Эконометрика

Кракашова Ольга Анатольевна

канд. экон. наук,
доцент кафедры СЭиОР РГЭУ (РИНХ)

Лекция № 3

Множественная регрессия и корреляция

Уравнение множественной регрессии

$$y = f(x_1, x_2, ..., x_m) + \varepsilon,$$

где y – зависимая переменная (результативный признак), x_i – независимые, или объясняющие, переменные (признаки-факторы).

Основная цель множественной регрессии– построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

Спецификация модели включает 2 этапа:

1) отбор факторов;

2) выбор вида уравнения регрессии.

Отбор факторов при построении уравнения множественной регрессии

Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:

1. Они должны быть количественно измеримы. Если необходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность;
2. Факторы не должны быть интеркоррелированы и тем более находиться в точной функциональной связи.

Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и не увеличивает показатель детерминации, но и приводит к статистической незначимости параметров регрессии по критерию Стьюдента.

Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. Если строится модель с набором m факторов, то для нее рассчитывается показатель детерминации R^2, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии m факторов. Влияние других, не учтенных в модели факторов, оценивается как $1-R^2$ с соответствующей остаточной дисперсией S^2.

При дополнительном включении в регрессию $m+1$ фактора коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться:

$$R^2_{m+1} \geq R^2_m \quad \text{и} \quad S^2_{m+1} \leq S^2_m.$$

Если же этого не происходит и данные показатели практически не отличаются друг от друга, то включаемый в анализ фактор x_{m+1} не улучшает модель и практически является лишним фактором.

Отбор факторов обычно осуществляется в две стадии:

1) подбираются факторы исходя из сущности проблемы;
2) на основе матрицы показателей корреляции определяют статистики для параметров регрессии.

Коэффициенты интеркорреляции (т.е. корреляции между объясняющими переменными) позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно коллинеарны, т.е. находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0,7$. Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии. Предпочтение при этом отдается не фактору, более тесно связанному с результатом, а тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами. В этом требовании проявляется специфика множественной регрессии как метода исследования комплексного воздействия факторов в условиях их независимости друг от друга.

Пусть, например, при изучении зависимости $y = \hat{f}(x_1, x_2, x_3)$ матрица парных коэффициентов корреляции оказалась следующей:

Таблица 2.1

	y	x_1	x_2	x_3
y	1	0,8	0,7	0,6
x_1	0,8	1	0,8	0,5
x_2	0,7	0,8	1	0,2
x_3	0,6	0,5	0,2	1

Очевидно, что факторы x_1 и x_2 дублируют друг друга. В анализ целесообразно включить фактор x_2, а не x_1, хотя корреляция x_2 с результатом y слабее, чем корреляция фактора x_1 с y $\left(r_{yx_2} = 0,7 < r_{yx_1} = 0,8\right)$, но зато значительно слабее межфакторная корреляция $r_{x_2x_3} = 0,2 < r_{x_1x_3} = 0,5$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

Включение в модель мультиколлинеарных факторов нежелательно
в силу следующих последствий:

1. Затрудняется интерпретация параметров множественной регрессии как характеристик действия факторов в «чистом» виде, ибо факторы коррелированы; параметры линейной регрессии теряют экономический смысл.
2. Оценки параметров ненадежны, обнаруживают большие стандартные ошибки и меняются с изменением объема наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования.

Для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами.

Если бы факторы не коррелировали между собой, то матрица парных коэффициентов корреляции между факторами была бы единичной матрицей, поскольку все недиагональные элементы $r_{x_i x_j}$ $(i \neq j)$ были бы равны нулю. Так, для уравнения, включающего три объясняющих переменных

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

матрица коэффициентов корреляции между факторами имела бы определитель, равный единице:

$$\mathrm{Det}\,\mathbf{R} = \begin{vmatrix} r_{x_1 x_1} & r_{x_1 x_2} & r_{x_1 x_3} \\ r_{x_2 x_1} & r_{x_2 x_2} & r_{x_2 x_3} \\ r_{x_3 x_1} & r_{x_3 x_2} & r_{x_3 x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

