УДК 381.3.06

МНОГОФАКТОРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ С ПОМОЩЬЮ MS EXCEL

Д. М. Смагин, Т. В. Горлова
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E–mail: kafius@kdu.edu.ua

Рассмотрены вопросы анализа множественной линейной регрессии с помощью встроенной функции ЛИНЕЙН MS Excel. Приведен подробный анализ проблемы выбора модели регрессии, оценки параметров выбранной модели, проверки статистических гипотез о параметрах модели и построения доверительных интервалов для этих параметров. Рассмотрен конкретный пример построения функции регрессии с проверкой значимости каждого выборочного коэффициента регрессии, а также статистические характеристики, рассчитываемые функцией ЛИНЕЙН. Дана оценка мере адекватности построенной функции регрессии исходным данным.

Ключевые слова: функция регрессии, линейная регрессия, статистическая гипотеза, выборочные параметры, коэффициент детерминации.

БАГАТОФАКТОРНИЙ РЕГРЕСІЙНИЙ АНАЛІЗ ЗА ДОПОМОГОЮ MS EXCEL

Д. М. Смагін, Т. В. Горлова
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E–mail: kafius@kdu.edu.ua

Розглянуто питання аналізу множинної лінійної регресії за допомогою вбудованої функції ЛИНЕЙН MS Excel. Наданий докладний аналіз проблеми вибору моделі регресії, оцінки параметрів регресійної моделі, перевірки статистичних гіпотез про параметри моделі і побудови довірчих інтервалів для цих параметрів. Розглянуто конкретний приклад побудови функції регресії з перевіркою значущості кожного вибіркового коефіцієнта регрессії, а також статистичні характеристики, що розраховуються функцією ЛИНЕЙН. Надана оцінка міри адекватності побудованої функції регресії початковим даним.

Ключові слова: функція регресії, лінійна регресія, статистична гіпотеза, вибіркові параметри, коефіцієнт детермінації.

АКТУАЛЬНОСТЬ РАБОТЫ. В настоящее время в различных сферах науки и техники к статистическим методам анализа проявляется повышенный интерес. Существует множество специализированных программных средств для статистических расчетов, но наибольшее распространение получила электронная таблица MS Excel, которая установлена практически на каждом компьютере. Выбор Excel как средства реализации статистического анализа (в том числе регрессионного анализа) вызван тем, что эта программа имеет порядка 80 встроенных статистических функций и надстройку *Пакет анализа*, в том числе для регрессионного анализа в Excel имеется две процедуры и восемь встроенных функций. Чтобы многие сложные статистические функции и процедуры MS Excel стали доступны и понятны широкому кругу пользователей–прикладников, возникла необходимость в рассмотрении конкретных примеров с применением этих инструментов, чтобы показать, как правильно их следует использовать в нужных целях.

Цель работы – регрессионный анализ на основе применения функции ЛИНЕЙН.

МАТЕРИАЛ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ. *Выбор модели регрессии*. Регрессионный анализ объединяет практические методы исследования формы корреляционной зависимости между случайными величинами X и Y по данным, полученным в ходе эксперимента [1, 2]. При этом величина X рассматривается как контролируемая величина, значения которой $x_1,x_2,\ldots,x_n$ задаются заранее, при планировании эксперимента, а соответствующие им наблюдаемые значения $y_1, y_2, \ldots, y_n$ рассматриваются как реализации случайной величины $Y= \varphi(x;\beta) + \varepsilon$. Здесь $\varphi(x;\beta)$ – некоторая детерминированная функция аргумента x (*функция регрессии*), $b = (b_0, b_1, \mathbf{K}\, b_m)$ – совокупность неизвестных параметров этой функции, m – число факторов, ε – нормальная случайная величина с нулевым математическим ожиданием и неизвестной дисперсией σ^2, порожденная неучтенными случайными факторами и случайными ошибками измерений. При этом контролируемая переменная X называется *регрессионной переменной* или фактором, а зависимая переменная Y называется *результативным признаком* или *откликом*.

