Тема 2. Множественная линейная регрессия

Цель и задачи

Цель контента темы 2 — дать представление о модели множественной линейной регрессии.

Задачи контента темы 2:

- Ввести понятие множественной линейной регрессии, дать спецификацию этой модели;
- Познакомить с методом наименьших квадратов (МНК) применительно к множественной линейной регрессии, вывести оценки параметров регрессии по МНК, дать им экономическую интерпретацию;
- Сформулировать основные предположения регрессионного анализа и статистические свойства оценок;
- Обсудить критерии качества множественной регрессии, сформулировать критерии проверки статистической значимости (оценок по отдельности и регрессии в целом), построить доверительные интервалы и прогнозы с помощью множественной линейной регрессии;
- Дать представление о проблеме мультиколлинеарности и регрессионных моделях с переменной структурой.

Оглавление.

§ 2.1. Спецификация модели множественной линейной регрессии

Экономические зависимости, как правило, содержат большое число одновременно и совокупно действующих факторов. В связи с этим часто возникает задача исследования зависимости одной пере-

менной от нескольких объясняющих переменных (факторов). Эта проблема решается при помощи множественного регрессионного анализа. Примерами подобных зависимостей являются следующие:

- показатель y — расходы фирмы за месяц, фактор x_1 — объем выпущенной продукции за месяц, x_2 — стоимость электроэнергии в этом месяце.
- показатель y — спрос на товар, факторы: x_1 — цена единицы товара, x_2, x_3 — цены товаров-заменителей.

Итак, при построении множественной регрессии, в отличие от случая парной регрессии, предполагают, что имеется несколько объясняющих факторов. Пусть y — изучаемый эконометрический показатель; $x_1, x_2, \mathbf{K}, x_k$ — объясняющие факторы.

Эконометрическая модель, приводящая к множественной регрессии, имеет следующий вид:

$$y = f(x_1, \mathbf{K}, x_k) + e \qquad (2.1.1)$$

где $f(x_1, x_2, \mathbf{K}, x_k)$ — неизвестная функциональная зависимость (*теоретическая регрессия*); e — случайное слагаемое (*возмущение*), представляющее собой совокупное действие не включенных в модель факторов, ошибки измерения.

Основная задача эконометрического моделирования — построение эмпирической множественной регрессии $\hat{f}(x_1, x_2, \mathbf{K}, x_k)$, являющейся *оценкой* теоретической регрессии (функции $f(x_1, \mathbf{K}, x_k)$):

$$\hat{y} = \hat{f}(x_1, \mathbf{K}, x_k), \qquad (2.1.2)$$

здесь $\hat{f}(x_1, \mathbf{K}, x_k)$ — *эмпирическая (выборочная) регрессия*, описывающая усредненную зависимость между изучаемым показателем и факторами. После построения выборочной множественной регрессии, так же как и в случае парной регрессии, обычно производится верификация модели — проверка статистической значимости и адекватности построенной парной регрессии имеющимся эмпирическим данным.

Экспериментальная основа построения множественной эмпирической регрессии — многомерная выборка $(x_{11}, x_{12}, \mathbf{K}, x_{1k}, y_1), \ldots, (x_{n1}, x_{n2}, \mathbf{K}, x_{nk}, y_n)$, где n — объем выборки (объем массива экспериментальных данных), k — число факторов, x_{ij} — i-е наблюдение объясняющей переменной x_j ($i = \overline{1,n}$, $j = \overline{1,k}$).

Аналогично случаю парной регрессии одна из важных задач спецификации модели множественной регрессии заключается в выборе функциональной зависимости.

Основные методы выбора функциональной зависимости f те же, что и в случае парной регрессии. Однако задача выбора функциональной зависимости для множественной регрессии оказывается бо-

лее сложной, чем в случае парной регрессии. Причина заключается в многомерной природе объясняющих переменных.

Так, *геометрический* метод, основанный на построении поля корреляции, менее нагляден, чем в случае парной регрессии, а в ряде случаев просто неприменим, что связано с трудностями графического изображения многомерных данных.

