УДК 378(075.8):330.4

ПРОВЕРКА ГИПОТЕЗ, ОТНОСЯЩИХСЯ К КОЭФФИЦИЕНТАМ РЕГРЕССИИ

РЕГРЕССИЯ КОЭФФИЦИЕНТТЕРІНЕ ҚАТЫНАСТЫ ГИПОТЕЗАНЫ ТЕКСЕРУ

VERIFICATION OF HYPOTHESES RELATED TO COEFFICIENTS OF REGRESSION

Б.Д. ШАРИПОВА, Ж. БЕГУЛИЕВА, I. ОМАРБАЙ
Б.Д. ШАРИПОВА, Ж. БЕГУЛИЕВА, I. ӨМАРБАЙ
B.D. SHARIPOVA, G. BEGULIEVA, I. OMARBAI

(Алматинский технологический университет)
(Алматы технологиялық университеті)
(Almaty Technological University)
E-mail: Birjan2103@mail.ru

В данной статье исследованы эконометрические показатели.

Новизна работы заключается в том, что получены оценки значимости коэффициента уравнения регрессии, качество модели и построения доверительного интервала.

Получены результаты по статье, которые определяют свойства коэффициентов уравнений регрессии и проверка гипотез относящихся к коэффициентам регрессии. Проведены примеры и анализы показателей регрессии. Рассмотрены условия и предложения, лежащие в основе регрессионного анализа.

Практическая значимость исследования работы — определение взаимозависимости показателей регрессии и использование показателей в дальнейших исследованиях.

Приведены примеры зависимости расходов на питание от личного дохода для модели регрессии без свободного члена и на этой основе рассчитаны стандартные ошибки коэффициента регрессии. Также дана оценка значимости коэффициента регрессии, качество модели и построен доверительный интервал. Определена взаимозависимость критериев.

Бұл мақалада эконометрикалық көрсеткiштер зерттелген.

Жұмыстың жасалынған есептеулерiнiң жаңалығы сол, регрессия теңдеуiнiң коэффициенттерiнiң маңыздылығының бағалаулары, модельдiң сапалығы және сенiмдiлiк интервалдарының құрылуы болып саналады.

Мақалада алынған шешiмдер теңдеудiң коэффициенттерiнiң қасиеттерiн анықтайды және теңдеудiң коэффициенттерiне қатысты гипотезаларын тексередi. Мысалдар қарастырылған және регрессия көрсеткiштерiнiң талдаулары жасалынған. Регрессия талдауларының негiзiнде орналасқан шарттары мен ұсыныстары қарастырылған.

Келтiрiлген мысалда қарастырылған регрессия моделiнiң бос мүшесiсiз болғандағы, тамақтануға кеткен шығын мен жеке кiрiстiң арасындағы байланыс келтiрiлген және соның негiзiнде регрессия коэффициенттерiнiң стандарттық қателiктерi есептелiнген. Сонымен қатар, регерссияның коэффициенттерiнiң маңыздылығы бағаланған, модельдiң сапасы мен сенiмдiлiк шектеулiктерi құрылған. Критерийлердiң өзара байланыстары анықталды.

In this article econometric indicators are investigated.

The novelty of work is that estimates of the importance of coefficient of the equation of regression, quality of model and creation of a confidential interval are received.

Results on article which define properties of coefficients of the equations of regression and check of the hypotheses relating to regression coefficients are received. Examples and analyses of indicators of regression are carried out. The conditions and offers which are the cornerstone of the regression analysis are considered.

The practical importance of a research of work, this definitions interdependence of indicators of regression and use of indicators in further researches.

Examples of dependence of expenses on food on personal income for regression model without free member are given and on this basis standard errors of coefficient of regression are calculated. An assessment of the importance of coefficient of regression, quality of model is also given and the confidential interval is constructed. The interdependence of criteria is defined.

Ключевые слова: регрессия, корреляция, гипотеза, доверительные интервалы, математическое ожидание, критерий, несмещенность, дисперсионный анализ, детерминация.