В регрессионном анализе принято различать парную и множественную регрессию. *Парная регрессия* описывает связь между случайной величиной Y и контролируемой исследователем неслучайной величиной x. Множественная регрессия описывает зависимость случайной величины Y от нескольких контролируемых переменных (факторов) $x_1,x_2,\ldots,x_m$.

Регрессионный анализ включает в себя следующие основные этапы:

– выбор модели регрессии;

– оценка параметров выбранной модели регрессии;

– проверка статистических гипотез о параметрах модели и построении доверительных интервалов для этих параметров.

Выбор модели регрессии состоит в выборе функции регрессии: $\overline{y}_x = \varphi(x;\beta)$.

В случае парной регрессии в качестве модели используется линейная функция:

$$\varphi(x; \beta) = \beta_0 + \beta_1 x.$$

В случае множественной линейной регрессии в качестве модели используется линейная функция:

$$\varphi(x;\beta) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m.$$

Эта функция линейна не только относительно контролируемых переменных $x_1, x_2, \ldots, x_m$, но и относительно параметров $\beta_0, \beta_1, \ldots, \beta m$.

Оценка параметров выбранной модели. Выбранная модель регрессии $\bar{y}_x = \varphi(x;\beta)$ всегда содержит некоторое число параметров $\beta_0, \beta_1, \ldots, \beta m$. Оценка этих параметров осуществляется методом наименьших квадратов. При использовании этого метода в качестве выборочных оценок параметров $\beta_0, \beta_1, \ldots, \beta m$ используются такие числа $b_0, b_1, \ldots, b_m$, которые минимизируют сумму $\sum_{i=1}^{n} [y_i - \varphi([x_i; b_0, b_1, \ldots, b_m)]^2$.

Оценки $b_0, b_1, \ldots, b_m$, называются *выборочными параметрами регрессии, о*ценка b_0 – *выборочной постоянной регрессии*, а оценки $b_1, \ldots, b_m$ – *выборочными коэффициентами регрессии.* Функция $\hat{y}(x) = \varphi(x; b_0, b_1, \ldots, b_m)$ называется *эмпирической (выборочной) функцией регрессии*.

Excel предоставляет несколько функций для работы с линейной регрессией,среди которых находится встроенная функция ЛИНЕЙН, обладающая большими возможностями. Эта функция возвращает выборочные параметры $b_0, b_1, \ldots, b_m$ функции регрессии и еще ряд дополнительных выборочных характеристик исследуемой регрессионной зависимости: коэффициент детерминации (R^2), значение критерия Фишера, стандартные ошибки коэффициентов b_i и ряда других показателей. Функция ЛИНЕЙН имеет следующий синтаксис:

ЛИНЕЙН (*известные_ значения_y; известные_ значения_x; конст; статистика*):

– *известные_ значения_y* – множество значений зависимой переменной y, полученных в ходе эксперимента. Этот аргумент может быть одним столбцом, одной строкой или прямоугольным диапазоном ячеек;

– *известные_ значения_x* – множество значений $x_1, x_2, \ldots, x_n$ каждой из контролируемых переменных (факторов) $x_1, x_2, \ldots, x_m$. Причем, если массив *известные_ значения_y* имеет один столбец, то каждый столбец массива *известные_ значения_x* интерпретируется как отдельная переменная. В случае множественной регрессии, если массив *известные_ значения_y* представляет собой вектор–столбец размера n×1, то массив *известные_ значения_x* должен иметь n строк и m столбцов. При этом каждый столбец массива должен содержать n экспериментальных значений каждого фактора;

– к*онст* – логическая переменная , определяющая, следует ли включить в уравнение регрессии постоянную регрессии b_0 (аргумент не обязательный). Если аргумент *конст* =1 (ИСТИНА) или опущен, то вычисляются все коэффициенты $b_0, b_1, \ldots, bm$. Если аргумент *конст*=0 (ЛОЖЬ), то предполагается, что b_0 =0.При этом выборочное уравнение регрессии принимает вид $\hat{y}(x) = b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$;

– *Статистика* – аргумент, принимающий логическое значение, которое указывает, требуется ли рассчитывать дополнительные статистические характеристики регрессии. Если этот аргумент имеет значение ИСТИНА или 1, то помимо коэффициентов $b_0, b_1, \ldots, bm$ функция ЛИНЕЙН выдает дополнительную информацию об исследуемой регрессионной зависимости (строки 2 – 5, строки табл. 1). Если аргумент Статистика имеет значение ЛОЖЬ , 0 или опущен, то функция возвращает только значения коэффициентов $b_0, b_1, \ldots, bm$.