Если же, наоборот, между факторами существует полная линейная зависимость и все коэффициенты корреляции равны единице, то определитель такой матрицы равен нулю:

$$\mathrm{Det}\,\mathbf{R} = \begin{vmatrix} r_{x_1 x_1} & r_{x_1 x_2} & r_{x_1 x_3} \\ r_{x_2 x_1} & r_{x_2 x_2} & r_{x_2 x_3} \\ r_{x_3 x_1} & r_{x_3 x_2} & r_{x_3 x_3} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0.$$

Существует ряд подходов преодоления сильной межфакторной корреляции:

1) самый простой путь устранения мультиколлинеарности состоит в исключении из модели одного или нескольких факторов;
2) связан с преобразованием факторов, при котором уменьшается корреляция между ними.

Одним из путей учета внутренней корреляции факторов является переход к совмещенным уравнениям регрессии, т.е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие. Так, если $y = f(x_1, x_2, x_3)$, то возможно построение следующего совмещенного уравнения:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + \varepsilon.$$

Рассматриваемое уравнение включает взаимодействие первого порядка (взаимодействие двух факторов). Возможно включение в модель и взаимодействий более высокого порядка, если будет доказана их статистическая значимость по F-критерию Фишера, но, как правило, взаимодействия третьего и более высоких порядков оказываются статистически незначимыми.

Наиболее широкое применение получили следующие методы построения уравнения множественной регрессии:

1. Метод исключения– отсев факторов из полного его набора.
2. Метод включения– дополнительное введение фактора.
3. Шаговый регрессионный анализ – исключение ранее введенного фактора.

При отборе факторов также рекомендуется пользоваться следующим правилом: **число включаемых факторов обычно в 6–7 раз меньше объема совокупности, по которой строится регрессия**. Если это соотношение нарушено, то число степеней свободы остаточной дисперсии очень мало. Это приводит к тому, что параметры уравнения регрессии оказываются статистически незначимыми, а F –критерий меньше табличного значения.

Метод наименьших квадратов(МНК). Свойства оценок на основе МНК

Ввиду четкой интерпретации параметров наиболее широко используется линейная функция. В линейной множественной регрессии $\hat{y}_x = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$ параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

Рассмотрим линейную модель множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m + \varepsilon . \qquad (2.1)$$

Классический подход к оцениванию параметров линейной модели множественной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных $\hat{y}$ минимальна:

$$\sum_i \left(y_i - \hat{y}_{x_i}\right)^2 \rightarrow \min. \qquad (2.2)$$

Как известно из курса математического анализа, для того чтобы найти экстремум функции нескольких переменных, надо вычислить частные производные первого порядка по каждому из параметров и приравнять их к нулю.

Итак. Имеем функцию $m+1$ аргумента:

$$S(a, b_1, b_2, \ldots, b_m) = \sum \left(y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m\right)^2.$$

Находим частные производные первого порядка:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum \left(y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m\right) = 0; \\ \dfrac{\partial S}{\partial b_1} = -2\sum x_1 \left(y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m\right) = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial S}{\partial b_m} = -2\sum x_m \left(y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m\right) = 0. \end{cases}$$

После элементарных преобразований приходим к системе линейных нормальных уравнений для нахождения параметров линейного уравнения множественной регрессии (2.1):

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 + \ldots + b_m\sum x_m = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1 x_2 + \ldots + b_m\sum x_1 x_m = \sum y x_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a\sum x_m + b_1\sum x_1 x_m + b_2\sum x_2 x_m + \ldots + b_m\sum x_m^2 = \sum y x_m. \end{cases} \qquad (2.3)$$

Для двухфакторной модели данная система будет иметь вид:

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1 x_2 = \sum y x_1, \\ a\sum x_2 + b_1\sum x_1 x_2 + b_2\sum x_2^2 = \sum y x_2. \end{cases}$$

Метод наименьших квадратов применим и к уравнению множественной регрессии в стандартизированном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_m t_{x_m} + \varepsilon, \qquad (2.4)$$

где t_y, t_{x_1}, ..., t_{x_m} – стандартизированные переменные: $t_y = \dfrac{y - \bar{y}}{\sigma_y}$, $t_{x_i} = \dfrac{x_i - \bar{x}_i}{\sigma_{x_i}}$, для которых среднее значение равно нулю: $\bar{t}_y = \bar{t}_{x_i} = 0$, а среднее квадратическое отклонение равно единице: $\sigma_{t_y} = \sigma_{t_{x_i}} = 1$; β_i – стандартизированные коэффициенты регрессии.

