

Эконометрика

Полковников Александр Александрович

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

Конспект лекций для студентов направления
"Экономика"

Частная корреляция

Ранжирование факторов, участвующих в множественной линейной регрессии, может быть проведено с помощью стандартизованных коэффициентов регрессии (β). Эта же цель может быть достигнута с помощью частных коэффициентов корреляции, характеризующих тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в уравнение регрессии.

Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет включения в анализ нового фактора к остаточной дисперсии, имевшей место до введение нового фактора в модель.

Предположим, что в модели имеются результат y и два нетривиальных фактора x_1, x_2 . Тогда величина

$$r_{yx_1|x_2} = \sqrt{\frac{RSS_{yx_2} - RSS_{yx_1x_2}}{RSS_{yx_2}}}$$

называется **частным коэффициентом корреляции по x_1 первого порядка**.

Воспользовавшись формулой

$$RSS = TSS (1 - R^2)$$

можно записать

$$r_{yx_1|x_2} = \sqrt{1 - \frac{1 - R_{yx_1x_2}^2}{1 - R_{yx_2}^2}}.$$

Используют и частные коэффициенты корреляции более высоких порядков:

$$r_{yx_1|x_2\dots x_m} = \sqrt{1 - \frac{1 - R_{yx_1\dots x_m}^2}{1 - R_{yx_2\dots x_m}^2}},$$

которые можно определить через частные коэффициенты более низких степеней по рекуррентной формуле

$$r_{yx_1|x_2\dots x_m} = \frac{r_{yx_1|x_2\dots x_{m-1}} - r_{yx_m|x_2\dots x_{m-1}} \cdot r_{x_1x_m|x_2\dots x_{m-1}}}{\sqrt{\left(1 - r_{yx_m|x_2\dots x_{m-1}}^2\right) \cdot \left(1 - r_{x_1x_m|x_2\dots x_{m-1}}^2\right)}}.$$

В частности, влияние на результат y фактора x_1 при постоянном воздействии фактора x_2 можно измерить с помощью

$$\begin{aligned} r_{yx_1|x_2} &= \sqrt{\frac{R_{yx_1x_2}^2 - r_{yx_2}^2}{1 - r_{yx_2}^2}} = \\ &= \sqrt{\frac{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2} - r_{yx_2}^2}{1 - r_{yx_2}^2}} = \\ &= \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2) \cdot (1 - r_{x_1x_2}^2)}}. \end{aligned}$$

А влияние на результат y фактора x_1 при постоянном воздействии факторов x_2 и x_3 можно измерить с помощью

$$r_{yx_1|x_2x_3} = \frac{r_{yx_1|x_2} - r_{yx_3|x_2} \cdot r_{x_1x_3|x_2}}{\sqrt{(1 - r_{yx_3|x_2}^2) \cdot (1 - r_{x_1x_3|x_2}^2)}}.$$

В частном случае, если стандартизированное уравнение регрессии имеет вид

$$\hat{y}^* = \beta_1 x_1^* + \beta_2 x_2^*$$

коэффициенты β могут быть определены по формулам

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2},$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}.$$

Сравнивая эти формулы с частными коэффициентами корреляции, имеем

$$r_{yx_1|x_2} = \beta_1 \sqrt{\frac{1 - r_{x_1 x_2}^2}{1 - r_{yx_2}^2}},$$

$$r_{yx_2|x_1} = \beta_2 \sqrt{\frac{1 - r_{x_1 x_2}^2}{1 - r_{yx_1}^2}}.$$

Частные коэффициенты подтверждают ранжирование факторов по стандартизированным коэффициентам регрессии. Зная частные коэффициенты корреляции, можно определить индекс детерминации

$$R_{yx_1 \dots x_m}^2 = 1 - (1 - r_{yx_1}^2) \cdot (1 - r_{yx_2|x_1}^2) \cdot \dots \cdot (1 - r_{yx_m|x_1 \dots x_{m-1}}^2).$$

Оценка надежности множественной регрессии

Для оценки значимости параметров уравнения множественной линейной регрессии используются критерий Стьюдента.

Обозначим **стандартную ошибку** параметра θ_j через m_j

$$m_j = \sqrt{\frac{RSS}{n - m - 1} (X^T X)^{-1}_{jj}}$$

При выполнении гипотезы $H_0: \theta_j = 0$ случайная величина

$$t_j = \frac{\hat{\theta}_j}{m_j}$$

имеет распределение Стьюдента с $(n - m - 1)$ степенями свободы.

Задаем уровень значимости α . Затем, сравним фактическое значение статистики t_j с квантилью $t_{1-\alpha/2}[n - m - 1]$ распределения Стьюдента.

Отвергаем гипотезу H_0 (считаем параметр θ_j значимым) при $|t_j| > t_{1-\alpha/2}[n - m - 1]$.