Отметим, что функция ЛИНЕЙН возвращает массив значений, поэтому должна задаваться в виде формулы массива, в противном случае будет выведено значение только коэффициента bm.

Таблица 1 – Выходной массив данных функции ЛИНЕЙН

bm	bm_{-1}	–	b_2	b_1	b_0
SEm	SEm_{-1}	–	SE_2	SE_1	SE_0
R^2	SE_y				
f	df				
SSрег	SSост				

Остальные ячейки этого массива заполняются значениями #Н/Д.

Таблица 2 – Статистические характеристики, рассчитываемые функцией ЛИНЕЙН

Статистика	Описание
$SEm, \ldots, SE_1, SE_0$	Среднеквадратические отклонения для выборочных параметров регрессии $b_0, b_1, \ldots, bm$
R^2	Коэффициент детерминации, который вычисляется по результатам сравнения фактических значений Y и значений $\hat{Y}$, получаемые из уравнения регрессии.
SE_y	Среднеквадратическое отклонение нормальной случайной величины ε. SE_y^2 – дисперсия ε
f	Расчетное (выборочное) значение статистики F, используется при оценке значимости R^2
df	Число степеней свободы знаменателя статистики F (df = n – p)
SSрег	Сумма квадратов, обусловленная регрессией, $SSрег = \Sigma(\hat{y}_i - \bar{y})^2$
SSост	Сумма квадратов остатков $SSост = \Sigma(y_i - \hat{y}_i)^2$

Проверка статистических гипотез о параметрах модели регрессии. Выборочный коэффициент детерминации R^2 и выборочные оценки b_0 , b_1 , …, b_m параметров регрессии , вычисленные по ограниченному числу экспериментальных данных, всегда содержат элементы случайности и, по существу , сами являются случайными величинами. В связи с этим возникает необходимость *проверки значимости* этих выборочных числовых характеристик.

Критерий проверки адекватности функции регрессии состоит в проверке значимости коэффициента детерминации R^2. Для этого выдвигается гипотеза H_0: $R_Г^2 = 0$ о том, что коэффициент детерминации $R_Г^2$ генеральной совокупности, из которой извлечена исследуемая выборка, равен нулю. Эта гипотез равносильна гипотезе H_0: $\beta_1 = \beta_2 = \ldots = \beta m = 0$ о том, что ни один из факторов, включенных в регрессию, не оказывает существенного (значимого) влияния на отклик Y. В качестве альтернативы рассматривается гипотеза H_1: $R_Г^2 \neq 0$, т.е. хотя бы один коэффициент регрессии $b_i \neq 0$. При проверке этой гипотезы используется статистика:

$$F = \frac{MS_{рег}}{MS_{ост}} = \frac{R^2}{1-R^2}\frac{(n-p)}{(p-1)},$$

где p – число параметров регрессии. В случае, когда в модель регрессии включена постоянная регрессии β_0, тогда p=m+1, если $\beta_0 = 0$, то p= m. Если выполнены все предпосылки регрессионного анализа, то статистика F имеет F – распределение (p – 1) и (n – p) степенями свободы.

Коэффициент детерминации R^2 и статистику F вычисляет функция ЛИНЕЙН (табл. 1).

При проверке гипотезы H_0 используется правосторонняя критическая область $\Omega[f\ (\alpha;p-1,n-p),\infty)$, где Fкр= f (α;p–1,n–p) – критическое значение порядка α распределения Фишера – Снедекера с (p – 1) и (n – p) степенями свободы.

