

## **ТЕМА 2. ПОСТРОЕНИЕ ОБОБЩЕННОЙ РЕГРЕССИОННОЙ МОДЕЛИ**

### **Расчет параметров модели в матричной форме**

Рассмотренная линейная парная регрессионная модель является частным случаем, так как на большинство экономических процессов и явлений оказывает влияние множество факторов. Например, уровень потребления населения зависит от величины доходов, уровня цен, экономической ситуации, состава семьи, склонности к сбережению и т.п. Производительность труда рабочего зависит от технического уровня производства, степени его механизации и автоматизации, квалификации рабочего, организационного уровня производства, морального климата в коллективе и других факторов.

Многофакторные модели требуют для своего построения большего объема информации и, как следствие, большего, чем однофакторные объема вычислений. В связи с этим для их разработки широко применяются пакеты прикладных программ с расчетами на ПЭВМ.

Этапы построения многофакторной модели такие же как и однофакторной. Напомним их:

- постановка задачи;
- спецификация модели;
- формирование массива информации;
- оценка параметров модели методом МНК;
- проверка модели и ее параметров на адекватность;
- анализ полученных результатов, прогнозирование.

Между отдельными этапами могут возникать обратные связи. В частности, после проверки модели на адекватность может несколько раз повторяться этап спецификации модели для того, чтобы обеспечить построение адекватной модели.

Ошибки спецификации можно разбить на три группы

- не включение в модель существенно влияющего фактора (факторов) – приводит к неточным (смещенным) оценкам;
- включение в модель несущественно влияющего фактора (факторов) – гарантирует неверное измерение количественной связи между показателями и, как результат, переменная, которая не должна присутство-

вать в модели, может показать существенный уровень связи с зависимой переменной;

- использование видов зависимостей, не соответствующих истинной форме связи, – приводит к получению смещенных оценок.

Для многофакторных моделей условия применения метода наименьших квадратов остаются теми же, что и для однофакторной. Рассмотрим расчет параметров модели в матричной форме.

С этой целью введем обозначения. Пусть

$Y$  – вектор-столбец наблюдений зависимой переменной  $y$ ,

$X$  – матрица наблюдений независимых переменных (включает в себя столбец единиц с целью расчета свободного члена уравнения регрессии),

$\hat{A}$  – вектор-столбец оцениваемых параметров регрессионной модели,

$u$  – вектор-столбец остатков (отклонений, ошибок).

Тогда, исходя из общей записи уравнения регрессии

$$Y = XA + u,$$

имеем

$$u = Y - \hat{Y},$$

где  $\hat{Y} = X\hat{A}$  – оценка уравнения регрессии.

Сумма квадратов остатков определяется из выражения

$$\sum u^2 = u'u = (Y - X\hat{A})'(Y - X\hat{A}) = Y'Y - 2\hat{A}'X'Y + \hat{A}'X'X\hat{A}.$$

В данном и в последующих выражениях символ  $(')$  определяет операцию транспонирования.

Для нахождения параметров матрицы  $\hat{A}$  продифференцируем полученное выражение по  $\hat{A}$  и приравняем производные к нулю

$$\frac{\partial(u'u)}{\partial \hat{A}} = -2X'Y + 2X'X\hat{A} = 0,$$

откуда

$$\hat{A} = (X'X)^{-1}X'Y.$$

Рассмотрим на **примере** расчет параметров уравнения регрессии. Пусть нам необходимо определить, как товарооборот сети магазинов зависит

от их торговой площади и среднедневной интенсивности потока покупателей на основе следующего набора данных

| Магазин | Товарооборот, сотен тыс. грн. | Торговая площадь, тыс. м <sup>2</sup> | Среднедневная интенсивность потока покупателей, тыс. чел. |
|---------|-------------------------------|---------------------------------------|-----------------------------------------------------------|
| 1       | 2,93                          | 0,31                                  | 10,24                                                     |
| 2       | 5,27                          | 0,98                                  | 7,51                                                      |
| 3       | 6,85                          | 1,21                                  | 10,81                                                     |
| 4       | 7,01                          | 1,29                                  | 9,89                                                      |
| 5       | 7,02                          | 1,12                                  | 13,72                                                     |
| 6       | 8,35                          | 1,49                                  | 13,92                                                     |
| 7       | 4,33                          | 0,78                                  | 8,54                                                      |
| 8       | 5,77                          | 0,94                                  | 12,36                                                     |
| 9       | 7,68                          | 1,29                                  | 12,27                                                     |
| 10      | 3,16                          | 0,48                                  | 11,01                                                     |
| 11      | 1,52                          | 0,24                                  | 8,25                                                      |
| 12      | 3,15                          | 0,55                                  | 9,31                                                      |

