Исследование линейной зависимости результирующей переменной от нескольких объясняющих переменных

Парные коэффициенты корреляции $r\left(y, x^{(i)}\right)$ не учитывают влияние на эту связь других переменных $x^{(j)}, j \neq i$ $\Rightarrow$ **необходим измеритель связи, очищенный от опосредованного влияния других переменных**, т. е. дающий оценку тесноты связи между y и $x^{(j)}$ при условии, что значения остальных переменных зафиксированы на некотором постоянном уровне.

Частные (очищенные) коэффициенты корреляции

Приведенные формулы справедливы для многомерного нормального закона и приближенно для линейных множественных связей $y = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)} + \varepsilon(X)$. Обозначим для удобства $y \equiv x^{(0)}$.

$r_{ij(-ij)} = \dfrac{-R_{ij}}{\left(R_{ii} R_{jj}\right)^{1/2}}$ – частный коэффициент корреляции между переменными $x^{(i)}$ и $x^{(j)}$ при фиксированнных значениях всех остальных переменных.

R_{kl} – алгебраическое дополнение для r_{kl} в определителе корреляционной матрицы

$$R = \begin{pmatrix} 1 & r_{01} & r_{02} & \ldots & r_{0p} \\ r_{10} & 1 & r_{12} & \ldots & r_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{p0} & r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}.$$

$R_{kl} = (-1)^{k+l} \det A_{kl}$, матрица A_{kl} получена из R вычеркиванием k-строки и l-столбца.

\#\# $r_{01(2)} = \dfrac{r_{01} - r_{02} r_{12}}{\sqrt{\left(1 - r_{02}^2\right)\left(1 - r_{12}^2\right)}}$ – формула, примененная к трехмерному признаку.

Свойства (проверка гипотез, доверительные интервалы) **частных коэффициентов корреляции** k-порядка (исключение влияния k переменных) такие же, как у парных коэффициентов корреляции с поправкой: **объем выборки уменьшается на k**.

\#\# $n = 37$ – число исследованных предприятий легкой промышленности,
$x^{(0)} \equiv y$ – качество ткани (в баллах),
$x^{(1)}$ – среднемесячное число профилактических наладок автоматической линии,
$x^{(2)}$ – среднемесячное число обрывов нити.
$\hat{r}_{01} = 0{,}105,\ \hat{r}_{02} = 0{,}024,\ \hat{r}_{12} = 0{,}996$.

$$R = \begin{pmatrix} 1 & 0{,}105 & 0{,}024 \\ 0{,}105 & 1 & 0{,}996 \\ 0{,}024 & 0{,}996 & 1 \end{pmatrix}.$$

$$r_{01(2)} = \frac{0{,}105 - 0{,}024 \times 0{,}996}{\sqrt{\left(1 - 0{,}996^2\right)\left(1 - 0{,}024^2\right)}} = 0{,}9079,$$

$$r_{02(1)} = \frac{-\left(0{,}105 \times 0{,}996 - 0{,}024\right)}{\sqrt{\left(1 - 0{,}996^2\right)\left(1 - 0{,}105^2\right)}} = -0{,}9068.$$

Связь имеется, что согласуется с профессиональными представлениями.
Найдем доверительный интервал для истинного значения $r_{01(2)}$.
Исключаем одну переменную $\Rightarrow$ $n = 37 - 1 = 36$.

$$z = \text{ФИШЕР}(0{,}9079) - \frac{0{,}9079}{2(36-1)} = 1{,}5022.$$

$$z \in \left[1{,}5022 - \frac{1{,}96}{\sqrt{36-3}};\ 1{,}5022 + \frac{1{,}96}{\sqrt{36-3}}\right],\quad z \in [1{,}1610;\ 1{,}8434],\quad r \in [0{,}8214;\ 0{,}9511].$$

$n = 20$ – число лет наблюдений за погодой,
$x^{(0)} \equiv y$ – урожайность кормовых трав,
$x^{(1)}$ – весеннее количество осадков,
$x^{(2)}$ – накопленная за весну сумма активных (выше +5,5^{0}C) температур.
$\hat{r}_{01} = 0{,}80,\ \hat{r}_{02} = -0{,}40,\ \hat{r}_{12} = -0{,}56$.

