4

Проверка статистической значимости коэффициентов

В какой мере можно доверять оценкам коэффициентов уравнения регрессии, вычисленным по данным конкретной выборке? Для определенности, рассуждения будем проводить для коэффициента β и его оценки b.

Могут быть выдвинуты две альтернативные гипотезы:

H_0: $\beta = \beta_0$ – нулевая гипотеза,

H_1: $\beta \neq \beta_0$ – альтернативная гипотеза.

Применим общую схему проверки гипотез. Если верна нулевая гипотеза H_0, то оценка $b \sim N(\beta_0, \sigma_b^2)$, и получаемая путем линейного преобразования случайная величина имеет стандартное нормальное распределение:

$$z = \frac{b - \beta_0}{\sigma_b} \sim N(0,1) \cdot$$

Однако, среднеквадратическое отклонение σ_b не известно. Поэтому воспользоваться стандартным нормальным распределением не удастся. Для парной линейной регрессии оценкой дисперсии случайного члена является величина

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2, \quad \text{где } e_i = y_i - \hat{y}_i, \quad i = 1,2,...,n. \qquad (4.6)$$

а для коэффициентов a,b оценки дисперсии вычисляются по формулам:

$$s_a^2 = \frac{s^2}{n}\left[1+\frac{\bar{x}^2}{Var(x)}\right], \quad s_b^2 = \frac{s^2}{nVar(x)}. \tag{4.7}$$

Квадратные корни s_a, s_b из оценок дисперсий s_a^2, s_b^2 называются *стандартными ошибками коэффициентов* a,b соответственно. Можно показать, что случайная величина

$$t = \frac{b-\beta_0}{s_b}$$

в предположении справедливости гипотезы H_0 имеет t-распределение Стьюдента с $n-2$ степенями свободы.

Зададимся уровнем значимости δ, равным 0.05 или 0.01, и по таблице определим критическое значение распределения Стьюдента с $n-2$ степенями свободы $t_c = t_{\delta/2,n-2}$. Для вычисленного значения t возможны две ситуации.

Первая ситуация. Величина t попадает в критическую об-ласть, т.е. $t < -t_c$ или $t > t_c$. Это означает, что произошло маловероятное событие, которое считается практически нереализуемым. Поэтому гипотеза H_0 отвергается в пользу альтернативной гипотезы H_1.

Вторая ситуация. Вычисленное значение t-статистики попадает в интервал от $-t_c$ до t_c. Это событие, вероятность которого равна $1-\delta$, т.е. 0.95 или 0.99. Нет оснований отвергать гипотезу H_0. Гипотеза H_0 не отвергается (принимается). Выражение «гипотеза H_0 не отвергается» неопределенно, однако, если какое-то решение надо принимать, и единственно разумное в этой ситуации сказать, что «гипотеза H_0 принимается». Отметим, что чем больше величина δ, тем *уже* интервал $\left[-z_c, z_c\right]$, и тем более

вероятна ошибка I рода – будет отвергнута истинная нулевая гипотеза H_0.

А чем меньше величина δ, тем шире интервал $[-z_c, z_c]$, и тем более вероятна ошибка II рода – не будет отвергнута ложная гипотеза H_0 (будет принята).

Наиболее важен случай, когда проверка гипотез проводится для $\beta_0 = 0$, т.е. тестируются гипотезы:

$$H_0: \ \beta = 0 \ - \text{нулевая гипотеза},$$
$$H_1: \ \beta \neq 0 \ - \text{альтернативная гипотеза}.$$

Фактически, это проверка наличия линейной связи между переменными. Даже если коэффициент β равен нулю, его оценка b, вообще говоря, не будет равной нулю, так как величина b вычисляется на основании формулы (4.5) по данным выборки.

В этом случае t-статистика представляет собой отношение оцененного коэффициента к его стандартной ошибке: $t = \dfrac{b}{s_b}$.

Аналогично проводится проверка гипотез для коэффициента a.

На практике экономисты обычно, как уже говорилось, полагают $\delta = 0.05$ или $\delta = 0.01$. Если гипотеза H_0 отвергается (не отвергается), то говорят, что коэффициент b значим (не значим) на 5%-ном или 1%-ном уровне соответственно.