Стандартизованные коэффициенты регрессии показывают, на сколько единиц изменится в среднем результат, если соответствующий фактор x_i изменится на одну единицу при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i можно сравнивать между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

Применяя МНК к уравнению множественной регрессии в стандартизированном масштабе, получим систему нормальных уравнений вида

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{yx_m} = \beta_1 r_{x_1x_m} + \beta_2 r_{x_2x_m} + \beta_3 r_{x_3x_m} + \ldots + \beta_m, \end{cases} \qquad (2.5)$$

где r_{yx_i} и $r_{x_ix_j}$ – коэффициенты парной и межфакторной корреляции.

Коэффициенты «чистой» регрессии b_i связаны со стандартизованными коэффициентами регрессии β_i следующим образом:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}. \qquad (2.6)$$

Поэтому можно переходить от уравнения регрессии в стандартизованном масштабе (2.4) к уравнению регрессии в натуральном масштабе переменных (2.1), при этом параметр a определяется как $a = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - \ldots - b_m\overline{x}_m$.

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет их использовать при отсеве факторов – из модели исключаются факторы с наименьшим значением β_i.

На основе линейного уравнения множественной регрессии

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_mx_m + \varepsilon \qquad (2.7)$$

могут быть найдены частные уравнения регрессии:

$$\begin{cases} \hat{y}_{x_1 \cdot x_2, x_3, \ldots, x_m} = f(x_1), \\ \hat{y}_{x_2 \cdot x_1, x_3, \ldots, x_m} = f(x_2), \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{y}_{x_m \cdot x_1, x_2, \ldots, x_{m-1}} = f(x_m), \end{cases} \qquad (2.8)$$

т.е. уравнения регрессии, которые связывают результативный признак с соответствующим фактором x_i при закреплении остальных факторов на среднем уровне. В развернутом виде систему (2.8) можно переписать в виде:

$$\begin{cases} y_{x_1 \cdot x_2, x_3, \ldots, x_m} = a + b_1x_1 + b_2\overline{x}_2 + b_3\overline{x}_3 + \ldots + b_m\overline{x}_m + \varepsilon, \\ y_{x_2 \cdot x_1, x_3, \ldots, x_m} = a + b_1\overline{x}_1 + b_2x_2 + b_3\overline{x}_3 + \ldots + b_m\overline{x}_m + \varepsilon, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_m \cdot x_1, x_2, \ldots, x_{m-1}} = a + b_1\overline{x}_1 + b_2\overline{x}_2 + b_3\overline{x}_3 + \ldots + b_mx_m + \varepsilon. \end{cases}$$

При подстановке в эти уравнения средних значений соответствующих факторов они принимают вид парных уравнений линейной регрессии, т.е. имеем

$$\begin{cases} \hat{y}_{x_1 \cdot x_2, x_3, \ldots, x_m} = A_1 + b_1x_1, \\ \hat{y}_{x_2 \cdot x_1, x_3, \ldots, x_m} = A_2 + b_2x_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{y}_{x_m \cdot x_1, x_2, \ldots, x_{m-1}} = A_m + b_mx_m, \end{cases} \qquad (2.9)$$

где

$$\begin{cases} A_1 = a + b_2\overline{x}_2 + b_3\overline{x}_3 + \ldots + b_m\overline{x}_m, \\ A_2 = a + b_1\overline{x}_1 + b_3\overline{x}_3 + \ldots + b_m\overline{x}_m, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_m = a + b_1\overline{x}_1 + b_2\overline{x}_2 + b_3\overline{x}_3 + \ldots + b_{m-1}\overline{x}_{m-1}. \end{cases}$$

В отличие от парной регрессии частные уравнения регрессии характеризуют изолированное влияние фактора на результат, ибо другие факторы закреплены на неизменном уровне. Эффекты влияния других факторов присоединены в них к свободному члену уравнения множественной регрессии. Это позволяет на основе частных уравнений регрессии определять частные коэффициенты эластичности:

$$Э_{y_{x_i}} = b_i \cdot \frac{x_i}{\hat{y}_{x_i \cdot x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m}}, \qquad (2.10)$$

где b_i – коэффициент регрессии для фактора x_i в уравнении множественной регрессии, $\hat{y}_{x_i \cdot x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m}$ – частное уравнение регрессии.

