## Свойства коэффициентов регрессии. Теорема Гаусса-Маркова

Предположим, что истинная модель регрессии между, например, расходами на питание ($y$) и располагаемым личным доходом ($x$) описывается следующим выражением:

$$y = \alpha + \beta x + u \tag{3.21}$$

и оценка регрессии

$$\hat{y} = 55{,}3 + 0{,}093x \tag{3.22}$$

Полученный результат можно истолковать следующим образом: коэффициент при $x$ (коэффициент наклона) показывает, что если $x$ увеличивается на одну единицу, то $y$ возрастает на 0,093 единицы. Как $x$, так и $y$ измеряются в млрд. долл. в постоянных ценах; таким образом, коэффициент наклона показывает, что если доход увеличивается на 1 млрд. долл., то расходы на питание возрастают на 93 млн. долл. Другими словами, из каждого дополнительного доллара дохода 9,3 цента будут израсходованы на питание.

Что можно сказать о постоянной в уравнении? Формально говоря, она показывает прогнозируемый уровень $y$, когда $x = 0$. Иногда это имеет смысл, иногда нет.

В рассматриваемом случае получается, что если доход был бы равен 0, то расходы на питание составили бы 55,3 млрд. долл. Такое толкование может быть правдоподобным в отношении отдельного человека, т.к. он может израсходовать на питание накопленные или одолженные средства. Однако вряд ли оно имеет какой то смысл применительно к совокупности.

### *Качество оценки: коэффициент $R^2$*

В парном регрессионном анализе мы пытаемся объяснить поведение $y$ путем определения регрессионной зависимости от соответственно выбранной независимой переменной $x$. После построения уравнения регрессии мы можем разбить значение $y_i$ в каждом наблюдении на две составляющих – $\hat{y}_i$ и $e_i$:

$$y_i = \hat{y}_i + e_i \tag{3.23}$$

Величина $\hat{y}_i$ – расчетное значение в наблюдении $i$ – это то значение, которое имел бы $y$ при условии, что уравнение регрессии было правильным, и отсутствии случайного фактора. Это, иными словами, величина $y$, спрогнозированная по значению $x$ в данном наблюдении. Тогда остаток $e_i$ есть расхождение между фактическим и спрогнозированным значениями величины $y$. Это та часть $y$, которую мы не можем объяснить с помощью уравнения регрессии.

Разброс значений $y$ в любой выборке можно суммарно описать с помощью выборочной дисперсии $Var(y)$. Мы должны уметь рассчитать величину этой дисперсии.

Используя (3.23), разложим дисперсию $y$.

$$Var(y) = Var(\hat{y} + e) = Var(\hat{y}) + Var(e) + 2Cov(\hat{y}, e). \quad (3.24)$$

Далее, оказывается, что $Cov(\hat{y}, e)$ должна быть равна нулю (попробуйте доказать это самостоятельно). Следовательно, мы получаем:

$$Var(y) = Var(\hat{y}) + Var(e). \quad (3.25)$$

Это означает, что мы можем разложить $Var(y)$ на две части: $Var(\hat{y})$ – часть, которая "объясняется" уравнением регрессии в вышеописанном смысле, и $Var(e)$ - "необъясненную" часть.

Согласно (3.25), $Var(\hat{y}) / Var(y)$ – это часть дисперсии $y$, объясненная уравнением регрессии. Это отношение известно, как коэффициент детерминации и его обычно обозначают $R^2$:

$$R^2 = \frac{Var(\hat{y})}{Var(y)}, \quad (3.26)$$

что равносильно

$$R^2 = 1 - \frac{Var(e)}{Var(y)}. \quad (3.27)$$

Максимальное значение коэффициента $R^2$ равно единице. Это происходит в том случае, если линия регрессии точно соответствует всем наблюдениям, так что $\hat{y}_i = y_i$ для всех $i$ и все остатки равны нулю. Тогда $Var(\hat{y}) = Var(y)$, $Var(e) = 0$ и $R^2 = 1$.

Если в выборке отсутствует видимая связь между $y$ и $x$, то коэффициент $R^2$ будет близок к нулю.

При прочих равных условиях желательно, чтобы коэффициент $R^2$ был как можно больше. Легко показать, что это не противоречит критерию, в соответствии с которым $a$ и $b$ должны быть выбраны таким образом, чтобы минимизировать сумму квадратов остатков. Отметим сначала, что

$$e_i = y_i - \hat{y}_i - a - bx_i, \quad (3.28)$$

откуда, беря среднее значение $e_i$ по выборке и используя уравнение (3.11), получим:

$$\bar{e} = \bar{y} - a - b\bar{x} = \bar{y} - \left[\bar{y} - b\bar{x}\right] - b\bar{x} = 0. \quad (3.29)$$

Следовательно,

$$Var(e) = \frac{1}{n}\sum (e_i - \bar{e})^2 = \frac{1}{n}\sum e_i^2 \quad . \qquad (3.30)$$

Отсюда следует, что принцип минимизации суммы квадратов остатков эквивалентен минимизации дисперсии остатков при условии выполнения (3.11). Однако если мы минимизируем $Var(e)$, то при этом в соответствии с (3.27) автоматически максимизируется коэффициент $R^2$.