Негізгі сөздер: регрессия, корреляция, гипотеза, сенімді интервалдар, математикалық күтілім, критерий, ауыстырмайтын, дисперсиялық талдау, детерминация.

Key words: regression, correlation, hypothesis, confidence intervals, expected value, criterion, nesmeschennost', analysis of variance, determinaciya.

Введение

Целью данного исследования, осуществляемого в настоящее время, является использование его результатов в будущем, или, иначе говоря, прогнозирование состояния изучаемого явления. При этом, желая изучать явление во взаимосвязи с другими явлениями или величинами, приходится выделять некоторые из них, влияющие на изучаемое, оценивать степень и «качество» влияния, то есть характер связи между изучаемым (основным в данном исследовании) и влияющими на него величинами качественного или количественного характера.

Эконометрика как система специфических методов начала развиваться с осознания своих задач – отражения особенностей экономических переменных и связей между ними. В уравнения регрессии начали включаться переменные не только первой, но и второй степени – с целью отразить свойство оптимальности экономических переменных: наличие значений, при которых достигается минимаксное воздействие на зависимую переменную.

Регрессионный анализ — статистический метод исследования взаимосвязи переменных. Проверки статистических гипотез — класс самых трудных задач в математической статистики.

Объекты и методы исследований

Объектами исследования является регрессионо-корреляционный анализ. В данной статье используются эконометрические методы. Рассмотрим свойства коэффициентов регрессии и проверка гипотез.

Величина Y в модели регрессии $Y=\alpha+\beta X+\varepsilon$ имеет две составляющие: *неслучайную*($\alpha+\beta X$) *и случайную*(ε).

Оценки коэффициентов регрессии (a,b) являются линейными функциями Y, и теоретически их также можно представить в виде двух составляющих [1].

Воспользовавшись разложением показателей

$$\mathrm{cov}(x,y) = \mathrm{cov}(x, \alpha+\beta X+\varepsilon) = \beta \mathrm{var}(x)+\mathrm{cov}(x,\varepsilon) \quad (1)$$

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum(\alpha + \beta x_i + \varepsilon_i) = \alpha + \beta\bar{x} + \frac{1}{n}\sum \varepsilon_{i,} \quad (2)$$

получим преобразованные соотношения для (a,b);

$$\begin{cases} b = \beta + \frac{\mathrm{cov}(x,\varepsilon)}{\mathrm{var}(x)}, \\ a = \alpha + \left[\frac{1}{n}\sum \varepsilon_i - \frac{\bar{x}\mathrm{cov}(x,\varepsilon)}{\mathrm{var}(x)}\right]. \end{cases} \quad (3)$$

Таким образом, коэффициенты (a,b) разложены на две составляющие: *неслучайную*, равную истинным значениям (α,β), и *случайную,* зависящую от ε.

На практике нельзя разложить коэффициенты регрессии на составляющие, так как значения (α,β) или фактические значения ε в выборке неизвестны.

Линейная регрессионная модель с двумя переменными имеет вид

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = \overline{1,n}), \quad (4)$$

где Y – объясняемая переменная, X – объясняющая переменная, ε – случайный член.

Методом исследований является метод наименьших квадратов.

Для того, чтобы регрессионный анализ, основанный на МНК, давал наилучшие из всех возможных результаты, должны выполняться определенные условия по условию Гаусса – Маркова.

При выполнении условий Гаусса – Маркова модель называется классической нормальной линейной регрессионной моделью [2].

Наряду с условиями Гаусса – Маркова обычно предполагается, что *случайный член распределен нормально*, т.е. $\varepsilon_i \sim N(0;\sigma^2)$.

Рассмотрим условия и предложения, лежащие в основе регрессионного анализа.

Первое условие означает, что случайный член не должен иметь систематического смещения. Если постоянный член включен в уравнение регрессии, то это условие выполняется автоматически.

Второе условие означает, что дисперсия случайного члена в каждом наблюдении имеет только одно значение.