Исходя из этого, имеем следующие матрицы исходных данных (матрицы  $X$  и  $X'$  содержат единичные векторы для расчета свободного члена  $\hat{a}_0$  уравнения регрессии)

$$X = \begin{pmatrix} 1 & 0,31 & 10,24 \\ 1 & 0,98 & 7,51 \\ 1 & 1,21 & 10,81 \\ 1 & 1,29 & 9,89 \\ 1 & 1,12 & 13,72 \\ 1 & 1,49 & 13,92 \\ 1 & 0,78 & 8,54 \\ 1 & 0,94 & 12,36 \\ 1 & 1,29 & 12,27 \\ 1 & 0,48 & 11,01 \\ 1 & 0,24 & 8,25 \\ 1 & 0,55 & 9,31 \end{pmatrix}, \quad Y = \begin{pmatrix} 2,93 \\ 5,27 \\ 6,85 \\ 7,01 \\ 7,02 \\ 8,35 \\ 4,33 \\ 5,77 \\ 7,68 \\ 3,16 \\ 1,52 \\ 3,15 \end{pmatrix},$$

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0,31 & 0,98 & 1,21 & 1,29 & 1,12 & 1,49 & 0,78 & 0,94 & 1,29 & 0,48 & 0,24 & 0,55 \\ 10,24 & 7,51 & 10,81 & 9,89 & 13,72 & 13,92 & 8,54 & 12,36 & 12,27 & 11,01 & 8,25 & 9,31 \end{pmatrix}.$$

Найдем составляющие формулы  $\hat{A} = (X'X)^{-1}X'Y$  для расчета параметров вектора  $\hat{A}$

$$\hat{A} = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{pmatrix}.$$

Рассчитаем произведение матриц  $X'X$

$$X'X = \begin{pmatrix} 12 & 10,68 & 127,83 \\ 10,68 & 11,4058 & 118,973 \\ 127,83 & 118,973 & 1410,14 \end{pmatrix}.$$

Найдем обратную по отношению к полученной матрицу одним из известных способов (методом Жордана–Гаусса, с помощью присоединенной матрицы и т.п.)

$$(X'X)^{-1} = \begin{pmatrix} 2,47452 & 0,18978 & -0,2403 \\ 0,18978 & 0,74547 & -0,0801 \\ -0,2403 & -0,0801 & 0,02925 \end{pmatrix}.$$

Рассчитаем произведение  $X'Y$

$$X'Y = \begin{pmatrix} 63,04 \\ 66,0611 \\ 704,692 \end{pmatrix}.$$

Наконец, подставив полученные значения выражений  $(X'X)^{-1}$  и  $X'Y$  в формулу для расчета  $\hat{A} = (X'X)^{-1} X'Y$ , найдем составляющие вектора  $\hat{A}$

$$\hat{A} = \begin{pmatrix} -0,8319 \\ 4,74295 \\ 0,17499 \end{pmatrix}.$$

Таким образом, регрессионная модель имеет вид:

$$\hat{y} = -0,8319 + 4,74295x_1 + 0,17499x_2$$

Интерпретация модели следующая: коэффициенты регрессии показывают, что при изменении размера торговой площади на одну тысячу м<sup>2</sup>, при прочих равных условиях, товарооборот увеличится на 4,74 сотен тыс. грн., а при увеличении среднедневного потока покупателей на одну тыс. чел., при прочих равных условиях, он возрастет на 0,175 сотен тыс. грн.

### **Свойства оценок параметров**

Рассчитанные оценки модели с помощью метода 1МНК должны обладать свойствами:

1. несмещенности;
2. обоснованности;
3. эффективности;
4. инвариантности.

Первое свойство предполагает, что оценка будет несмещенной, если ее математическое ожидание равно  $A$ :

$$M(\hat{A}) = A.$$

Данное заключение вытекает из следующих рассуждений. Используя уравнение регрессии в общем виде  $Y = XA + u$ , путем преобразований (умножая левую и правую части на  $(X'X)^{-1}X'$ ) приведем его к виду

$$(X'X)^{-1}X'Y = (X'X)^{-1}X'XA + (X'X)^{-1}X'u.$$

Так как  $(X'X)^{-1}X'X = E$ , то имеем

$$M[(X'X)^{-1}X'Y] = M[A + (X'X)^{-1}X'u]$$

или

$$M[\hat{A}] = A + (X'X)^{-1}X'M(u).$$

А, учитывая, что  $M(u) = 0$  в соответствии с первым условием применения метода МНК, то

$$M(\hat{A}) = A.$$

Таким образом, оценки параметров модели будут несмещенными, если математическое ожидание их выборочных значений, найденных при многократном повторении выборки, не отличается от истинного значения.

Смещение определяется как

$$Q = M(\hat{A}) - A.$$

Наличие смещенности (или ее отсутствие) может быть проверено с помощью отношения среднеквадратического отклонения к абсолютной величине оценки

$$\frac{\hat{\sigma}_{a_i}}{|\hat{a}_i|}.$$

В случае если величина отношения превышает 10%, то делают вывод о смещенности оценки.

Второе свойство может быть определено следующим образом: выборочная оценка  $\hat{A}$  называется обоснованной, если для любого сколь угодно малого числа  $\varepsilon > 0$  выполняется соотношение

$$\lim_{n \rightarrow \infty} P\{|\hat{A} - A| < \varepsilon\} = 1.$$

Иначе, оценка считается обоснованной, если она удовлетворяет закону больших чисел.