Используя расчетные (выборочные) значения f статистики F, проверим гипотезу H_0 об отсутствии регрессионной зависимости между переменной y и x. Практическая реализация проверки гипотезы H_0 с помощью MS Excel не вызывает затруднений, нам остается найти только квантиль F – распределения. Для этого зададим уровень значимости α и с помощью формулы =FРАСПОБР(α; p–1, n–p), вычислим Fкр = f (α;p–1,n–p) – *критическое значение порядка a* , F – распределения с (p–1) и (n–p) степенями свободы. Тогда правосторонняя критическая область определяется неравенством f > Fкр, а область принятия нулевой гипотезы – неравенством

f < Fкр, если согласно данным функции ЛИНЕЙН расчетное значение f статистики F превышает ее критическое значение Fкр, т.е. f > Fкр, тогда нулевая гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1. В этом случае коэффициент детерминации R^2 значимо отличается от нуля, функция регрессии статистически значима и адекватно описывает исходные данные.

Проверка значимости параметров регрессии. Предположим, что значения коэффициентов b_i и их среднеквадратические отклонения SEi (i = 0,1,…,m) уже вычислены функцией ЛИНЕЙН. При проверке значимости параметров регрессии b_i , i = 0,1,…,m , выдвигается гипотеза $H_{0(i)}$: βi = 0 о том, что фактор x_i , не оказывает заметного влияния на отклик Y. В качестве альтернативы рассматривается гипотеза $H_{0(i)}$: βi ≠0. Критерии проверки гипотез о значимости коэффициентов функции регрессии строятся на том основании, что отношение вычисленного коэффициента к его среднеквадратическому отклонению T= b_i/ SE_i (i = 0,1,…,m) имеет распределение Стьюдента с (n– p) степенями свободы.

При проверке рассматриваемой гипотезы используется двухсторонняя критическая область Ω(|T| ≥ t(α/2;n–p)), где Ткр = t(α/2;n–p) – критическое значение порядка α/2 распределения Стьюдента с (n–p) степенями свободы При проверке гипотез $H_{0(i)}$: βi = 0: .

– вычисляется критериальная статистика, как модуль отношения значения этого коэффициента к его среднеквадратическому отклонению, т.е. величина t_i = | bi / SEi | , i = 0,1,…,m;

– по заданному уровню значимости α вычисляется критическое значение Ткр статистики Т как квантиль порядка α/2 распределения Стьюдента с помощью формулы = СТЬЮДРАСПОБР(α/2;(n–p)).

Если t_i < Ткр принимается гипотеза $H_{0(i)}$, что коэффициент βi = 0. В противном случае считается, что этот коэффициент значимо отличается от нуля.

Пример. Найти выборочное уравнение линейной регрессии Y на X по данным, приведенным в корреляционной таблице на рис. 1 в диапазоне A1:E12

Проанализируем эти данные с помощью функции ЛИНЕЙН. Для этого:

1. Выделим диапазон H3: L7, введем в него формулу массива =ЛИНЕЙН(E2 : E12 ; A2 : D12 ; 1 ; 1) и нажмем комбинацию клавиш Ctrl+Shift+Enter. В диапазоне H3:L7 появятся характеристики исследуемой регрессионной зависимости (табл. 3).

2. Используя результаты, приведенные в диапазоне H3:L3, запишем выборочное уравнение множественной регрессии:

$$\hat{y}(x) = 56{,}67999 + 0{,}025524\,x_1 + 12{,}60916\,x_2 + 2{,}718465x_3 - 0{,}23168x_4. \quad (1)$$

Большое значение коэффициента детерминации R^2 =0,9965 (ячейка H5) свидетельствует о сильной зависимости отклика Y от включенных в анализ контролируемых переменных (факторов).

Используя дополнительные характеристики (SE_i, R^2, f, SS – ячейки H4 : L7) исследуемой регрессионной зависимости, можно оценить значимость коэффициента детерминации R^2 и коэффициентов β_1 , β_2 , β_3 , β_4 уравнения регрессии (1).

Оценим вначале значимость коэффициента детерминации R^2 . Зададимся уровнем значимости α = 0,05.С помощью функции FРАСПОБР(0,05; 4; 6) вычислим критическое значение F – распределения с числом степеней свободы 4 и 6: f (0.05; 4; 6) = 4,534 (число степеней свободы числителя (p–1) = 4, число степеней свободы знаменателя df = 6 (ячейка I6)). Сравним расчетное (выборочное) значение f = 434.74 статистики F (ячейка H6) с ее критическим значением f(0.05;4;6)= 4.534. Расчетное значение 434.74 значительно превышает критическое значение 4.534. Это означает, что коэффициент детерминации R^2 значимо отличается от нуля, поэтому гипотезу H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ об отсутствии регрессионной зависимости между переменными y и x_1, x_2, x_3, x_4 следует отклонить как противоречащую фактическим данным наблюдения.