Под дисперсией σ^2 имеется в виду возможное поведение случайного члена до того, как сделана выборка. Величина σ^2 неизвестна, и одна из задач регрессионного анализа состоит в ее оценке.

Условие *независимости* дисперсии случайного члена от номера наблюдения называется гомоскедастичностью (что означает одинаковый разброс). Зависимость дисперсии случайного члена от номера наблюдения называется *гетероскедастичностью.*

Характерные диаграммы рассеяния для случаев гомоскедастичности показаны на рисунке 1 *а* и *б* соответственно.

Если условие гомоскедастичности не выполняется, то оценки коэффициентов регрессии будут *неэффективными*, хотя и *несмещенными*.

Существуют специальные методы диагностирования и устранения гетероскедастичности.

Если условие независимости случайных членов не выполняется, то оценки коэффициентов регрессии, полученные по МНК, оказываются *неэффективными*, хотя и *несмещенными*.

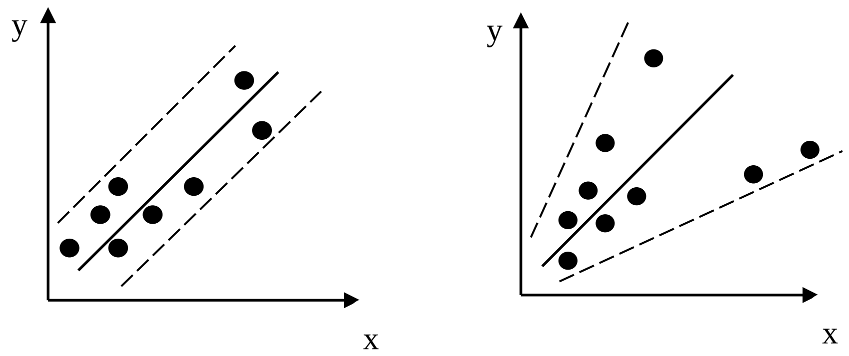

Рисунок 1 – Характерные диаграммы рассеяния для случаев гомоскедастичности.

Если условие о неслучайности объясняющей переменной не выполняется, то оценки коэффициентов регрессии оказываются *смещенными* и *несостоятельными*.

В регрессионном анализе часто вместо условия *о неслучайности объясняющей переменной* используется более слабое *условие о независимости (некоррелированности) распределений объясняющей переменной и случайного члена*. Получаемые при этом оценки коэффициентов регрессии обладают теми же основными свойствами, что и оценки, полученные при использовании условия о неслучайности объясняющей переменной.

Предположение о нормальности распределения случайного члена необходимо для проверки значимости параметров регрессии и для их интервального оценивания.

Если условия регрессионного анализа выполняются, то оценки (a,b), сделанные с помощью МНК, являются наилучшими линейными несмещенными оценками, т.е. обладают следующими свойствами:

$$M[\operatorname{cov}(x,\varepsilon)] = M\left[\frac{1}{n}\sum(x_i-\bar{x})\varepsilon_i\right] = \frac{1}{n}\sum(x_i-\bar{x})M(\varepsilon_i) = 0, \qquad (5)$$

$$D[\operatorname{cov}(x,\varepsilon)] = D\left[\frac{1}{n}\sum(x_i-\bar{x})\varepsilon_i\right] = \frac{1}{n^2}\sum(x_i-\bar{x})^2 D(\varepsilon_i) = \frac{\operatorname{var}(x)}{n}\sigma^2,$$

$$M\left[\frac{\sum\varepsilon_i}{n}\right] = \frac{\sum M(\varepsilon_i)}{n} = 0, \qquad (6)$$

$$D\left[\frac{\sum \varepsilon_i}{n}\right] = \frac{\sum D(\varepsilon_i)}{n^2} = \frac{\sigma^2}{n}, \tag{7}$$

$$M\left[\frac{\bar{x}\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \frac{\bar{x}}{\operatorname{var}(x)} M[\operatorname{cov}(x,\varepsilon)] = 0, \tag{8}$$

$$D\left[\frac{\bar{x}\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \frac{(\bar{x})^2}{\operatorname{var}^2(x)} D[\operatorname{cov}(x,\varepsilon)] = \frac{(\bar{x})^2\sigma^2}{n\operatorname{var}(x)}. \tag{9}$$