Третье свойство предполагает, что выборочные оценки вектора  $\hat{A}$  будут только тогда эффективными, когда их дисперсии будут наименьшими. Данное свойство определяется **теоремой Гаусса-Маркова**: функция оценивания по методу 1МНК покомпонентно минимизирует дисперсию всех линейных несмещенных функций вектора оценок  $\hat{A}$ :

$$\sigma_{\hat{A}}^2 \leq \sigma_{\bar{A}}^2 \text{ для } j = \overline{1, m},$$

где  $\sigma_{\hat{A}}^2$  – дисперсия оценок  $\hat{A}$ , определенная методом 1МНК,  $\sigma_{\bar{A}}^2$  – дисперсия оценок  $\bar{A}$ , определенных другими способами.

Исходя из этого, функция оценивания 1МНК  $\hat{A}$  в классической линейной модели является лучшей линейно несмещенной функцией оценивания (с англ. *BLUE – Best Linear Unbiased Estimator*).

Величина дисперсии оценок параметров зависит от количества наблюдений, спецификации модели и точности метода оценивания параметров.

Четвертое свойство дает возможность использовать функции от выборочных оценок. По определению: оценка  $\hat{A}$  параметров  $A$  называется инвариантной, если для произвольно заданной функции  $g$  оценка параметров функции  $g(A)$  представляется в виде  $g(\hat{A})$ .

Так, зная коэффициент множественной детерминации, можно найти коэффициент множественной корреляции, используя значение выборочной дисперсии, можно определить среднеквадратическое отклонение искомой величины и т.д.

### **Коэффициенты корреляции, детерминации, эластичности**

Для оценки тесноты связи между рассматриваемыми в модели показателями рассчитываются коэффициенты парной корреляции. Расчет коэффициентов парной корреляции может быть осуществлен по формулам

$$r_{y/x_j} = \frac{\text{cov}(x_j, y)}{\sqrt{\text{var}(x_j) \text{var}(y)}}; \quad r_{x_l/x_j} = \frac{\text{cov}(x_j, x_l)}{\sqrt{\text{var}(x_j) \text{var}(x_l)}}$$

или нижеописанным способом.

С этой целью исходные данные предварительно нормируются

$$y_i^* = \frac{y_i - \bar{y}}{\sigma_y}, \quad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}},$$

где  $y_i^*$  – нормированное значение  $i$ -го наблюдения зависимой переменной

$x_{ij}^*$  – нормированное значение  $i$ -го наблюдения  $j$ -й независимой переменной.

Далее, используя векторы нормированных значений переменных  $Y^*$  и  $X_j^*$ , рассчитывают коэффициенты парной корреляции

$$r_{y/x_j} = \frac{1}{n} (Y^*)' X_j^* \quad \text{и} \quad r_{x_l/x_j} = \frac{1}{n} (X_l^*)' X_j^*.$$

Для расчета коэффициентов парной корреляции найдем нормированные значения наблюдений переменных модели. Занесем данные в таблицу

| Наблюдение | $y^*$     | $x_1^*$   | $x_2^*$   |
|------------|-----------|-----------|-----------|
| 1          | -1,100378 | -1,457381 | -0,205323 |
| 2          | 0,007894  | 0,226145  | -1,564188 |
| 3          | 0,756214  | 0,804073  | 0,078396  |
| 4          | 0,831994  | 1,005091  | -0,379536 |
| 5          | 0,836730  | 0,577927  | 1,526856  |
| 6          | 1,466645  | 1,507636  | 1,626407  |
| 7          | -0,437309 | -0,276400 | -1,051502 |
| 8          | 0,244704  | 0,125636  | 0,849913  |
| 9          | 1,149319  | 1,005091  | 0,805115  |
| 10         | -0,991445 | -1,030218 | 0,177947  |
| 11         | -1,768183 | -1,633272 | -1,195851 |
| 12         | -0,996182 | -0,854327 | -0,668233 |

Рассчитаем коэффициенты парной корреляции

$$r_{y/x_1} = \frac{1}{12} \cdot (-1,1003 \quad 0,0079 \quad \dots \quad -0,9962) \cdot \begin{pmatrix} -1,4574 \\ 0,2261 \\ \vdots \\ -0,8543 \end{pmatrix} = 0,9843,$$

$$r_{y/x_2} = \frac{1}{12} (-1,1003 \quad 0,0079 \quad \dots \quad -0,9962) \cdot \begin{pmatrix} -0,2053 \\ -1,5642 \\ \vdots \\ -0,6682 \end{pmatrix} = 0,65141,$$

$$r_{x_1/x_2} = \frac{1}{12} \cdot (-1,4574 \quad 0,2261 \quad \dots \quad -0,8543) \cdot \begin{pmatrix} -0,2053 \\ -1,5642 \\ \vdots \\ -0,6682 \end{pmatrix} = 0,5424.$$

С учетом этого матрица коэффициентов парной корреляции выглядит следующим образом:

$$r = \begin{pmatrix} 1 & 0,9843 & 0,65141 \\ 0,9843 & 1 & 0,5424 \\ 0,65141 & 0,5424 & 1 \end{pmatrix}.$$

