

Эконометрика

Полковников Александр Александрович

Волжский политехнический институт (филиал)
ФГБОУ ВПО "Волгоградский государственный технический университет"

Конспект лекций для студентов направления
"Экономика"

Проверка гипотез. Доверительные интервалы

Если остаток e имеет нормальное распределение, то TSS/De имеет распределение χ^2 с $df = n - 1$ степенью свободы (degree of freedom).

Из $(n - 1)$ -й степени свободы на вариацию, объясняемую регрессией, приходится $m = 1$ степень свободы, где $m + 1 = 2$ — количество параметров регрессии (a и b). Остальные $n - 2$ степени свободы приходятся на остаточную вариацию.

В пакетах статистического анализа (например, в Excel) в распечатках результатов приводятся значения сумм квадратов отклонений на одну степень свободы, так называемые **средние квадраты** (mean squares):

$$EMS = \frac{1}{m} ESS, \quad RMS = \frac{1}{n - m - 1} RSS.$$

В случае парной линейной регрессии:

$$EMS = ESS, \quad RMS = \frac{1}{n - 2} RSS.$$

Далее имеет смысл выяснить, сколь близким к нулю должно быть значение R^2 , чтобы можно было говорить об отсутствии линейной связи между переменными.

Для этого рассмотрим F -статистику:

$$F = \frac{ESS}{RSS/(n-2)},$$

имеющую распределение Фишера с числом степеней свободы 1 и $(n-2)$.

Величина F связана с коэффициентом детерминации:

$$F = (n-2) \frac{ESS}{RSS} = (n-2) \frac{R^2}{1-R^2}.$$

Проверим гипотезу H_0 о незначимости уравнения регрессии, т. е. о близости к нулю коэффициента детерминации.

Задаем уровень значимости α . Вычисляем квантиль $F_{1-\alpha}[1; n - 2]$ уровня $(1 - \alpha)$ распределения Фишера со степенями свободы 1 и $n - 2$.

Если $F > F_{1-\alpha}[1; n - 2]$, то гипотеза H_0 отвергается, уравнение регрессии значимо.

Если $F < F_{1-\alpha}[1; n - 2]$, то гипотеза H_0 не отвергается, уравнение регрессии не значимо.

Статистические пакеты, выполняющие регрессионный анализ, приводят помимо вычисленного значения F и соответствующее ему P -значение, т. е. вероятность

$$P(F_{1-\alpha}[1; n-2] > F).$$

Правило отвержения гипотезы H_0 при превышении F -статистикой порогового уровня $F_{1-\alpha}[1; n-2]$ соответствует отвержению этой гипотезы при выполнении неравенства

$$P\text{-значение} < \alpha.$$

В линейной регрессии оценивается не только значимость уравнения в целом, но и значимость отдельных параметров уравнения.

Для проверки гипотезы $H_0: b = 0$ о несущественности коэффициента регрессии b при альтернативе $H_1: b \neq 0$ рассмотрим стандартную ошибку коэффициента регрессии m_b :

$$m_b = \sqrt{\frac{RMS}{nS_x^2}} = \sqrt{\frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{(n-2) \sum_{k=1}^n (x_k - \bar{x})^2}}.$$

Статистика $t_b = b/m_b$ имеет распределение Стьюдента с $(n-2)$ степенями свободы, т. к.

$$t_b^2 = \frac{b^2}{m_b^2} = \frac{b^2 n S_x^2}{RMS} = (n-2) \frac{ESS}{RSS} = F$$

имеет распределение Фишера со степенями свободы 1 и $(n-2)$.

Данный критерий был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны, статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student».

Задаем уровень значимости α . Вычислим квантиль $t_{1-\alpha/2}[n-2]$ распределения Стьюдента уровня $(1 - \alpha/2)$ с $(n - 2)$ степенями свободы.

Если $|t_b| > t_{1-\alpha/2}[n-2]$, то гипотезу о несущественности коэффициента регрессии отклоняем. Коэффициент b существенно отличается от нуля.

Если $|t_b| < t_{1-\alpha/2}[n-2]$, то гипотезу о несущественности коэффициента регрессии не отклоняем. Коэффициент b не существенно отличается от нуля.

На основе оценки b метода наименьших квадратов можно построить доверительный интервал для коэффициента регрессии с доверительной вероятностью $\beta = 1 - \alpha$:

$$\left[b - t_{1-\alpha/2}[n-2] \cdot m_b; b + t_{1-\alpha/2}[n-2] \cdot m_b \right] = \\ \left[b - t_{(1+\beta)/2}[n-2] \cdot m_b; b + t_{(1+\beta)/2}[n-2] \cdot m_b \right].$$

Стандартная ошибка параметра a определяется по формуле

$$m_a = \sqrt{\frac{RMS \cdot (S_x^2 + (\bar{x})^2)}{nS_x^2}} = \sqrt{\frac{\sum_{k=1}^n (y_k - \hat{y}_k)^2}{n-2} \cdot \frac{\sum_{k=1}^n x_k^2}{n \sum_{k=1}^n (x_k - \bar{x})^2}}.$$

Гипотеза $H_0: a=0$ о незначимости коэффициента a проверяется при помощи статистики $t_a = a/m_a$, имеющей распределение Стьюдента с $(n-2)$ степенями свободы.

Задается уровень значимости α . Вычисляется квантиль $t_{1-\alpha/2}[n-2]$ распределения Стьюдента уровня $(1-\alpha/2)$ с $(n-2)$ степенями свободы.

