] от индекса цен (Индексы потребительских цен на товары и услуги 2020 г.) по данным 16 регионов РФ [12].

Построенная по всем наблюдениям (табл. 2) модель имеет вид:

$$\hat{Y} = 218{,}163 - 1{,}979 \cdot X.$$

Из приведенного протокола (рис. 3) можно сделать вывод, что параметры модели значимы, коэффициент детерминации достаточно высокий 0,87. Для выбранных регионов увеличение индекса цен на один процент приводит в среднем к уменьшению числа родившихся примерно на два человека.

Для оценки качества этой модели построено 16 уравнений регрессии, в каждом из которых последовательно удалялось по одному наблюдению. Оценки параметров этих моделей, прогноз на 16-е наблюдение, ошибка прогноза и квадрат ошибки приведены в табл. 2. Сумма квадратов ошибок – это и есть CV – равна **14,824.**

Другой подход к вычислению статистики CV с помощью H матрицы приведен в табл. 3 и 4. В табл. 3 приведен фрагмент матрицы H, вычисленной с помощью матричных преобразований по формуле (2).

Таблица 2

Исходные данные и результаты перекрестной проверки по 16 моделям

№	Регион	X	**Y**	b1	b0	Y^	e	e^2
1	Тульская область	106,1	7,4	-1,946	214,656	8,229	-0,8	**0,687**
2	Пензенская область	106,12	7,4	-1,946	214,745	8,187	-0,8	**0,619**
3	Ивановская область	105,79	7,6	-1,949	215,064	8,857	-1,3	**1,581**
4	Саратовская область	106,69	7,7	-2,041	224,634	6,847	0,9	**0,727**
5	Рязанская область	106,01	7,9	-1,962	216,414	8,371	-0,5	**0,221**
6	Новгородская область	105,57	8,2	-1,966	216,866	9,271	-1,1	**1,146**
7	Воронежская область	106,93	8,2	-2,160	237,051	6,091	2,1	**4,447**
8	Курская область	105,77	8,3	-1,967	216,884	8,844	-0,5	**0,296**
9	Липецкая область	106,14	8,3	-1,990	219,285	8,048	0,3	**0,063**
10	Республика Карелия	106,06	8,5	-1,991	219,347	8,204	0,3	**0,087**
11	Республика Татарстан	104,78	10,6	-1,985	218,751	10,779	-0,2	**0,032**
12	Ханты-Мансийский автономный округ	103,89	12,3	-2,001	220,460	12,575	-0,3	**0,076**
13	Тюменская область	104,22	12,3	-1,949	214,981	11,809	0,5	**0,241**
14	Ямало-Ненецкий автономный округ	103,36	12,9	-2,079	228,768	13,832	-0,9	**0,869**
15	Республика Алтай	104,16	13,3	-1,882	207,809	11,781	1,5	**2,306**
16	Республика Саха (Якутия)	103,95	13,4	-1,889	208,604	12,207	1,2	**1,423**
								14,824

	Коэффициент	*Ст. ошибка*	*t-статистика*	*p-значение*	
const	218,163	21,4957	10,15	**<0,0001**	***
X	−1,97937	0,204037	−9,701	**<0,0001**	***
Среднее завис. перемен	9,643750	Ст. откл. завис. перем		2,355976	
Сумма кв. остатков	10,78190	Ст. ошибка модели		**0,877574**	
R-квадрат	**0,870502**	Исправ. R-квадрат		0,861252	
F(1, 14)	**94,10996**	P-значение (F)		1,36e-07	
Лог. правдоподобие	−19,54526	Крит. Акаике		43,09052	
Крит. Шварца	44,63570	Крит. Хеннана-Куинна		43,16965	

Рис. 3. Оценка параметров модели регрессии зависимости рождаемости от индекса цен на данных 16 регионов РФ

В табл. 4 приведены остатки, полученные при построении модели регрессии по всем наблюдениям, диагональные элементы матрицы *H* и *CV* критерий.