Эмпирический метод, как и в случае парной регрессии, основан на *методе наименьших квадратов.* Выбирается некоторая параметрическая функциональная зависимость $f(x_1, \mathbf{K}, x_k)$. Для построения по выборке оценки $\hat{f}(x)$ этой зависимости чаще всего используется *метод наименьших квадратов (МНК)*.

Согласно методу наименьших квадратов значения параметров функции $\hat{f}(x_1, \mathbf{K}, x_k)$, которая является оценкой функции $f(x_1, \mathbf{K}, x_k)$ по выборке (будем обозначать их через a, b_1, $b_2, \ldots, b_k$), выбираются таким образом, чтобы сумма квадратов отклонений выборочных значений y_i от значений $\hat{f}(x_{i1}, \mathbf{K}, x_{ik})$ была минимальной

$$\sum_{i=1}^{n} \left(y_i - \hat{f}(x_{i1}, \mathbf{K}, x_{ik}) \right)^2 \xrightarrow[a, b_1, \mathbf{K}, b_k]{} \min, \qquad (2.1.3)$$

минимум ищется по параметрам a, b_1, $b_2, \ldots, b_k$ которые входят в зависимость $\hat{f}(x_1, \mathbf{K}, x_k)$.

Найденные значения параметров, которые минимизируют указанную сумму квадратов разностей, называются *оценками неизвестных параметров регрессии по методу наименьших квадратов (оценками МНК)*. Выборочная регрессия $\hat{y} = \hat{f}(x_1, \mathbf{K}, x_k)$ (или $\hat{y}_i = \hat{f}(x_{i1}, \mathbf{K}, x_{ik}), i = 1, \mathbf{K}, n$), в которую подставлены найденные значения, уже не содержит неизвестных параметров и является оценкой теоретической регрессии. Именно эту зависимость $\hat{f}(x_1, \mathbf{K}, x_k)$ рассматривают как эмпирическую усредненную зависимость изучаемого показателя от объясняющих факторов.

После нахождения эмпирического уравнения регрессии вычисляются значения $\hat{y}_i = \hat{f}(x_{i1}, \mathbf{K}, x_{ik}), i = 1, \mathbf{K}, n$ и *остатки* $e_i = y_i - \hat{y}_i$, $i = \overline{1, n}$. По величине остаточной суммы квадратов $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, как и в случае парной регрессии, можно судить о качестве соответствия эмпирической функции $\hat{f}(x_1, \mathbf{K}, x_k)$ имеющимся в наличии статистическим наблюдениям. Перебирая разные функциональные зависимости и, каждый раз, действуя подобным образом можно практически подобрать наиболее подходящую функцию для описания имеющихся данных.

Аналогично случаю парной регрессии если y — расходы фирмы, $x_1, \mathbf{K}, x_k$ — объемы выпущенной продукции за месяц, то применяя аналитический метод, можно получить следующую модель:

$$y = a + b_1 x_1 + \mathbf{K} + b_k x_k + e$$

где a — условно-постоянные расходы, $b_1 x_1 + b_2 x_2 + \mathbf{K} + b_k x_k$ — условно-переменные расходы.

В практике эконометрического анализа часто используют *линейную множественную регрессию*. В модели множественной линейной регрессии зависимость 2.1.1 между переменными представляется в виде

$$y = a + b_1 x_1 + \mathbf{K} + b_k x_k + e\,, \qquad (2.1.4)$$

т.е. теоретическая регрессия имеет вид $f(x_1, \mathbf{K}, x_k) = a + b_1 x_1 + \mathbf{K} + b_k x_k$.

На основе выборочных наблюдений оценка теоретической регрессии — выборочная (эмпирическая) регрессия $\hat{y}$ строится в виде:

$$\hat{y} = a + b_1 x_1 + \mathbf{K} + b_k x_k\,, \qquad (2.1.5)$$

где $a, b_1, b_2, \ldots, b_k$ являются оценками параметров $a, b_1, b_2, \ldots, b_k$ теоретической регрессии.