П р и м е р 4.1. Рассмотрим линейную модель зависимости объема продаж Q от затрат на рекламу X:

$$Q = \alpha + \beta X + \varepsilon .$$

По формулам (4.5) для 9 наблюдений получены оценки коэффициентов: $a = 122$, $b = 32.7$. Найдены также их стандартные ошибки: $s_a = 57.2$, $s_b = 8.4$. Результаты обычно записывают в следующей форме:

$$\hat{Q} = \quad 122 \quad + \quad 32.7\,X$$

(57.2) (8.4)

Стандартные ошибки записывают в круглых скобках под соответствующими коэффициентами. Вычислим t-статистики:

$$t_a = \frac{122}{57.2} = 2.13, \quad t_b = \frac{32.7}{8.4} = 3.89.$$

По таблице распределения Стьюдента найдем критические значения t-статистики с 9-2=7 степенями свободы для $\delta = 0.05$ и $\delta = 0.01$:

$$t_{0.025,7} = 2.365, \quad t_{0.005,7} = 3.499.$$

Сравнивая с ними вычисленные значения t-статистики, получим, что коэффициент $a = 122$ не значим (гипотеза H_0 не отвергается) при 5%-ном уровне, а коэффициент $b = 32.7$ значим (гипотеза H_0 отвергается) при 1%-ном уровне.

ЗАДАЧИ

1. В следующей таблице приведены выборочные данные по двум показателям. Оцените по методу наименьших квадратов коэффициенты парной линейной регрессии с зависимой переменной *y* и объясняющей переменной *x*.

x	6	5	4	3	4	2
y	1	4	3	6	7	9

2. По данным предыдущей задачи вычислите стандартные ошибки коэффициентов уравнения регрессии и проверьте гипотезу $H_0 : \beta = 0$.

4. Стремясь оптимизировать свою прибыль, фирма провела обследование и собрала данные о количестве занятых работников (*L*) и объеме производства (*Q*), приведенные в следующей таблице:

L	4	5	6	7	8	9
Q	12	14	19	20	25	29

Постройте линейную зависимость объема производства от численности занятых работников. Проверьте статистическую значимость коэффициентов.

5. По данным предыдущей задачи сделайте прогноз ожидаемого значения объема производства, постройте 95%-ный доверительный интервал для ожидаемого объема производства и его индивидуального значения, если фирма наймет 10 работников.

6. Получена выборка 20 наблюдений значений двух переменных x, y и найдено, что

$$\sum_{i=1}^{20} x_i = 50, \quad \sum_{i=1}^{20} y_i = 120, \quad \sum_{i=1}^{20} x_i^2 = 250, \quad \sum_{i=1}^{20} x_i y_i = 450.$$

Оцените коэффициенты парной линейной регрессии y на x.

7. Пусть в дополнение к условиям задачи 6 известно, что

$$\sum_{i=1}^{20} y_i^2 = 250.$$

Найдите несмещенную оценку дисперсии случайного члена рег-рессии, стандартные ошибки оцененных коэффициентов регрессии и выполните проверку их значимости на основе t-распределения Стьюдента.

8. Продавец мороженого для определения наиболее выгодной для него цены в течение 10 дней варьировал цену и фиксировал количества проданных порций. Результаты представлены в следующей таблице.

i	1	2	3	4	5	6	7	8	9	10
p	12	13	10	15	17	16	11	18	20	19
Q	120	110	140	105	100	104	130	90	80	85

Постройте линейную регрессию количества Q проданных порций мороженого на цену p и определите оптимальную цену, при которой выручка максимальная.

9. В таблице приведены значения денежной массы М2 и валового внутреннего продукта Казахстана в млрд тенге. Постройте регрессию денежного агрегата М2 на ВВП и оцените статистическую значимость оцененных коэффициентов. Приведите их экономическую интерпретацию.

Год	**Денежная масса М2**	**ВВП**
2005	1516	7591
2006	2815	10214

2007	3554	12850
2008	4620	16053
2009	5335	17008
2010	6570	21816
2011	7968	29380
2012	8547	32194
2013	8678	37085
2014	7968	40755

10. Для выборки с 20 наблюдениями значений пары переменных *x*, *y* известны следующие величины:

$$\sum_{i=1}^{20} x_i = 120,\ \sum_{i=1}^{20} y_i = 200, \sum_{i=1}^{20} x_i^2 = 800, \sum_{i=1}^{20} y_i^2 = 2500,\ r_{xy} = 0.8.$$

Оцените коэффициенты парной линейной регрессии *y* на *x*.

11. По 25 наблюдениям оценена регрессия y^ = 2.4 + 0.8x. Известно, что среднее выборочное объясняющей переменной равно 6.6, стандартная ошибка коэффициента при ней равна 0,3. Найдите оптимальный прогноз и доверительный интервал для ожидаемого значения зависимой переменной для значения $x_0 = 7.2$, если оценка дисперсии случайного члена регрессии $s^2 = 0.25$, критическое значение статистики Стьюдента равно 2,069.