Метод перекрестной проверки с исключением одного наблюдения реализован в Gretl. При анализе построенной модели (рис. 3) в меню следует выбрать значимость наблюдений (рис. 4) и в качестве дополнительной информации для команды leverage будет получен критерий CV (табл. 5).

Таблица 3

Фрагмент матрицы H

	1	2	3	4	…	13	14	15	16
1	**0,093**	0,094	0,081	0,117	…	0,017	-0,018	0,014	0,006
2	0,094	**0,095**	0,081	0,119	…	0,015	-0,021	0,013	0,004
3	0,081	0,081	**0,073**	0,095	…	0,035	0,015	0,034	0,029
…	…	…	**…**	…	…	…	…	…	…
13	0,017	0,015	0,035	-0,019	…	**0,131**	0,183	0,135	0,148
14	-0,018	-0,021	0,015	-0,082	…	0,183	**0,276**	0,190	0,212
15	0,014	0,013	0,034	-0,024	…	0,135	0,190	**0,139**	0,152
16	0,006	0,004	0,029	-0,039	…	0,148	0,212	0,152	**0,168**

Таблица 4

Вычисление критерия перекрестной проверки на основе матрицы H

№	Остатки	h_i	*CV*
1	-0,752	0,093	0,687
2	-0,712	0,095	0,619
3	-1,165	0,073	1,581
4	0,716	0,160	0,727
5	-0,430	0,086	0,221
6	-1,001	0,065	1,146
7	1,691	0,198	4,447
8	-0,505	0,072	0,296
9	0,227	0,097	0,063
10	0,269	0,090	0,087
11	-0,165	0,080	0,032
12	-0,226	0,177	0,076
13	0,427	0,131	0,241
14	-0,675	0,276	0,869
15	1,308	0,139	2,306
16	0,993	0,168	1,423
			14,8237

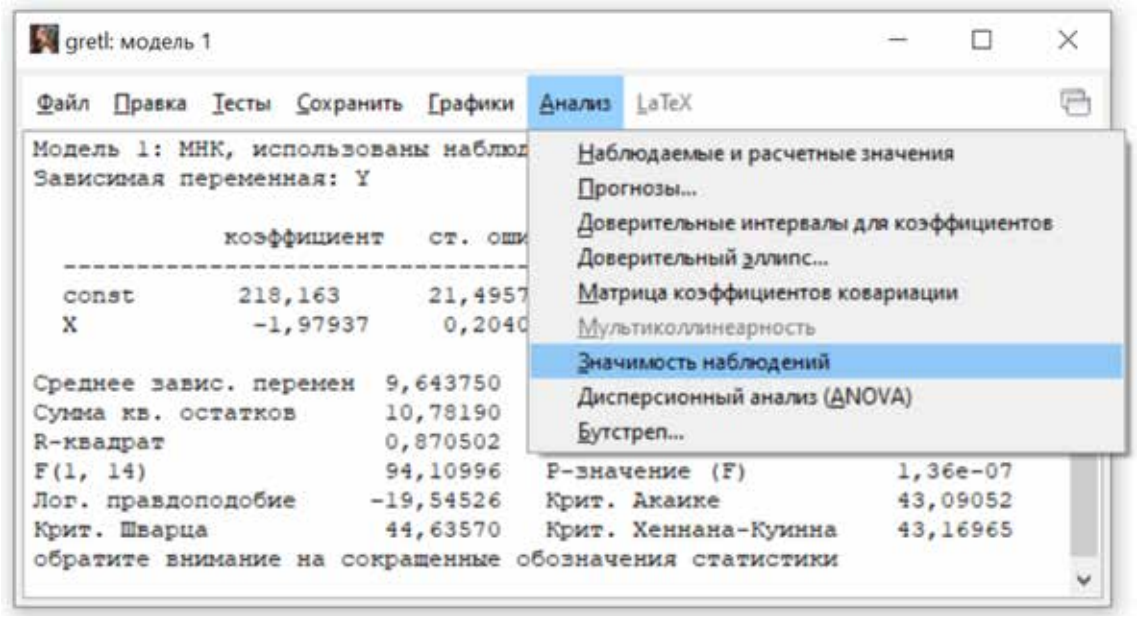


Рис. 4. Оценка параметров модели регрессии в Gretl и выбор вида анализа построенной модели