$$R = \begin{pmatrix} 1 & 0{,}80 & -0{,}40 \\ 0{,}80 & 1 & -0{,}56 \\ -0{,}40 & -0{,}56 & 1 \end{pmatrix}.$$

$$r_{01(2)} = \frac{0{,}80 - 0{,}56 \times 0{,}40}{\sqrt{\left(1 - 0{,}56^2\right)\left(1 - 0{,}40^2\right)}} = 0{,}759,$$

$$r_{02(1)} = \frac{-\left(0{,}40 - 0{,}80 \times 0{,}56\right)}{\sqrt{\left(1 - 0{,}56^2\right)\left(1 - 0{,}80^2\right)}} = 0{,}097.$$

$r_{01(2)} \in [0{,}448;\ 0{,}898],\ \ r_{02(1)} \in [-0{,}376;\ 0{,}526]$.

Множественный коэффициент корреляции

Множественный коэффициент корреляции – это коэффициент корреляции между y и линейной функцией регрессии y по x, т. е. между y и линейной комбинацией $x^{(1)}, \ldots, x^{(p)}$, для которой значение коэффициента корреляции максимально.

$$R_{y.X} = r\left(y, f(X)\right) = r\left(y,\ \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}\right).$$

Свойства множественного коэффициента корреляции (МКК):

1. $K_d(y;X) = R^2_{y.X} = 1 - \dfrac{D\varepsilon}{Dy}$.
2. Вычисление МКК по корреляционной матрице: $R^2_{y.X} = 1 - \dfrac{|R|}{|R_{00}|}$.
3. Вычисление МКК по частным КК: $R^2_{y.X} = 1 - \left(1 - r^2_{01}\right)\left(1 - r^2_{02(1)}\right)\left(1 - r^2_{03(12)}\right)\ldots\left(1 - r^2_{0p(123\ldots(p-1))}\right)$.

$\hat{R}^2_{y.\left(x^{(1)}x^{(2)}\right)} = 1 - \left(1 - \hat{r}^2_{01}\right)\left(1 - \hat{r}^2_{02(1)}\right) = 1 - \left(1 - 0{,}105^2\right)\left(1 - (-0{,}9068)^2\right) = 0{,}8243,$

$$\hat{R}^2_{y.\left(x^{(1)}x^{(2)}\right)} = 1 - \left(1 - \hat{r}^2_{02}\right)\left(1 - \hat{r}^2_{01(2)}\right) = 1 - \left(1 - 0{,}024^2\right)\left(1 - 0{,}9079^2\right) = 0{,}8243,$$

$$\hat{R}_{y.\left(x^{(1)}x^{(2)}\right)} = \sqrt{0{,}8243} = 0{,}9079.$$

4. МКК мажорирует все парные и частные КК, характеризующие статистическую связь y: $R^2_{y.X} \geq r^2_{0j\left(I_j\right)}$, где I_j – любое подмножество $\{1\ldots p\}$, не содержащее j.
5. Присоединение новой переменной не может уменьшить величины R (вне зависимости от порядка присоединения): $R^2_{y.x^{(1)}} \leq R^2_{y.\left(x^{(1)},x^{(2)}\right)} \leq R^2_{y.\left(x^{(1)},x^{(2)},x^{(3)}\right)} \leq \ldots \leq R^2_{y.\left(x^{(1)},x^{(2)},\ldots x^{(p)}\right)}$.

Проверка гипотезы об отсутствии множественной линейной связи

Гипотеза о статистической независимости y и $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ $\quad H_0 : R_{y.X} = 0$.

1. Выбираем уровень значимости α.

2. Сравниваем эмпирическое и критическое значение критерия:

$$F_{\text{эмп}} = \frac{\hat{R}^2_{yx}}{1 - \hat{R}^2_{yx}} \frac{n - p - 1}{p}, \quad F_{\text{крит}} = \text{FРАСПОБР}(\alpha;\ p;\ n - p - 1).$$

Если $F_{\text{эмп}} > F_{\text{крит}}$, то гипотеза об отсутствии множественной линейной связи отвергается с вероятностью ошибки $\alpha \Rightarrow$ связь есть.