Наряду с частными коэффициентами эластичности могут быть найдены средние по совокупности показатели эластичности:

$$\overline{Э}_i = b_i \cdot \frac{\overline{x}_i}{\overline{y}_{x_i}}, \qquad (2.11)$$

Рассмотрим **пример**[4] (для сокращения объема вычислений ограничимся только десятью наблюдениями). Пусть имеются следующие данные (условные) о сменной добыче угля на одного рабочего y (т), мощности пласта x_1 (м) и уровне механизации работ x_2 (%), характеризующие процесс добычи угля в 10 шахтах.

Таблица 2.2

№	1	2	3	4	5	6	7	8	9	10
x_1	8	11	12	9	8	8	9	9	8	12
x_2	5	8	8	5	7	8	6	4	5	7
y	5	10	10	7	5	6	6	5	6	8

Предполагая, что между переменными y, x_1, x_2 существует линейная корреляционная зависимость, найдем уравнение регрессии y по x_1 и x_2.

Для удобства дальнейших вычислений составляем таблицу ($\varepsilon = y - \hat{y}_x$):

Таблица 2.3

№	x_1	x_2	y	x_1^2	x_2^2	y^2	$x_1 \cdot x_2$	$x_1 \cdot y$	$x_2 \cdot y$	$\hat{y}_x$	ε^2
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	8	5	5	64	25	25	40	40	25	5,13	0,016
2	11	8	10	121	64	100	88	110	80	8,79	1,464
3	12	8	10	144	64	100	96	120	80	9,64	0,127
4	9	5	7	81	25	49	45	63	35	5,98	1,038
5	8	7	5	64	49	25	56	40	35	5,86	0,741
6	8	8	6	64	64	36	64	48	48	6,23	0,052
7	9	6	6	81	36	36	54	54	36	6,35	0,121
8	9	4	5	81	16	25	36	45	20	5,61	0,377
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
9	8	5	6	64	25	36	40	48	30	5,13	0,762
10	12	7	8	144	49	64	84	96	56	9,28	1,631
Сумма	94	63	68	908	417	496	603	664	445	68	6,329
Среднее значение	9,4	6,3	6,8	90,8	41,7	49,6	60,3	66,4	44,5	–	–
σ^2	2,44	2,01	3,36	–	–	–	–	–	–	–	–
σ	1,56	1,42	1,83	–	–	–	–	–	–	–	–

Для нахождения параметров уравнения регрессии в данном случае необходимо решить следующую систему нормальных уравнений:

$$\begin{cases} 10a + 94b_1 + 63b_2 = 68, \\ 94a + 908b_1 + 603b_2 = 664, \\ 63a + 603b_1 + 417b_2 = 445. \end{cases}$$

Откуда получаем, что $a = -3,54$, $b_1 = 0,854$, $b_2 = 0,367$. Т.е. получили следующее уравнение множественной регрессии:

$$\hat{y}_x = -3,54 + 0,854 \cdot x_1 + 0,367 \cdot x_2.$$

Оно показывает, что при увеличении только мощности пласта x_1 (при неизменном x_2) на 1 м добыча угля на одного рабочего y увеличится в среднем на 0,854 т, а при увеличении только уровня механизации работ x_2 (при неизменном x_1) на 1% – в среднем на 0,367 т.

Найдем уравнение множественной регрессии в стандартизованном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \varepsilon,$$

при этом стандартизованные коэффициенты регрессии будут

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 0,854 \cdot \frac{1,56}{1,83} = 0,728,$$

$$\beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 0,367 \cdot \frac{1,42}{1,83} = 0,285.$$

Т.е. уравнение будет выглядеть следующим образом:

$$\hat{t}_y = 0,728 \cdot t_{x_1} + 0,285 \cdot t_{x_2}.$$

Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что мощность пласта оказывает большее влияние на сменную добычу угля, чем уровень механизации работ.