В противном случае при $|t_j| < t_{1-\alpha/2}[n - m - 1]$ считаем параметр θ_j не значимым и влияние фактора x_j на результат y не существенным.

Значимость дополнительных факторов, включаемых в уравнение регрессии, можно оценить с помощью **частного F-критерия**:

$$F_j = \frac{RSS_1 - RSS_2}{RSS_2 / (n - m - 1)},$$

где

RSS_1 — остаточная сумма квадратов для модели без фактора x_j ,

RSS_2 — остаточная сумма квадратов для модели с фактором x_j .

Разделив числитель и знаменатель на полную сумму квадратов TSS , получим

$$F_j = \frac{R^2_{yx_1 \dots x_j \dots x_m} - R^2_{yx_1 \dots x_{j-1} x_{j+1} \dots x_m}}{R^2_{yx_1 \dots x_j \dots x_m} / (n - m - 1)}.$$

Можно доказать, что $F_j = t_j^2$.

Величина F_j имеет распределение Фишера числом степеней свободы 1 и $(n - m - 1)$. Для уровня значимости α вычисляем квантиль $F_{1-\alpha}[1; n - m - 1]$ распределения Фишера уровня $1 - \alpha$.

Если $F_j > F_{1-\alpha}[1; n - m - 1]$, то включение фактора x_j в модель оправдано.

Если $F_j < F_{1-\alpha}[1; n - m - 1]$, то включение фактора x_j в модель не оправдано.

Помимо оценки значимости отдельных факторов F -критерий Фишера используется для проверки значимости уравнения регрессии в целом. Случайная величина

$$F = \frac{EMS}{RMS} = \frac{ESS/m}{RSS/(n-m-1)} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$$

имеет распределение Фишера с числом степеней свободы m и $(n-m-1)$.

Уравнение регрессии значимо, если фактическое значение F -критерия больше табличного $F_{1-\alpha}[m; n-m-1]$, где α — уровень значимости.

Прогнозирование в множественной регрессии

Прогнозирование по модели множественной регрессии проводится аналогично прогнозированию по уравнению парной регрессии.

Пусть имеем линейную модель $y = X\theta + e$ с $(m + 1)$ фактором, $\mathbb{E}e = 0$, $\mathbb{D}e = \sigma^2 I_n$. Известны значения результата и факторов в n наблюдениях. Требуется дать прогноз результата y в $(n + 1)$ -м наблюдении, если значения факторов в $(n + 1)$ -м наблюдении известны.

Точечный прогноз результата находится по уравнению регрессии:

$$\hat{y}_{n+1} = X_{n+1} \hat{\theta} = X_{n+1} (X^T X)^{-1} X^T y,$$

где

\hat{y}_{n+1} — прогнозное значение результата y ,

$X_{n+1} = (1; x_{n+1,1}; \dots; x_{n+1,m})$ — вектор-строка значений факторов x_1, \dots, x_m , для которых строим прогноз,

$\hat{\theta}$ — оценка параметров регрессии методом наименьших квадратов.

Математическое ожидание и дисперсия величины \hat{y}_{n+1} составляют:

$$\mathbb{E}(\hat{y}_{n+1}) = X_{n+1}\theta.$$

$$\begin{aligned}\mathbb{D}(\hat{y}_{n+1}) &= \text{Cov}(X_{n+1}\hat{\theta}) = X_{n+1} \text{Cov}(\hat{\theta}) X_{n+1}^T = \\ &= \sigma^2 X_{n+1} (X^T X)^{-1} X_{n+1}^T.\end{aligned}$$

Вычислим характеристики **ошибки прогноза** $y_{n+1} - \hat{y}_{n+1}$:

$$\mathbb{E}(y_{n+1} - \hat{y}_{n+1}) = X_{n+1}\theta - X_{n+1}\theta = 0,$$

$$\begin{aligned}\mathbb{D}(y_{n+1} - \hat{y}_{n+1}) &= \sigma^2 + \sigma^2 X_{n+1} (X^T X)^{-1} X_{n+1}^T \\ &= \sigma^2 (1 + \nu),\end{aligned}$$

где $\nu = X_{n+1} (X^T X)^{-1} X_{n+1}^T$.

Величина

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{\sigma^2 (1 + \nu)}}$$

имеет стандартное нормальное распределение, а величина

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{\frac{RSS}{n-m-1} (1 + \nu)}}$$

имеет распределение Стьюдента с $n - m - 1$ степенями свободы.

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

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

где

$$m_{\hat{y}} = \sqrt{RMS} \cdot \sqrt{(1 + \nu)}$$

$$RMS = RSS / (n - m - 1),$$

$$\nu = X_{n+1} (X^T X)^{-1} X_{n+1}^T.$$

В частности, для парной линейной регрессии ($m = 1$) получаем:

$$\nu = \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{nS_x^2}.$$

СПАСИБО ЗА ВНИМАНИЕ!