Вычислим математическое ожидание и дисперсию оценок b и a:

$$M(b) = \beta + M\left[\frac{\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \beta + \frac{M[\operatorname{cov}(x,\varepsilon)]}{\operatorname{var}(x)} = \beta, \tag{10}$$

$$D(b) = D\left[\frac{\operatorname{cov}(x,\varepsilon)}{\operatorname{var}(x)}\right] = \frac{D[\operatorname{cov}(x,\varepsilon)]}{\operatorname{var}^2(x)} = \frac{\sigma^2}{n\operatorname{var}(\mathrm{x})}; \tag{11}$$

Рассчитаем стандартные ошибки коэффициентов регрессии.

Полученные теоретические дисперсии $D(a)$, $D(b)$ зависят от дисперсии σ^2 случайного члена.

По данным выборки отклонения ε_i, а следовательно, и их дисперсии σ^2 неизвестны, поэтому они заменяются наблюдаемыми остатками ε_i и их выборочной дисперсией.

Пример. Построим зависимость расходов на питание y от личного дохода x для модели регрессии *без свободного члена* и рассчитаем стандартную ошибку коэффициента регрессии.

Исходные данные и расчетные показатели даны в таблице 1:

Таблица 1 – Регресионная зависимость между переменными x и y

Годы	2012	2013	2014	2015	2016
X	2	6	10	14	18
Y	1	2	4	11	12

Коэффициенты b = 0.642; a = 0. Следовательно, $\hat{y} = 0{,}642x$.

Остаточная дисперсия S^2 и стандартная ошибка регрессии S равны соответственно

S = 1.843

Стандартная ошибка коэффициента регрессии $S_b = 0.071$.

Статистические свойства МНК – оценок (a,b)

Пусть выполняется условие нормальности распределения случайного члена: $\varepsilon_i \sim N(0;\sigma^2)$. Тогда МНК – оценки коэффициентов регрессии также имеют нормальное распределение, поскольку являются линейными функциями от ε_i, т.е.

$$a \sim N\left(\alpha; \frac{\sigma^2\overline{x^2}}{n\operatorname{var}(x)}\right); \quad b \sim N\left(\beta; \frac{\sigma^2}{n\operatorname{var}(x)}\right).$$

Если условие нормальности распределения случайного члена не выполняется, то оценки (a,b) имеют асимптотически нормальное распределение.

Проверим гипотезу, относящуюся к коэффициентам регрессии (A,B)

Проверка гипотезы $H_0: \beta = \beta_0$

Пусть в теоретической зависимости: $Y = \alpha + \beta x + \varepsilon$

случайный член ε распределен нормально с неизвестной дисперсией σ^2.

Задача заключается в проверке гипотезы H_0 на основании выборочных данных.

Пусть по выборочным данным получена оценка b.

В качестве критерия проверки гипотезы H_0 принимают случайную величину

$$t = \frac{b - \beta_0}{S_b},$$

которая имеет распределение Стьюдента с $v = n - 2$ степенями свободы.

Вычисляется наблюдаемое значение критерия *t*. По таблице критических точек распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $v = n - 2$ находят критическую точку $t_{\text{кр}}$.

Сравнивая наблюдаемое значение критерия с критическим, можно принять или отвергнуть нулевую гипотезу.

Результаты оценивания регрессии совместимы не только с конкретной гипотезой $H_0: \beta = \beta_0$, но и с некоторым их множеством.

Посредине интервала лежит величина b. Границы интервала одинаково стоят от b, зависят от выбора уровня значимости и являются случайными числами.

Доверительный интервал покрывает значение параметра β с заданной вероятностью (1-α), т.е. $P(b - t_{\text{кр}} s_b < \beta < b + t_{\text{кр}} s_b) = 1 - \alpha$.