Сравнивать влияние факторов на результат можно также при помощи средних коэффициентов эластичности (2.11):

$$\overline{Э}_i = b_i \cdot \frac{\overline{x}_i}{\overline{y}_{x_i}}.$$

Вычисляем:

$$\overline{Э}_1 = 0,854 \cdot \frac{9,4}{6,8} = 1,18, \qquad \overline{Э}_2 = 0,367 \cdot \frac{6,3}{6,8} = 0,34.$$

Т.е. увеличение только мощности пласта (от своего среднего значения) или только уровня механизации работ на 1% увеличивает в среднем сменную добычу угля на 1,18% или 0,34% соответственно. Таким образом, подтверждается большее влияние на результат y фактора x_1, чем фактора x_2.

Проверка существенности факторов и показатели качества регрессии

Практическая значимость уравнения множественной регрессии оценивается с помощью показателя множественной корреляции и его квадрата– показателя детерминации.

Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком или, иначе, оценивает тесноту совместного влияния факторов на результат.

Независимо от формы связи показатель множественной к

$$R_{yx_1x_2\ldots x_m} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}},$$

где σ^2_y – общая дисперсия результативного признака; $\sigma^2_{ост}$ – остаточная дисперсия.

Границы изменения индекса множественной корреляции от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака со всем набором исследуемых факторов. Величина индекса множественной корреляции должна быть больше или равна максимальному парному индексу корреляции:

$$R_{yx_1x_2\ldots x_m} \geq r_{yx_i(\max)} \quad \left(i = \overline{1,m}\right).$$

Сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

Расчет индекса множественной корреляции предполагает определение уравнения множественной регрессии и на его основе остаточной дисперсии:

$$\sigma^2_{ост} = \frac{1}{n}\sum\left(y - \hat{y}_{x_1x_2-x_m}\right)^2. \qquad (2.13)$$

Можно пользоваться следующей формулой индекса множественной детерминации:

$$R^2_{yx_1x_2-x_m} = 1 - \frac{\sum\left(y - \hat{y}_{x_1x_2-x_m}\right)^2}{\sum\left(y - \bar{y}\right)^2}. \qquad (2.14)$$

При линейной зависимости признаков формула индекса множественной корреляции может быть представлена следующим выражением:

$$R_{yx_1x_2-x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}}, \qquad (2.15)$$

где β_i – стандартизованные коэффициенты регрессии; r_{yx_i} – парные коэффициенты корреляции результата с каждым фактором.

Формула индекса множественной корреляции для линейной регрессии получила название *линейного коэффициента множественной корреляции*, или, что то же самое, *совокупного коэффициента корреляции*.

Возможно также при линейной зависимости определение совокупного коэффициента корреляции через матрицу парных коэффициентов корреляции:

$$R_{yx_1x_2\ldots x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}, \qquad (2.16)$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \ldots & r_{yx_p} \\ r_{yx_1} & 1 & r_{x_1x_2} & \ldots & r_{x_1x_p} \\ r_{yx_2} & r_{x_2x_1} & 1 & \ldots & r_{x_2x_p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{yx_p} & r_{x_px_1} & r_{x_px_2} & \ldots & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & \ldots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{x_px_1} & r_{x_px_2} & \ldots & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Скорректированный индекс множественной корреляции содержит поправку на число степеней свободы, а именно остаточная сумма квадратов $\sum\left(y-\hat{y}_{x_1x_2-x_m}\right)^2$ делится на число степеней свободы остаточной вариации $(n-m-1)$, а общая сумма квадратов отклонений $\sum(y-\bar{y})^2$ на число степеней свободы в целом по совокупности $(n-1)$.

Формула скорректированного индекса множественной детерминации имеет вид:

$$\hat{R}^2 = 1-\frac{\sum\left(y-\hat{y}\right)^2/(n-m-1)}{\sum(y-\bar{y})/(n-1)}, \qquad (2.17)$$

где m – число параметров при переменных x; n – число наблюдений.