Если $|t_a| > t_{1-\alpha/2}[n-2]$, то гипотезу о несущественности параметра a отклоняем. Коэффициент a существенно отличается от нуля.

Если $|t_a| < t_{1-\alpha/2}[n-2]$, то гипотезу о несущественности параметра a не отклоняем. Коэффициент a не существенно отличается от нуля.

На основе оценки a метода наименьших квадратов можно построить доверительный интервал для свободного члена регрессии с доверительной вероятностью $\beta = 1 - \alpha$:

$$\begin{aligned} & \left[a - t_{1-\alpha/2}[n-2] \cdot m_a; a + t_{1-\alpha/2}[n-2] \cdot m_a \right] = \\ & \left[a - t_{(1+\beta)/2}[n-2] \cdot m_a; a + t_{(1+\beta)/2}[n-2] \cdot m_a \right]. \end{aligned}$$

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции:

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}}.$$

Далее, используется статистика $t_r = r/m_r$, имеющая распределение Стьюдента с $(n - 2)$ степенями свободы. При $|r|$ близких к 1 распределение величины t_r существенно отличается от распределения Стьюдента. Чтобы обойти это затруднение Рональд Фишер предложил ввести вспомогательную величину

$$z = \frac{1}{2} \ln \frac{1 + r}{1 - r},$$

со стандартной ошибкой $m_z = 1/\sqrt{n - 3}$, и использовать статистику $t_z = z/m_z$, имеющую распределение Стьюдента с $(n - 2)$ степенями свободы.

Уравнение регрессии дает точечный прогноз результата y , т. е.

$$\hat{y}_{n+1} = a + bx_{n+1},$$

где x_{n+1} — соответствующее значение фактора x .

Интервальную оценку для величины y с доверительной вероятностью β можно записать в виде:

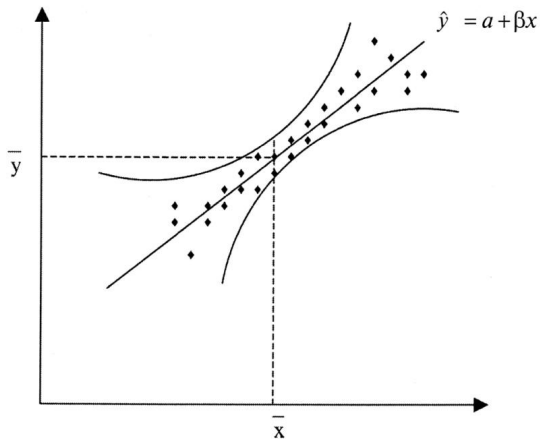
$$a + bx_{n+1} \pm t_{(1+\beta)/2}[n-2] \cdot m_{\hat{y}},$$

где

$$m_{\hat{y}} = \sqrt{RMS} \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{nS_x^2}\right)}.$$

Величину \sqrt{RMS} часто называют **стандартной ошибкой величины y** .

Как показано на рисунке границы доверительного интервала приблизительно представляют собой гиперболы. Самое “узкое” значение интервала находится в точке \bar{x} .



Линеаризация нелинейной регрессии

Различают два класса нелинейных регрессий.

- I Регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные относительно параметров.

Полиномиальная $y = a + bx + cx^2 + dx^3 + e$

Гиперболическая $y = a + \frac{b}{x} + e$

- II Регрессии нелинейные по оцениваемым параметрам.

Степенная $y = a \cdot x^b \cdot e$

Показательная $y = a \cdot b^x \cdot e$

Нелинейные регрессии первого класса после замены факторных переменных оценивают методом наименьших квадратов. Например, для гиперболической регрессии

$$\hat{y} = a + \frac{b}{x}$$

замена переменной x на $z = 1/x$ приводит к линейной модели

$$\hat{y} = a + bz.$$

Иначе обстоит дело с существенно нелинейными моделями второго класса. Для них необходимо проводить линеаризацию и заменять помимо факторных еще и результирующую переменную.

Например, для степенной регрессии

$$y = a \cdot x^b$$

можно провести логарифмирование

$$\lg y = \lg a + b \lg x.$$

Замены $w = \lg y$ и $z = \lg x$ приводят к линейной модели

$$w = A + bz,$$

где $A = \lg a$ ($a = 10^A$).

Для оценки тесноты связи между x и y используют **индекс корреляции** R_{yx} и **индекс детерминации** R_{yx}^2 :

$$R_{yx} = \sqrt{1 - \frac{RSS}{TSS}},$$

где

$$RSS = \sum_{k=1}^n (y_k - \hat{y}_k)^2, \quad TSS = \sum_{k=1}^n (y_k - \bar{y})^2.$$

После линеаризации нелинейных регрессий **первого класса** для оценки тесноты связи между новыми переменными может быть использован линейный коэффициент корреляции r_{yz} . Докажем, что

$$\begin{aligned} r_{yz}^2 &= R_{yx}^2. \\ r_{yz}^2 &= b^2 \frac{S_z^2}{S_y^2} = \frac{\sum_{k=1}^n (\hat{y}_k - \bar{y})^2}{\sum_{k=1}^n (y_k - \bar{y})^2} = \\ &= \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = R_{yx}^2. \end{aligned}$$

Иначе обстоит дело для существенно нелинейных регрессий второго класса. Для них коэффициент корреляции не совпадает с индексом корреляции.

СПАСИБО ЗА ВНИМАНИЕ!