Выбор объясняющих переменных $x_1, \mathbf{K}, x_k$ является основным моментом *спецификации модели* множественной линейной регрессии (иногда выбор объясняющих переменных и называют спецификацией модели). Иногда, исходя из экономической теории, предыдущих исследований, заранее известен характер зависимости, определен список объясняющих переменных. В этом случае задача состоит лишь в оценивании неизвестных параметров зависимости.

Но на практике чаще встречается случай, когда имеется достаточное число наблюдений различных показателей (независимых, объясняющих) переменных, но нет априорной модели, позволяющей однозначно определить состав объясняющих переменных. В этом случае используют различные эмпирические процедуры *пошагового отбора факторов.* Суть этой процедуры в том, что сначала рассматривается только одна объясняющая переменная, имеющая с зависимой переменной y наиболее тесную корреляционную связь. На следующем шаге в регрессионную модель включается новая объясняющая переменная, таким образом, чтобы улучшить «качество» модели (для проверки используется скорректированный коэффициент детерминации 2.4.5, коэффициенты частной корреляции 2.4.7, значение F-статистики 2.5.3 и т.д.). Следует только иметь в виду то, что подобные пошаговые процедуры не гарантируют получение наилучшего набора факторов.

§ 2.2. Оценка параметров. Метод наименьших квадратов. Экономическая интерпретация.

Рассматривается модель множественной линейной регрессии

$$y_i = a + b_1 x_{i1} + \mathbf{K} + b_k x_{ik} + e_i, \quad i = 1, \mathbf{K}, n$$

На основе эмпирических наблюдений построим оценку теоретической регрессии — найдем выборочное уравнение регрессии

$$\hat{y}_i = a + b_1 x_{i1} + \mathbf{K} + b_k x_{ik}, \quad i = \overline{1, n}.$$

Оценки $a, b_1, b_2, \mathbf{K}, b_k$ параметров a , b_1, b_2,…, b_k определяются по *методу наименьших квадратов* из соотношения:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (a + b_1 x_{i1} + \mathbf{K} + b_k x_{ik}) \right)^2 \xrightarrow[a, b_1, b_2, \mathbf{K}, b_k]{} \min, \qquad (2.2.1)$$

т.е. значения $a, b_1, b_2, \mathbf{K}, b_k$ выбираются таким образом, чтобы минимизировать сумму квадратов отклонений выборочных (эмпирических) значений показателя y_i от расчетных $\hat{y}_i$.

Введем обозначения

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \mathbf{K} & x_{1k} \\ 1 & x_{21} & x_{22} & \mathbf{K} & x_{2k} \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ 1 & x_{n1} & x_{n2} & \mathbf{K} & x_{nk} \end{pmatrix}, \quad \mathbf{X}^T = \begin{pmatrix} 1 & 1 & \mathbf{K} & 1 \\ x_{11} & x_{21} & \mathbf{K} & x_{n1} \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ x_{1k} & x_{2k} & \mathbf{K} & x_{nk} \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} a \\ b_1 \\ \text{M} \\ b_k \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \text{M} \\ y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} e_1 \\ e_2 \\ \text{M} \\ e_n \end{pmatrix}, \qquad (2.2.2)$$

здесь и далее знаком ^{T} обозначается операция транспонирования матрицы.

Используя эти обозначения, модель множественной регрессии 2.1.4 может быть записана в матричной форме:

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}. \qquad (2.2.3)$$

Применяя тот же прием, что и в случае парной регрессии (вычисляем частные производные по неизвестным параметрам и приравниваем их к нулю) приходим к системе так называемых *нормальных уравнений* метода наименьших квадратов. В матричной форме система нормальных уравнений записывается следующим образом:

$$\mathbf{X}^T \mathbf{X}\mathbf{b} = \mathbf{X}^T \mathbf{Y}, \qquad (2.2.4)$$

Для того чтобы найти вектор оценок $\mathbf{b}$, т.е. система (2.2.2) имела решение, необходимо, чтобы матрица $\mathbf{X}^T\mathbf{X}$ была *неособенной (невырожденной)*, т.е. $|\mathbf{X}^T\mathbf{X}| \neq 0$. Для этого векторы значений объясняющих переменных (т.е. столбцы матрицы $\mathbf{X}$) должны быть линейно незави-

симы, т.е. ранг матрицы должен быть равен числу ее столбцов $r(\mathbf{X}) = k+1$.