Поскольку $\dfrac{\sum\left(y-\hat{y}_{x_1x_2-x_m}\right)^2}{\sum(y-\bar{y})^2}=1-R^2$, то величину скорректированного индекса детерминации можно представить в виде:

$$\hat{R}^2 = 1-\left(1-R^2\right)\cdot\frac{n-1}{n-m-1}. \qquad (2.17а)$$

Чем больше величина m, тем сильнее различия $\hat{R}^2$ и R^2.

Как было показано выше, ранжирование факторов, участвующих во множественной линейной регрессии, может быть проведено через стандартизованные коэффициенты регрессии (β-коэффициенты). Эта же цель может быть достигнута с помощью частных коэффициентов корреляции (для линейных связей). Кроме того, частные показатели корреляции широко используются при решении проблемы отбора факторов: целесообразность включения того или иного фактора в модель можно доказать величиной показателя частной корреляции.

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при элиминировании (устранении влияния) других факторов, включенных в уравнение регрессии.

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

В общем виде при наличии m факторов для уравнения

$$y = a + b_1x_1 + b_2x_2 \ldots + b_mx_m + \varepsilon$$

коэффициент частной корреляции, измеряющий влияние на y фактора x_j, при неизменном уровне других факторов, можно определить по формуле:

$$r_{yx_j\cdot x_1x_2-x_{j-1}x_{j+1}-x_m} = \sqrt{1-\frac{1-R^2_{yx_1x_2-x_j-x_m}}{1-R^2_{yx_1x_2-x_{j-1}x_{j+1}-x_m}}}, \qquad (2.18)$$

где $R^2_{yx_1x_2-x_j-x_m}$ – множественный коэффициент детерминации всех m факторов с результатом; $R^2_{yx_1x_2-x_{j-1}x_{j+1}-x_m}$ – тот же показатель детерминации, но без введения в модель фактора x_j.

При двух факторах формула (2.18) примет вид:

$$r_{yx_1\cdot x_2} = \sqrt{1-\frac{1-R^2_{yx_1x_2}}{1-r^2_{yx_2}}}; \qquad r_{yx_2\cdot x_1} = \sqrt{1-\frac{1-R^2_{yx_1x_2}}{1-r^2_{yx_1}}}. \qquad (2.18а)$$

Порядок частного коэффициента корреляции определяется количеством факторов, влияние которых исключается. Например, $r_{yx_1 \cdot x_2}$ – коэффициент частной корреляции первого порядка. Соответственно коэффициенты парной корреляции называются коэффициентами нулевого порядка. Коэффициенты частной корреляции более высоких порядков можно определить через коэффициенты частной корреляции более низких порядков по рекуррентной формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}} - r_{yx_m \cdot x_1 x_2 \ldots x_{m-1}} \cdot r_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}}{\sqrt{\left(1 - r^2_{yx_m \cdot x_1 x_2 \ldots x_{m-1}}\right) \cdot \left(1 - r^2_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}\right)}}. \quad (2.19)$$

При двух факторах данная формула примет вид:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}; \; r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}. \quad (2.19а)$$

Для уравнения регрессии с тремя факторами частные коэффициенты корреляции второго порядка определяются на основе частных коэффициентов корреляции первого порядка. Так, по уравнению $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$ возможно исчисление трех частных коэффициентов корреляции второго порядка:

$$r_{yx_1 \cdot x_2 x_3}, \; r_{yx_2 \cdot x_1 x_3}, \; r_{yx_3 \cdot x_1 x_2},$$

каждый из которых определяется по рекуррентной формуле. Например, при $i = 1$ имеем формулу для расчета $r_{yx_1 \cdot x_2 x_3}$:

$$r_{yx_1 \cdot x_2 x_3} = \frac{r_{yx_1 \cdot x_2} - r_{yx_3 \cdot x_2} \cdot r_{x_1 x_3 \cdot x_2}}{\sqrt{\left(1 - r^2_{yx_3 \cdot x_2}\right)\left(1 - r^2_{x_1 x_3 \cdot x_2}\right)}}. \quad (2.20)$$