Коэффициент множественной детерминации может быть рассчитан следующим образом:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^n (Y - \bar{Y})^2} = \frac{\hat{A}' X' Y}{Y' Y}$$

или

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (Y - \hat{Y})^2}{\sum_{i=1}^n (Y - \bar{Y})^2} = 1 - \frac{u' u}{Y' Y}.$$

Данные формулы не учитывают числа степеней свободы и при увеличении в модели количества независимых переменных значение коэффициента множественной детерминации не уменьшается, а даже растет. В связи с этим в чистом виде он не может быть использован для проверки адекватности модели, так как при его расчете необходимо учитывать степени свободы (а с до-



бавлением каждой независимой переменной в модель без увеличения числа наблюдений они уменьшаются на единицу). В связи с этим используются формулы для расчета скорректированного коэффициента множественной детерминации  $\bar{R}^2$ . В частности

$$\bar{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k}.$$

Коэффициент множественной корреляции равен квадратному корню из коэффициента множественной детерминации

$$R = \sqrt{R^2}.$$

Он характеризует тесноту связи всех независимых переменных и зависимой. Его величина всегда положительна (в отличие от коэффициента парной корреляции).

Применение формул для расчета коэффициента множественной детерминации требует дополнительных расчетов составляющих этих формул.

| Магазин | $y$  | $y - \bar{y}$ | $(y - \bar{y})^2$ | $\hat{y}$ | $\hat{y} - \bar{y}$ | $(\hat{y} - \bar{y})^2$ | $(y - \bar{y})(\hat{y} - \bar{y})$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---------|------|---------------|-------------------|-----------|---------------------|-------------------------|------------------------------------|---------------|-------------------|
| 1       | 2,93 | -2,32         | 5,40              | 2,43024   | -2,82               | 7,97                    | 6,55896                            | 0,50          | 0,2498            |
| 2       | 5,27 | 0,02          | 0,00              | 5,1303    | -0,12               | 0,02                    | -0,00205                           | 0,14          | 0,0195            |
| 3       | 6,85 | 1,60          | 2,55              | 6,79864   | 1,55                | 2,39                    | 2,46734                            | 0,05          | 0,0026            |
| 4       | 7,01 | 1,76          | 3,09              | 7,01708   | 1,76                | 3,11                    | 3,09833                            | -0,01         | 0,0001            |
| 5       | 7,02 | 1,77          | 3,12              | 6,88099   | 1,63                | 2,65                    | 2,87553                            | 0,14          | 0,0193            |
| 6       | 8,35 | 3,10          | 9,59              | 8,67087   | 3,42                | 11,68                   | 10,58300                           | -0,32         | 0,1030            |
| 7       | 4,33 | -0,92         | 0,85              | 4,36195   | -0,89               | 0,79                    | 0,82303                            | -0,03         | 0,0010            |
| 8       | 5,77 | 0,52          | 0,27              | 5,78927   | 0,54                | 0,29                    | 0,27690                            | -0,02         | 0,0004            |
| 9       | 7,68 | 2,43          | 5,89              | 7,43355   | 2,18                | 4,75                    | 5,29068                            | 0,25          | 0,0607            |
| 10      | 3,16 | -2,09         | 4,38              | 3,37128   | -1,88               | 3,54                    | 3,93974                            | -0,21         | 0,0446            |
| 11      | 1,52 | -3,73         | 13,94             | 1,75001   | -3,50               | 12,27                   | 13,07905                           | -0,23         | 0,0529            |
| 12      | 3,15 | -2,10         | 4,42              | 3,40581   | -1,85               | 3,41                    | 3,88594                            | -0,26         | 0,0654            |
| Сумма   |      |               | 53,50             |           |                     | 52,88                   | 52,87650                           |               | 0,6194            |

В результате применения любой из ранее описанных формул получаем значение коэффициента множественной детерминации

$$R^2 = 0,98842.$$

Скорректированный коэффициент множественной детерминации для нашего примера будет равен

$$\bar{R}^2 = 0,98585.$$

Рассчитаем коэффициент множественной корреляции

$$R = \sqrt{0,98842} = 0,99419314.$$

Приведенные результаты позволяют сделать вывод о высокой тесноте связи между отобранными факторами и товарооборотом магазинов. Именно вариация данных факторов на 98,585% объясняет вариацию зависимой переменной.