Таблица 5

Вычисление критерия перекрестной проверки в Gretl

	Остатки u	леверидж 0<=h<=1	Воздействие u*h/(1-h)	DFFITS
1	-0,7518	0,093	-0,07728	-0,286
2	-0,71221	0,095	-0,074644	-0,273
3	-1,1654	0,073	-0,09197	-0,402
4	0,71603	0,160	0,1365	0,386
5	-0,42994	0,086	-0,040616	-0,153
6	-1,0009	0,065	-0,069815	-0,316
7	1,6911	0,198	0,41773	1,260
8	-0,50499	0,072	-0,039302	-0,163
9	0,22738	0,097	0,024302	0,086
10	0,26903	0,090	0,02662	0,098
11	-0,16457	0,080	-0,014278	-0,056
12	-0,22621	0,177	-0,048696	-0,127
13	0,42698	0,131	0,064405	0,197
14	-0,67528	0,276*	-0,25712	-0,554
15	1,3082	0,139	0,21044	0,687
16	0,99255	0,168	0,20025	0,569

('*' указывает на точку левериджа)

Критерий перекрестной проверки = 14,8237

Затем в Gretl была построена двухфакторная модель, спецификация которой имеет вид:

$$Y = b_0 + b_1 X + b_2 X^2 + \varepsilon .$$

После оценки параметров модели в Gretl был вычислен критерий перекрестной проверки *CV*. Несмотря на то, что коэффициент детерминации двухфакторной модели 0,88 больше коэффициента детерминации однофакторной модели (0,87), а стандартная ошибка меньше (0,81 против 0,88), критерий перекрестной проверки *CV*, равный 20,65, больше значения *CV* для однофакторной модели, равного 14,82. В качестве лучшей модели для прогнозирования выбираем однофакторную модель.