Рассчитанные по рекуррентной формуле частные коэффициенты корреляции изменяются в пределах от –1 до +1, а по формулам через множественные коэффициенты детерминации – от 0 до 1. Сравнение их друг с другом позволяет ранжировать факторы по тесноте их связи с результатом. Частные коэффициенты корреляции дают меру тесноты связи каждого фактора с результатом в чистом виде. Если из стандартизованного уравнения регрессии $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \beta_3 t_{x_3} + \varepsilon$ следует, что $\beta_1 > \beta_2 > \beta_3$, т.е. по силе влияния на результат порядок факторов таков: x_1, x_2, x_3, то этот же порядок факторов определяется и по соотношению частных коэффициентов корреляции, $r_{yx_1 \cdot x_2 x_3} > r_{yx_2 \cdot x_1 x_3} > r_{yx_3 \cdot x_1 x_2}$.

В эконометрике частные коэффициенты корреляции обычно не имеют самостоятельного значения. Их используют на стадии формирования модели. Так, строя многофакторную модель, на первом шаге определяется уравнение регрессии с полным набором факторов и рассчитывается матрица частных коэффициентов корреляции. На втором шаге отбирается фактор с наименьшей и несущественной по t -критерию Стьюдента величиной показателя частной корреляции. Исключив его из модели, строится новое уравнение регрессии. Процедура продолжается до тех пор, пока не окажется, что все частные коэффициенты корреляции существенно отличаются от нуля. Если исключен несущественный фактор, то множественные коэффициенты детерминации на двух смежных шагах построения регрессионной модели почти не отличаются друг от друга, $R^2_{m+1} \approx R^2_m$, где m – число факторов.

Из приведенных выше формул частных коэффициентов корреляции видна связь этих показателей с совокупным коэффициентом корреляции. Зная частные коэффициенты корреляции (последовательно первого, второго и более высокого порядка), можно определить совокупный коэффициент корреляции по формуле:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1-\left(1-r^2_{yx_1}\right)\cdot\left(1-r^2_{yx_2x_1}\right)\cdot\left(1-r^2_{yx_3x_1x_2}\right)\cdot\ldots\cdot\left(1-r^2_{yx_mx_1x_2\ldots x_{m-1}}\right)}. \quad (2.21)$$

В частности, для двухфакторного уравнения формула (2.21) принимает вид:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1-\left(1-r^2_{yx_1}\right)\cdot\left(1-r^2_{yx_2x_1}\right)}. \quad (2.21а)$$

При полной зависимости результативного признака от исследуемых факторов коэффициент совокупного их влияния равен единице. Из единицы вычитается доля остаточной вариации результативного признака $\left(1-r^2\right)$, обусловленная последовательно включенными в анализ факторами. В результате подкоренное выражение характеризует совокупное действие всех исследуемых факторов.

Значимость уравнения множественной регрессии в целом, так же как и в парной регрессии, оценивается с помощью F-критерия Фишера:

$$F = \frac{S_{\text{факт}}}{S_{\text{ост}}} = \frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m}, \quad (2.22)$$

где $S_{\text{факт}}$ – факторная сумма квадратов на одну степень свободы; $S_{\text{ост}}$ – остаточная сумма квадратов на одну степень свободы; R^2 – коэффициент (индекс) множественной детерминации; m – число параметров при переменных x (в линейной регрессии совпадает с числом включенных в модель факторов); n – число наблюдений.

Оценивается значимость не только уравнения в целом, но и фактора, дополнительно включенного в регрессионную модель. Необходимость такой оценки связана с тем, что не каждый фактор, вошедший в модель, может существенно увеличивать долю объясненной вариации результативного признака. Кроме того, при наличии в модели нескольких факторов они могут вводиться в модель в разной последовательности. Ввиду корреляции между факторами значимость одного и того же фактора может быть разной в зависимости от последовательности его введения в модель. Мерой для оценки включения фактора в модель служит частный F-критерий, т.е. F_{x_j}.