Коэффициенты эластичности для линейной многофакторной модели рассчитываются по формуле

$$E_{y/x_j} = \frac{\partial y}{\partial x_j} \cdot \frac{\bar{x}_j}{\bar{y}}.$$

Вычислим коэффициенты эластичности, для чего сначала найдем средние величины используемых в модели показателей

$$\bar{y} = 5,25333, \quad \bar{x}_1 = 0,89, \quad \bar{x}_2 = 10,6525,$$

Искомые коэффициенты равны

$$E_{y/x_1} = \frac{\partial y}{\partial x_1} \cdot \frac{\bar{x}_1}{\bar{y}} = \hat{a}_1 \cdot \frac{\bar{x}_1}{\bar{y}} = 4,74295 \cdot \frac{0,89}{5,25333} = 0,80353,$$

$$E_{y/x_2} = \frac{\partial y}{\partial x_2} \cdot \frac{\bar{x}_2}{\bar{y}} = \hat{a}_2 \cdot \frac{\bar{x}_2}{\bar{y}} = 0,17499 \cdot \frac{10,6525}{5,25333} = 0,35483.$$

Таким образом, увеличение торговой площади на 1%, при прочих равных условиях, повлечет за собой рост товарооборота на 0,80353%, а увеличение среднедневной интенсивности потока покупателей также на 1%, при прочих равных условиях, повлечет рост товарооборота на 0,35483%.

### **Дисперсионно-ковариационная матрица**

Для оценки значимости параметров регрессионной модели и доверительных интервалов их изменения рассчитывается дисперсионно-ковариационная матрица  $Var(\hat{A})$

$$Var(\hat{A}) = \hat{\sigma}_u^2 (X'X)^{-1} = \begin{pmatrix} \hat{\sigma}_{a_0}^2 & cov(\hat{a}_0, \hat{a}_1) & \cdots & cov(\hat{a}_0, \hat{a}_m) \\ cov(\hat{a}_1, \hat{a}_0) & \hat{\sigma}_{a_1}^2 & \cdots & cov(\hat{a}_1, \hat{a}_m) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{a}_m, \hat{a}_0) & cov(\hat{a}_m, \hat{a}_1) & \cdots & \hat{\sigma}_{a_m}^2 \end{pmatrix}$$

Главная диагональ матрицы представлена оценками дисперсий  $\hat{\sigma}_{a_j}$   $j$ -х оценок параметров, вне главной диагонали находятся оценки ковариаций между  $\hat{a}_j$  и  $\hat{a}_i$ .

Чем большими будут значения дисперсий и ковариаций, тем менее существенной может оказаться значимость модели, и тем более широкими будут доверительные интервалы для оцененных параметров и зависящих от них интервалов прогноза зависимой переменной.

Составляющая формулы дисперсионно-ковариационной матрицы – оценка дисперсии ошибок  $\hat{\sigma}_u^2$  в матричной форме рассчитывается следующим образом:

$$\hat{\sigma}_u^2 = \frac{u'u}{n-k} = \frac{Y'Y - \hat{A}'X'Y}{n-k},$$

где  $n-k$  – число степеней свободы.

Для того чтобы осуществить расчет интересующей нас матрицы, определим предварительно расчетные значения  $\hat{y}_i$  и величины остатков  $u_i$

$$\hat{Y} = \begin{pmatrix} 2,43 \\ 5,13 \\ 6,799 \\ 7,017 \\ 6,881 \\ 8,671 \\ 4,362 \\ 5,789 \\ 7,434 \\ 3,372 \\ 1,75 \\ 3,406 \end{pmatrix}, \quad u = \begin{pmatrix} 0,5 \\ 0,14 \\ 0,051 \\ -0,007 \\ 0,139 \\ -0,321 \\ -0,032 \\ -0,019 \\ 0,246 \\ -0,211 \\ -0,23 \\ -0,256 \end{pmatrix}.$$

Рассчитаем составляющие формулы оценки дисперсии ошибок для нашего примера

$$Y'Y = 384,666,$$

$$\hat{A}'X'Y = 384,04664,$$

$$Y'Y - \hat{A}'X'Y = 0,61936.$$

С учетом этого имеем

$$\hat{\sigma}_u^2 = 0,06882.$$

Дисперсионно-ковариационная матрица для нашей задачи имеет вид

$$\begin{aligned} Var(\hat{A}) &= \hat{\sigma}_u^2 (X'X)^{-1} = 0,06882 \cdot \begin{pmatrix} 2,47452 & 0,18978 & -0,2403 \\ 0,18978 & 0,74547 & -0,0801 \\ -0,2403 & -0,0801 & 0,02925 \end{pmatrix} = \\ &= \begin{pmatrix} 0,17029 & 0,01306 & -0,0165 \\ 0,01306 & 0,0513 & -0,0055 \\ -0,0165 & -0,0055 & 0,00201 \end{pmatrix} \end{aligned}$$

Таким образом,

$$\begin{aligned} \hat{\sigma}_{a_0}^2 &= 0,17029 & \text{и, соответственно,} & \hat{\sigma}_{a_0} = 0,41266, \\ \hat{\sigma}_{a_1}^2 &= 0,0513, & & \hat{\sigma}_{a_1} = 0,2265, \\ \hat{\sigma}_{a_2}^2 &= 0,00201, & & \hat{\sigma}_{a_2} = 0,04487. \end{aligned}$$

С помощью найденных среднеквадратических отклонений и оценок параметров  $\hat{a}_j$  определим, являются ли найденные оценки смещенными, для чего рассчитаем отношения

$$\begin{aligned} \frac{\hat{\sigma}_{a_0}}{|\hat{a}_0|} &= \frac{0,41233}{|-0,8319|} = 0,496, \\ \frac{\hat{\sigma}_{a_1}}{|\hat{a}_1|} &= \frac{0,2265}{|4,74295|} = 0,048, \\ \frac{\hat{\sigma}_{a_2}}{|\hat{a}_2|} &= \frac{0,04487}{|0,17499|} = 0,256. \end{aligned}$$

Стандартные ошибки оценок относительно уровней параметров составляют, соответственно, 49,6%, 4,8% и 25,6%, что позволяет сделать вывод о смещенности оценок  $\hat{a}_0$  и  $\hat{a}_2$ .