Таблица 3 – Корреляционная характеристика исследуемой регрессии

	A	B	C	D	E	F	G	H	I	J	K	L
1	x_1	x_2	x_3	x_4	y							
2	2310	2	2	20	142			b_4	b_3	b_2	b_1	b_0
3	2333	2	2	12	144		b_i	-0,23168	2,718465	12,60916	0,025524	56,67999
4	2356	3	1,5	33	151		SEi	0,013682	0,546243	0,412994	0,005591	12,59965
5	2379	3	2	43	150		R^2	0,996562	1,001653	#Н/Д	#Н/Д	#Н/Д
6	2400	2	3	53	139		f	434,7386	6	#Н/Д	#Н/Д	#Н/Д
7	2425	4	2	23	169		SS	1744,707	6,019849	#Н/Д	#Н/Д	#Н/Д
8	2448	2	1,5	99	126			$SS_{рег}$	$SS_{ост}$			
9	2470	2	2	34	142							
10	2494	3	3	23	163							
11	2517	4	4	55	169							
12	2540	2	3	22	149							

Оценим теперь значимость выборочных коэффициентов регрессии β_1, β_2, β_3, β_4, Зададимся уровнем значимости α =0,05. С помощью функции СТЬЮДРАСПОБР(0,05;6) вычислим критическое значение распределения Стьюдента с (n–p) = 6 степенями свободы порядка $\alpha/2$ = 0,025, т.е. величину Ткр = t (0.025;6) = 2,447. Число степеней свободы вычислено по формуле n – p = 11 – 5 = 6, где n – число экспериментов, p – количество параметров (количество выборочных коэффициентов регрессии).

Используя формулу $t = b_i / SEi$, вычислим расчетное значение t статистики T для каждого из выборочных коэффициентов регрессии b_i:

bi	b_4	b_3	b_2	b_1	b_0
t	–16,933	4,97666	30,53109	4,565229	4,498537

ВЫВОДЫ. Все расчетные значения статистики T превышают по абсолютной величине ее критическое значение 2,447. Это означает, что контролируемые переменные (факторы), используемые в уравнении регрессии (**), существенно влияют на отклик Y. В заключение отметим, что встроенная функция ЛИНЕЙН MS Excel позволяет просто и эффективно осуществлять многофакторный регрессионный анализ с оценкой значимости выборочных коэффициентов уравнения регрессии и коэффициента детерминации.

ЛИТЕРАТУРА

1. Макарова Н. В. Трофимец В. Я. Статистика в Excel. – М.: Финансы и статистика, 2002.

2. Минько А.А.Статистический анализ в MS Excel. – М.: Издательский дом ”Вильямс”, 2004.

MULTIVARIATE REGRESSION MS EXCEL ANALYSIS

D. Smagin, T. Gorlova

Kremenchuk Mykhailo Ostrohradskyi National University

vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E–mail: kafius@kdu.edu.ua

The problems of multiple linear regression analysis using the built-in LINEST MS Excel are considered. The detailed analysis of the problems of regression model choosing, estimation of parameters of the selected model, and also verification of statistical hypotheses of model’s parameters is presented. A specific example of definition of regression function with test of significance of every sampling coefficient of the regression sample is discussed. The conformity degree of the regression function built to the original data is estimated.

Key words: regression function, linear regression, statistical hypothesis, sampling parameters.

REFERENCES

1. Makarova N.V. Trofimets V.Y. *Statistics in Excel.* – Moscow: Finance and Statistics, 2002. – 348 p. [in Russian]

2. Minko A.A. *MS Excel statistical analysis.* – Moscow: Publishing House "Williams", 2004. – 448 p. [in Russian]

Стаття надійшла 16.10.2012.
Рекомендовано до друку
д.т.н., проф. Гученко М.І.