Проверка гипотезы $H_0: \beta = 0$

Пусть по выборке получена оценка коэффициента регрессии b.

Гипотеза $H_0: \beta = 0$ используется для установления значимости коэффициента регрессии b.

Соответствующая t-статистика есть оценка коэффициента регрессии, деленная на ее стандартную ошибку, т.е.

$$t = \frac{b - 0}{s_b} = \frac{b}{s_b}.$$

Величина t имеет распределение Стьюдента с $v = n - 2$ степенями свободы.

Из сравнения *значимость* t с заданным *стандартным уровнем значимости* получаем:

- если значимость t больше стандартного уровня, то b не значим;
- если значимость t меньше стандартного уровня, то b значим [3].

Коэффициенты регрессии a= *0* и b = 0.642. Направление связи между y и x определяет знак коэффициента регрессии b = 0.642, т.е. связь является прямой и положительной. Коэффициент b = 0.642 показывает, что при увеличении душевного дохода на 1 усл. ед. расходы на питание увеличиваются на b = 0.642 усл.ед.

Оценим значимость коэффициентов полученной модели. Значимость коэффициентов (a,b) проверяется по t- тесту:

P-значение(a) = 0,00080<0,01<0,05;

P-значение(b) = 0,00016<0,01<0,05.

Результаты и их обсуждение

Коэффициенты (a,b) регрессии значимы при 1%-ном уровне, а тем более при 5%-ном уровне значимости. Таким образом, коэффициенты регрессии значимы и модель адекватна исходным данным.

Результаты оценивания регрессии совместимы не только с полученными значениями коэффициентов регрессии, но и с некоторым их множеством (доверительным интервалом). С вероятностью 95% доверительные интервалы для коэффициентов есть (38,16 – 93,68) для a и (0,0728 – 0,142) для b.

Качество модели оценивается коэффициентом детерминации R^2. Величина $R^2 = 0{,}884$ означает, что фактором душевого дохода можно объяснить 88,4% вариации (разброса) расхода на питание.

Значимость R^2 проверяется по ***F***-тесте: *Значимость* F = 0,00016<0,01<0,05.

Следовательно, R^2 значим при 1%-ном уровне, а тем более при 5%-ном уровне значимости.

В случае парной линейной регрессии коэффициент корреляции можно определить как $r = \sqrt{R^2} = 0{,}94$. Полученное значение коэффициента корреляции свидетельствует, что связь между расходами на питание и душевным доходом очень тесная.

Определим взаимозависимость критериев.

В парном регрессионном анализе эквивалентны t – критерий для $H_0: \beta = 0$; t – критерий для $H_0: \rho = 0$; F – критерий для R^2:

$$t_b = \frac{b}{s_b}, t_r = r\sqrt{\frac{n-2}{1-r^2}}, F = \frac{R^2(n-2)}{1-R^2}.$$

Связь между критериями выражается равенством $t_b = t_r = \sqrt{F}$, причем для критических значений критериев при любом уровне значимости $(t_b)_{\text{кр}} = (t_r)_{\text{кр}} = \sqrt{F_{\text{кр}}}$, и эти критерии дают один и тот же результат.

Выводы

Коэффициенты (a,b) значимы при 1% - ном уровне, а тем более при 5%-ном уровне значимости. Таким образом, коэффициенты регрессии значимы и модель адекватна исходным данным.

Результаты оценивания регрессии совместимы не только с полученными значениями коэффициентов регрессии, но и с некоторым их множеством (доверительным интервалом). С вероятностью 95% доверительные интервалы для коэффициентов.

СПИСОК ЛИТЕРАТУРЫ

1. Шарипова Б.Д., Жаркынбаев С. Эконометрия. - Алматы: Әрекет-Принт, 2009. - 147 с.

2. Доугерти. Введение в эконометрику/ пер. с англ.- М.: ИНФРА-М,2001. - 220 с.

3. Бородич С.А. Эконометрика. - Минск: Новое знание, 2012. - 310 с.