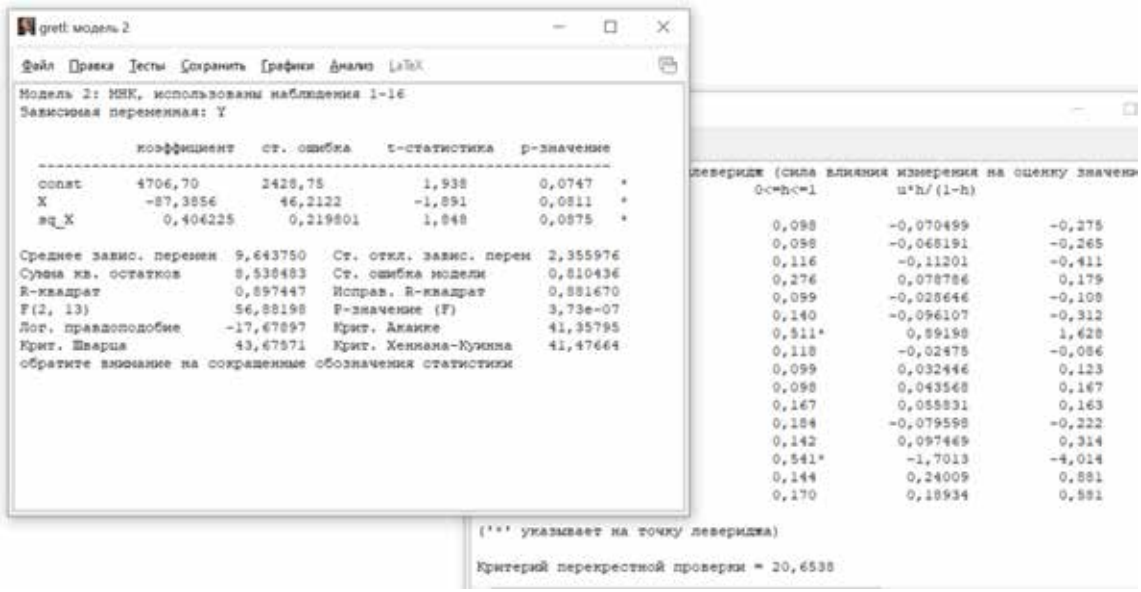


Рис. 5. Оценка параметров и критерия перекрестной проверки двухфакторной модели

Заключение

Рассмотрев некоторые аспекты оценки качества линейных регрессионных моделей, а именно проблему проверки адекватности моделей, можно сделать следующие выводы.

Оценка качества моделей регрессии должна выполняться по нескольким направлениям: оценка значимости всего уравнения регрессии, оценка значимости параметров модели регрессии, оценка точности модели, проверка выполнения предпосылок МНК, и только при положительных результатах по этим пунктам осуществлять проверку качества прогнозов, получаемых на базе обучающей выборки путем сравнения этих прогнозов с реальными значениями из контролирующей выборки, т.е. проверять адекватность модели, ее способность к построению точных прогнозов.

В качестве инструментария могут использоваться различные методы перекрестной проверки, реализованные в R или Gretl, а при небольших выборках можно использовать Excel.

Список литературы

1. Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование: учебное пособие. 3-е изд., перераб. и доп. М.: Вузовский учебник: Инфра-М, 2019. 389 с.

2. Демидова О.А., Малахов Д.И. Эконометрика: учебник и практикум для вузов. М.: Юрайт, 2022. 334 с.

3. Грин У.Г. Эконометрический анализ. Кн. 1 / пер. с англ.; под науч. ред. С.С. Синельникова, М.Ю. Турунцевой. М.: Издательский дом «Дело» РАНХиГС, 2016. 760 с.

4. Кеннеди П. Путеводитель по эконометрике / пер. с англ.; под науч. ред. В.П. Носко. М.: Издательский дом «Дело» РАНХиГС, 2016. 528 с.

5. Айвазян С.А., Фантаццини Д. Эконометрика – 2: продвинутый курс с приложениями в финансах: учебник. М.: Магистр, НИЦ ИНФРА-М, 2018. 944 с.

6. Бабешко Л.О., Бич М.Г., Орлова И.В. Эконометрика и эконометрическое моделирование. 2-е изд., испр. и доп. М.: ООО «Научно-издательский центр ИНФРА-М», 2021. 385 с. DOI: 10.12737/1141216.

7. Единый архив экономических и социологических данных. URL: http://sophist.hse.ruhttps://urait.ru/bcode/380873 (дата обращения: 05.02.2022).

8. Бабешко Л.О. Эконометрическое моделирование спроса на электроэнергию: проверка адекватности // Фундаментальные исследования. 2018. № 12–1. С. 47–52.

9. Salkever, David S. The use of dummy variables to compute predictions, prediction errors, and confidence intervals. Journal of Econometrics, Elsevie. 1976. Vol. 4 (4). P. 393–397.

10. Kurt S. Riedel. A Sherman – Morrison – Woodbury Identity for Rank Augmenting Matrices with Application to Centering", SIAM Journal on Matrix Analysis and Applications. 1992. No. 13. P. 659-662. DOI: 10.1137/0613040 preprint MR1152773.

11. Естественное движение населения в разрезе субъектов Российской Федерации за январь – февраль 2020 года. URL: https://www.gks.ru/free_doc/2019/demo/edn12-19.htm (дата обращения: 05.02.2022).

12. Федеральная служба государственной статистики. URL: https://www.gks.ru/dbscripts/cbsd_internal/DBInet.cgi?pl=1902001(дата обращения: 05.02.2022).