*Альтернативное представление коэффициента* $R^2$

На интуитивном уровне представляется очевидным, что чем больше соответствие, обеспечиваемое уравнением регрессии, тем больше должен быть коэффициент корреляции для фактических и прогнозных значений $y$ и наоборот. Покажем, что $R^2$ фактически равен квадрату такого коэффициента корреляции между $y$ и $\hat{y}$, который мы обозначим $r_{y,\hat{y}}$ (заметим, что $Cov(e, \hat{y}) = 0$):

$$r_{y,\hat{y}} = \frac{Cov(y, \hat{y})}{\sqrt{Var(y)Var(\hat{y})}} = \frac{Cov(\{\hat{y} + e\}, \hat{y})}{\sqrt{Var(y)Var(\hat{y})}} = \frac{Cov(\hat{y}, \hat{y}) + Cov(e, \hat{y})}{\sqrt{Var(y)Var(\hat{y})}} =$$

$$= \frac{Var(\hat{y})}{\sqrt{Var(y)Var(\hat{y})}} = \frac{\sqrt{Var(\hat{y})}}{\sqrt{Var(y)}} = \sqrt{R^2} \quad . \qquad (3.31)$$

*Предположения о случайной составляющей*

Для того, чтобы регрессионный анализ, основанный на обычном методе наименьших квадратов, давал наилучшие из всех возможные результаты, случайная составляющая должна удовлетворять четырем условиям, известным как условия Гаусса – Маркова.

*1-е условие Гаусса – Маркова*: $E(u_i) = 0$ *для всех наблюдений*

Итак, математическое ожидание случайной составляющей в любом наблюдении должно быть равно нулю.

Фактически, если уравнение регрессии включает постоянный член, то обычно бывает разумно предположить, что это условие выполняется автоматически, т.к. роль константы состоит в определении любой систематической тенденции в $y$, которую не учитывают объясняющие переменные, включенные в уравнение регрессии.

*2-е условие Гаусса – Маркова*: $pop.\mathrm{var}(u_i)$ *постоянна для всех наблюдений*

Эта постоянная дисперсия обычно обозначается $\sigma_u^2$, или часто в более краткой форме $\sigma^2$, а условие записывается следующим образом:

$$pop.\mathrm{var}(u_i) = \sigma_u^2 \text{ для всех } i \, . \qquad (4.7)$$

Т.к. $E(u_i)=0$ и $pop.\mathrm{var}(u_i)=E(u_i)^2$ условие можно переписать в виде

$$E(u_i^2)=\sigma_u^2 \text{ для всех } i. \qquad (4.8)$$

Величина $\sigma_u$, конечно, неизвестна. Одна из задач регрессионного анализа состоит в оценке стандартного отклонения случайной составляющей.

Если рассматриваемое условие не выполняется, то коэффициенты регрессии, найденные по обычному методу наименьших квадратов, будут неэффективны, и можно получить более надежные результаты путем применения модифицированного метода регрессии.

*3-е условие Гаусса – Маркова*: $pop.\mathrm{cov}(u_i,u_j)=0 \quad (i\neq j)$

Это условие предполагает отсутствие систематической связи между значениями случайной составляющей в любых двух наблюдениях. Случайные составляющие должны быть абсолютно независимы друг от друга.

В силу того, что $E(u_i)=E(u_j)=0$, данное условие можно записать следующим образом:

$$E(u_iu_j)=0 \quad (i\neq j). \qquad (4.9)$$

Если это условие не будет выполнено, то регрессия, оцененная по обычному методу наименьших квадратов, вновь даст неэффективные результаты. В главе 8 рассматриваются возникшие здесь проблемы и пути их преодоления.

*4-е условие Гаусса – Маркова: случайная составляющая должна быть распределена независимо от объясняющих переменных*

Значение любой независимой переменной в каждом наблюдении должно считаться экзогенным, полностью определяемым внешними причинами, учитываемыми в уравнении регрессии.

Если это условие выполнено, то теоретическая ковариация между независимой переменной и случайной составляющей равна нулю. Т.к. $E(u_i)=0$, то

$$pop.\mathrm{cov}(x_i,u_i)=E\{(x_i-\bar{x})(u_i)\}=E(x_iu_i)-\bar{x}E(u_i)=E(x_iu_i). \qquad (4.10)$$

Следовательно, данное условие можно записать также в виде

$$E(x_iy_i)=0. \qquad (4.11)$$