Частный F-критерий построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительно включенного фактора, с остаточной дисперсией на одну степень свободы по регрессионной модели в целом. В общем виде для фактора x_j частный F-критерий определится как

$$F_{x_j} = \frac{R^2_{yx_1\ldots x_j\ldots x_m} - R^2_{yx_1\ldots x_{j-1}x_{j+1}\ldots x_m}}{1-R^2_{yx_1\ldots x_j\ldots x_m}}\cdot\frac{n-m-1}{1}, \quad (2.23)$$

где $R^2_{yx_1\ldots x_j\ldots x_m}$ – коэффициент множественной детерминации для модели с полным набором факторов, $R^2_{yx_1\ldots x_{j-1}x_{j+1}\ldots x_m}$ – тот же показатель, но без включения в модель фактора x_j, n – число наблюдений, m – число параметров в модели (без свободного члена).

Фактическое значение частного F-критерия сравнивается с табличным при уровне значимости α и числе степеней свободы: 1 и $n-m-1$. Если фактическое значение F_{x_j} превышает $F_{\text{табл}}\left(\alpha, k_1, k_2\right)$, то дополнительное включение фактора x_j в модель статистически оправданно и коэффициент чистой регрессии b_j при факторе x_j статистически значим. Если же фактическое значение F_{x_j} меньше табличного, то дополнительное включение в модель фактора x_j не увеличивает существенно долю объясненной вариации признака y,

следовательно, нецелесообразно его включение в модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

Для двухфакторного уравнения частные F-критерии имеют вид:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-3), \quad F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-3). \quad (2.23а)$$

С помощью частного F-критерия можно проверить значимость всех коэффициентов регрессии в предположении, что каждый соответствующий фактор x_i вводился в уравнение множественной регрессии последним.

Частный F-критерий оценивает значимость коэффициентов чистой регрессии. Зная величину F_{x_i}, можно определить и t-критерий для коэффициента регрессии при i-м факторе, t_{b_i}, а именно:

$$t_{b_i} = \sqrt{F_{x_i}}. \quad (2.24)$$

Оценка значимости коэффициентов чистой регрессии по t-критерию Стьюдента может быть проведена и без расчета частных F-критериев. В этом случае, как и в парной регрессии, для каждого фактора используется формула:

$$t_{b_i} = \frac{b_i}{m_{b_i}}, \quad (2.25)$$

где b_i – коэффициент чистой регрессии при факторе x_i, m_{b_i} – средняя квадратическая (стандартная) ошибка коэффициента регрессии b_i.

Для уравнения множественной регрессии $\hat{y} = a + b_1x_1 + b_2x_2 + \ldots + b_mx_m$ средняя квадратическая ошибка коэффициента регрессии может быть определена по следующей формуле:

$$m_{b_i} = \frac{\sigma_y \sqrt{1 - R^2_{yx_1 \ldots x_m}}}{\sigma_{x_i} \sqrt{1 - R^2_{x_ix_1 \ldots x_m}}} \cdot \frac{1}{\sqrt{n-m-1}}, \quad (2.26)$$

где σ_y – среднее квадратическое отклонение для признака y, σ_{x_i} – среднее квадратическое отклонение для признака x_i, $R^2_{yx_1 \ldots x_m}$ – коэффициент детерминации для уравнения множественной регрессии, $R^2_{x_ix_1 \ldots x_m}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии; $n-m-1$ – число степеней свободы для остаточной суммы квадратов отклонений.

Как видим, чтобы воспользоваться данной формулой, необходимы матрица межфакторной корреляции и расчет по ней соответствующих коэффициентов детерминации $R^2_{x_ix_1 \ldots x_m}$. Так, для уравнения $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$ оценка значимости коэффициентов регрессии b_1, b_2, b_3 предполагает расчет трех межфакторных коэффициентов детерминации: $R^2_{x_1 \cdot x_2x_3}$, $R^2_{x_2 \cdot x_1x_3}$, $R^2_{x_3 \cdot x_1x_2}$.

Взаимосвязь показателей частного коэффициента корреляции, частного F-критерия и t-критерия Стьюдента для коэффициентов чистой регрессии может использоваться в процедуре отбора факторов. Отсев факторов при построении уравнения регрессии методом исключения практически можно осуществлять не только по частным коэффициентам корреляции, исключая на каждом шаге фактор с наименьшим незначимым значением частного коэффициента корреляции, но и по величинам t_{b_i} и F_{x_i}. Частный F-критерий широко используется и при построении модели методом включения переменных и шаговым регрессионным методом.