### **Проверка значимости модели и ее параметров**

Проверка значимости модели осуществляется по критерию Фишера. Как известно,  $F$ -статистика Фишера рассчитывается по формуле

$$F_{\text{расч.}} = \frac{MSR}{MSE}, \text{ где } MSR = \frac{R^2}{k-1} \text{ и } MSE = \frac{1-R^2}{n-k}, \text{ откуда}$$

$$F_{\text{расч.}} = \frac{R^2(n-k)}{(1-R^2)(k-1)}.$$

Полученное значение сравнивается с табличным (критическим) значением критерия для уровня значимости  $\alpha$  и числа степеней свободы  $v_1 = k - 1$  и  $v_2 = n - k$ . Если

$$F_{\text{расч.}} > F_{(k-1, n-k, \alpha)},$$

то уравнение считается значимым.

Проверим значимость уравнения регрессии по критерию Фишера. В нашем случае

$$F_{\text{расч.}} = \frac{R^2(n-k)}{(1-R^2)(k-1)} = \frac{0,98842(12-3)}{(1-0,98842)(3-1)} = 384,1781.$$

Табличное значение  $F$ -критерия при  $\alpha = 0,05$ ,  $v_1 = k - 1 = 3 - 1 = 2$  и  $v_2 = n - k = 12 - 3 = 9$  равно

$$F_{(2,9,0.05)} = 4,26.$$

Так как  $384,1781 > 4,26$ , то уравнение считается значимым.

### **Стандартизированные коэффициенты регрессии ( $\beta$ -коэффициенты), коэффициенты частной детерминации**

Коэффициенты регрессии отражают соотношение между зависимой и независимыми переменными в тех единицах измерения, в каких они исчисляются. В силу этого, независимые переменные нельзя проранжировать по степени их влияния на результативный показатель по показателям коэффициента регрессии. Проблема может быть решена путем расчета стандартизированных коэффициентов регрессии (или  $\beta$ -коэффициентов), которые пока-

зывают влияние независимых переменных на зависимую в относительных единицах измерения

$$\beta_j = \hat{a}_j \frac{\sigma_{x_j}}{\sigma_y}.$$

В терминах стандартизированных коэффициентов регрессии строятся также уравнения регрессии (т.н. стандартизированное уравнение регрессии)

$$\check{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m.$$

Для условий нашего примера рассчитаем стандартизированные коэффициенты регрессии:

$$\beta_1 = 4,74295 \cdot \frac{0,39797}{2,1114} = 0,89399,$$

$$\beta_2 = 0,17449 \cdot \frac{2,00903}{2,1114} = 0,1665.$$

По величинам  $\beta$ -коэффициентов можно сделать вывод, что влияние первого показателя на  $y$  значительно выше, чем второго.

Построим стандартизированное уравнение множественной регрессии:

$$\check{y} = 0,89399x_1 + 0,1665x_2.$$

Помимо коэффициента множественной детерминации в многофакторной модели может быть рассчитан коэффициент частной детерминации. **Коэффициент частной детерминации**  $\Delta R_j^2$  показывает предельный (граничный) вклад  $j$ -го регрессора (независимой переменной) в  $R^2$  (общую вариацию).

Он также показывает, на какую величину уменьшится коэффициент множественной детерминации, если  $j$ -ю переменную исключить из модели

$$\Delta R_j^2 = r_{y/x_j} \cdot \beta_j$$

и

$$R^2 = \Delta R_1^2 + \Delta R_2^2$$

Частные коэффициенты детерминации равны:

$$\Delta R_1^2 = 0,984 \cdot 0,894 = 0,87996,$$

$$\Delta R_2^2 = 0,651 \cdot 0,1665 = 0,1084.$$

Исходя из этого, почти 88% процентов общей вариации переменной  $y$  объясняется вариацией показателя торговой площади ( $x_1$ ) и почти 11% – вариацией среднедневного потока покупателей ( $x_2$ ).

$$R^2 = \Delta R_1^2 + \Delta R_2^2 = 0,87996 + 0,1084 = 0,9884.$$

**Проверка значимости параметров модели  $\hat{a}_0, \hat{a}_1, \hat{a}_2, \dots, \hat{a}_n$  и расчет доверительных интервалов**

Аналогично парной зависимости, в многофакторной модели проверка значимости параметров  $\hat{a}_0, \hat{a}_1, \hat{a}_2, \dots, \hat{a}_m$  осуществляется по критерию Стьюдента. С этой целью для каждого из показателей рассчитывается  $t$ -статистика Стьюдента

$$t_{\hat{a}_0} = \frac{|\hat{a}_0|}{\hat{\sigma}_{a_0}},$$

$$t_{\hat{a}_1} = \frac{|\hat{a}_1|}{\hat{\sigma}_{a_1}},$$

$$t_{\hat{a}_2} = \frac{|\hat{a}_2|}{\hat{\sigma}_{a_2}},$$

$$\vdots$$

$$t_{\hat{a}_n} = \frac{|\hat{a}_n|}{\hat{\sigma}_{a_n}}.$$

Значения  $\hat{\sigma}_{a_j}$  вычисляются на основе информации из дисперсионно-ковариационной матрицы  $Var(\hat{A})$ .