Кроме того, должно быть выполнено условие $n > k+1$, другими словами число имеющихся наблюдений каждой из объясняющих переменных должно, по крайней мере, на единицу превосходить число объясняющих переменных. На практике часто считается, что при оценивании параметров множественной линейной регрессии для обеспечения статистической надежности требуется, чтобы число наблюдений, по крайней мере, в 3 раза превосходило число оцениваемых параметров

Итак, если матрица $\mathbf{X}^T\mathbf{X}$ невырождена, то решением системы уравнений 2.2.4 является вектор $\mathbf{b} = (a, b_1, \mathbf{K}, b_k)^T$:

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \qquad (2.2.5)$$

Найденное решение — вектор $\mathbf{b} = (a, b_1, \mathbf{K}, b_k)^T$, называется *оценкой наименьших квадратов неизвестных параметров* $a, b_1, \mathbf{K}, b_k$.

Таким образом, *эмпирическая (выборочная) множественная линейная регрессия* имеет вид:

$$\hat{y} = a + b_1x_1 + b_2x_2 + \mathbf{K} + b_kx_k, \qquad (2.2.6)$$

где коэффициенты $a, b_1, b_2, \mathbf{K}, b_k$ определяются по формуле 2.2.5.

Экономическая интерпретация коэффициентов $b_1, b_2, \mathbf{K}, b_k$ регрессии при объясняющих факторах та же, что и в случае парной регрессии. Коэффициент b_i показывает, на сколько единиц изменится в среднем показатель y, если фактор x_i, соответствующий этому коэффициенту, увеличится на одну единицу, в то время как остальные факторы останутся неизменными.

В матричной форме выборочное уравнение множественной линейной регрессии 2.2.6 можно записать в виде

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}, \qquad (2.2.7)$$

где $\hat{\mathbf{Y}} = (\hat{y}_1, \mathbf{K}, \hat{y}_n)^T$, $\mathbf{X}$ и $\mathbf{b}$ определяются в 2.2.2, вектор оценок $\mathbf{b}$ вычислен в соответствии с 2.2.5.

Значения показателя, вычисленные по выборочной регрессии для значений объясняющих факторов, содержащихся в выборке

$$\hat{y}_i = a + b_1x_{i1} + b_2x_{i2} + \mathbf{K} + b_kx_{ik},\ i = 1, \mathbf{K}, n \qquad (2.2.6)$$

или, в матричной форме,

$$\hat{y}_i = \mathbf{X}_i^T\mathbf{B},$$

где $\mathbf{X}_i^T = (1, x_{i1}, x_{i2}, \mathbf{K}, x_{ik})$ — вектор значений объясняющих переменных (совпадает с i-ой строкой матрицы $\mathbf{X}$ из 2.2.2).

Как и в случае парной линейной регрессии, особое значение для проверки статистической значимости парной линейной регрессии имеют *остатки* (разности между истинными значениями показателя и значениями, вычисленными по уравнению линейной регрессии):

$$e_i = y_i - \hat{y}_i, \quad i = 1, \mathbf{K}, n. \tag{2.2.8}$$

§ 2.3. Основные предположения регрессионного анализа. Теорема Гаусса-Маркова Статистические свойства оценок.

Основные предположения регрессионного анализа в случае множественной регрессии содержательно остаются теми же, что и в случае парной регрессии. В формулировках следует лишь учесть многомерный характер данных, так в многомерном случае аналогом дисперсии случайной компоненты является *ковариационная матрица* вектора возмущений:

$$\boldsymbol{\Sigma}_e = \begin{pmatrix} E(e_1^2) & E(e_1e_2) & \mathbf{K} & E(e_1e_n) \\ E(e_2e_1) & E(e_2^2) & \mathbf{K} & E(e_2e_n) \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ E(e_ne_1) & E(e_ne_2) & \mathbf{K} & E(e_n^2) \end{pmatrix} \tag{2.3.1}$$

В матричной форме основные предположения (условия Гаусса-Маркова), образующие первую группу предположений, могут быть записаны следующим образом:

Условие 2.3.1. Вектор возмущений $\boldsymbol{\varepsilon} = (e_1, \mathbf{K}, e_n)$ является случайным вектором.