Расчет доверительных интервалов осуществляется аналогично тому, как это было сделано для парной линейной зависимости

$$a_0^* = \hat{a}_0 \pm t_{\alpha/2} \cdot \hat{\sigma}_{a_0},$$

$$a_1^* = \hat{a}_1 \pm t_{\alpha/2} \cdot \hat{\sigma}_{a_1},$$

$$a_2^* = \hat{a}_2 \pm t_{\alpha/2} \cdot \hat{\sigma}_{a_2},$$

$$\vdots$$

$$a_m^* = \widehat{a}_m \pm t_{\alpha/2} \cdot \widehat{\sigma}_{a_m}.$$

Оценим значимость параметров регрессионного уравнения:

$$t_{\widehat{a}_0} = \frac{|\widehat{a}_0|}{\widehat{\sigma}_{a_0}} = \frac{|-0,831946|}{0,412663} = 2,016$$

$$t_{\widehat{a}_1} = \frac{|\widehat{a}_1|}{\widehat{\sigma}_{a_1}} = \frac{|4,742948|}{0,226498} = 20,94$$

$$t_{\widehat{a}_2} = \frac{|\widehat{a}_2|}{\widehat{\sigma}_{a_2}} = \frac{|0,174988|}{0,044868} = 3,9$$

Для  $\alpha = 0,05$  и  $n - k = 12 - 3 = 9$   $t_{(\alpha, n-k)} = 2,262$ . Сравнивая рассчитанные значения  $t$ -статистики с табличным, можно сделать вывод о значимости на данном уровне только двух параметров  $\widehat{a}_1$  и  $\widehat{a}_2$ .

Если задать уровень доверия  $\alpha = 0,1$ , то при  $n - k = 12 - 3 = 9$  значение критерия Стьюдента составит  $t_{(\alpha, n-k)} = 1,833$ , и показатель  $\widehat{a}_0$  также можно считать значимым.

Определим доверительные интервалы для параметров модели  $\widehat{a}_0$ ,  $\widehat{a}_1$  и  $\widehat{a}_2$ . В расчетах будем использовать табличное значение критерия Стьюдента с  $\alpha = 0,05$  и  $n - k = 12 - 3 = 9$ :  $t_{(\alpha/2, n-k)} = 2,262$ .

$$a_0^* = \widehat{a}_0 \pm t_{\alpha/2} \cdot \widehat{\sigma}_{a_0} = -0,8319 \pm 2,262 \cdot 0,412663 = -0,8319 \pm 0,933.$$

Откуда

$$-1,765 \leq a_0^* \leq 0,101,$$

$$a_1^* = \widehat{a}_1 \pm t_{\alpha/2} \cdot \widehat{\sigma}_{a_1} = 4,7429 \pm 2,262 \cdot 0,2265 = 4,7439 \pm 0,512,$$

$$4,23 \leq a_1^* \leq 5,255,$$

$$a_2^* = \widehat{a}_2 \pm t_{\alpha/2} \cdot \widehat{\sigma}_{a_2} = 0,175 \pm 2,262 \cdot 0,045 = 0,175 \pm 0,1015,$$



$$0,0735 \leq a_2^* \leq 0,2765.$$

### **Прогноз значений зависимой переменной $y$ и построение доверительных интервалов прогноза**

Различают точечный (для математического ожидания  $\hat{y}$ ) и интервальный (для любого значения переменной) прогнозы и, соответственно, доверительные интервалы для точечного и интервального прогнозов. Чтобы рассчитать данные величины находят среднюю (стандартную) ошибку прогноза.

Для точечного прогноза:

$$\begin{aligned}\hat{\sigma}_{n+1}^2 &= \hat{\sigma}_u^2 \cdot X'_{n+1} (X'X)^{-1} X_{n+1}, \\ \hat{\sigma}_{n+1} &= \hat{\sigma}_u \cdot \sqrt{X'_{n+1} (X'X)^{-1} X_{n+1}}.\end{aligned}$$

С учетом этого доверительный интервал находится из выражения

$$\hat{Y}_{n+1} - t_{\alpha/2} \hat{\sigma}_u \sqrt{X'_{n+1} (X'X)^{-1} X_{n+1}} \leq M(Y_{n+1}^*) \leq \hat{Y}_{n+1} + t_{\alpha/2} \hat{\sigma}_u \sqrt{X'_{n+1} (X'X)^{-1} X_{n+1}}.$$

Для интервального прогноза стандартная ошибка прогноза находится из выражения

$$\hat{\sigma}_{n+1(i)} = \hat{\sigma}_u \cdot \sqrt{1 + X'_{n+1} (X'X)^{-1} X_{n+1}}.$$

Доверительный интервал рассчитывается по формуле

$$\hat{Y}_{n+1} - t_{\alpha/2} \hat{\sigma}_u \sqrt{1 + X'_{n+1} (X'X)^{-1} X_{n+1}} \leq Y_{n+1(i)}^* \leq \hat{Y}_{n+1} + t_{\alpha/2} \hat{\sigma}_u \sqrt{1 + X'_{n+1} (X'X)^{-1} X_{n+1}}.$$

Найдем прогнозное значение  $y$  и доверительные интервалы, соответственно, для точечного и интервального прогнозов.

Пусть

$$X_{n+1} = X_{13} = \begin{pmatrix} 1 \\ 1,2 \\ 9 \end{pmatrix},$$

тогда прогнозное значение  $\hat{y}_{13}$  составит:

$$\hat{y}_{13} = -0,831946 + 4,742948 \cdot 1,2 + 0,174988 \cdot 9 = 6,434.$$

Значение  $\hat{\sigma}_u$  равно

$$\hat{\sigma}_u = \sqrt{\hat{\sigma}_u^2} = \sqrt{0,068818} = 0,26233,$$

отсюда

$$\begin{aligned}\hat{\sigma}_{13} &= 0,26233 \cdot \sqrt{(1 \quad 1,2 \quad 9) \cdot \begin{pmatrix} 2,474519 & 0,189775 & -0,240327 \\ 0,189775 & 0,745468 & -0,080098 \\ -0,240327 & -0,080098 & 0,029253 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1,2 \\ 9 \end{pmatrix}} = \\ &= 0,26233 \cdot \sqrt{0,31694} = 0,26233 \cdot 0,56298 = 0,14769.\end{aligned}$$

Для расчета доверительного интервала для точечного прогноза будем использовать табличное значение критерия Стьюдента с  $\alpha = 0,05$  и  $n - k = 12 - 3 = 9$ :  $t_{(\alpha/2, n-k)} = 2,262$ .

Итак, имеем:

$$\begin{aligned}6,434 - 2,262 \cdot 0,14769 &\leq M(y_{13}^*) \leq 6,434 + 2,262 \cdot 0,14769, \\ 6,1 &\leq M(y_{13}^*) \leq 6,769.\end{aligned}$$

Рассчитаем доверительный интервал для интервального прогноза:

$$\begin{aligned}\hat{\sigma}_{13(i)} &= 0,26233 \cdot \sqrt{1 + (1 \quad 1,2 \quad 9) \cdot \begin{pmatrix} 2,474519 & 0,189775 & -0,240327 \\ 0,189775 & 0,745468 & -0,080098 \\ -0,240327 & -0,080098 & 0,029253 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1,2 \\ 9 \end{pmatrix}} = \\ &= 0,26233 \cdot \sqrt{1 + 0,3169} = 0,26233 \cdot 1,1476 = 0,301\end{aligned}$$

Откуда имеем

$$\begin{aligned}6,434 - 2,262 \cdot 0,301 &\leq y_{13}^* \leq 6,434 + 2,262 \cdot 0,301, \\ 5,754 &\leq y_{13}^* \leq 7,115.\end{aligned}$$

Таким образом, прогнозные значения среднего товарооборота попадают в более узкий интервал, чем отдельные значения данного показателя.

### **Процедура многошагового регрессионного анализа**

Построение модели множественной регрессии может быть осуществлено поэтапно с использованием процедуры многошагового регрессионного анализа. В ее рамках имеется два направления:

- 1) по мере добавления в модель независимых переменных;
- 2) по мере исключения из модели многофакторной регрессии несущественно влияющих на зависимую переменную независимых показателей.

Первое направление разбивается на шаги:

- 1.1. Расчет коэффициентов парной корреляции  $r_{y/x_j}$ ;
- 1.2. Выбор среди рассчитанных коэффициентов  $r_{y/x_j}$  наибольшего (по абсолютной величине). Включение в модель соответствующего показателя;
- 1.3. Построение модели парной регрессии, оценка значимости ее параметров;
- 1.4. Последовательное дополнение модели независимыми переменными по мере уменьшения значений  $r_{y/x_j}$ . Построение моделей, оценка их значимости.

Процедура оканчивается тогда, когда в модель будут включены все значимые переменные.

Второе направление также разбивается на шаги:

- 1.1. Построение модели множественной регрессии с включением в нее всего набора независимых переменных;
- 1.2. Оценка значимости параметров модели по критерию Стьюдента. Исключение из модели наименее значимой переменной;
- 1.3. Пересчет параметров модели для оставшегося набора независимых переменных. Оценка значимости параметров и т.д.

Процедура оканчивается тогда, когда в модели останутся только существенно влияющие на  $y$  переменные.