Условие 2.3.2. Математическое ожидание возмущений равно нулю: $E(\boldsymbol{\varepsilon}) = \mathbf{0}_n$ (здесь $\mathbf{0}_n$ — нулевой вектор размера n).

Условия 2.3.3, 2.3.4. Дисперсия возмущений постоянна и возмущения некоррелированы: $\boldsymbol{\Sigma}_e = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = s^2\mathbf{I}_n$ (здесь $\mathbf{I}_n$ — единичная матрица n-го порядка).

Условие 2.3.5. Величины e_i взаимно независимы со значениями объясняющих переменных.

Условие 2.3.6. $r(\mathbf{X}) = k + 1$.

При выполнении предположений из первой группы основных предположений регрессионного анализа справедлива теорема:

Теорема 2.3.1. (Гаусса-Маркова) Если выполнены предпосылки 2.3.1–2.3.6, то оценки метода наименьших квадратов $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ регрессионной модели 2.2.3 являются наиболее эффективными (в смысле минимума дисперсии линейных комбинаций оценок параметров) в классе линейных несмещенных оценок.

Вторую группу образует следующее предположение.

Условия 2.3.7. $\boldsymbol{\varepsilon}$ — нормально распределенный случайный вектор.

После построения уравнения выборочной регрессии, наблюдаемые значения y_i можно представить в виде

$$y_i = \hat{y}_i + e_i, \quad i = \overline{1,n}, \tag{2.3.2}$$

где $\hat{y}_i$ определяются по формуле 2.2.6, в матричной форме это представление имеет вид

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e} = \mathbf{Xb} + \mathbf{e},$$

где $\mathbf{e} = (e_1, e_2, \mathbf{K}, e_n)^T$ — *вектор остатков.* Остатки e_i являются, в отличие от возмущений e_i, наблюдаемыми величинами, с помощью которых можно оценить воздействие неучтенных факторов и ошибок наблюдений.

Можно показать, что статистика (*выборочная остаточная дисперсия*)

$$S_{ост}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} = \frac{\mathbf{e}^T\mathbf{e}}{n-k-1} \tag{2.3.3}$$

является несмещенной оценкой дисперсии s^2.

При выполнении условий Гаусса-Маркова первой и второй групп (1.5.1–1.5.6) справедливы утверждения:

Утверждение 2.3.1. Статистика $\frac{a-a}{m_a}$ распределена по закону Стьюдента с $n-k-1$ степенями свободы, здесь

$$m_a = S_{ост}\sqrt{\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]_{11}}, \tag{2.3.4}$$

представляет собой *стандартную ошибку коэффициента* a, $\left[\left(X^T X\right)^{-1}\right]_{11}$ — первый элемент, стоящий на главной диагонали матрицы $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$.

Утверждение 2.3.2. Статистика $\frac{b_i - b_i}{m_{b_i}}$ распределена по закону Стьюдента с $n-k-1$ степенями свободы, здесь

$$m_{b_i} = S_{ост}\sqrt{\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]_{i+1\,i+1}}, \tag{2.3.5}$$

представляет собой *стандартную ошибку коэффициента* b_i, $\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]_{i+1\,i+1}$ — $i+1$-й элемент, стоящий на главной диагонали матрицы $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$.

§ 2.4. Показатели качества регрессии. Коэффициент детерминации. Коэффициенты парной и частной корреляции.

Как и в случае парной регрессионной модели (см. § 1.7) в случае множественной регрессии выборочная дисперсия наблюдений y_i может быть представлена в виде суммы:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2, \tag{2.4.1}$$

в которой первое слагаемое представляет собой часть, «объясненную» регрессионным уравнением (или обусловленную регрессией), а второе слагаемое — «необъясненную» часть, характеризующую влияние неучтенных факторов и т.п. Аналогично 1.7.2 справедливо равенство

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2.4.2}$$

части которой имеют точно такой же смысл, как и в случае парной регрессии.

Коэффициент детерминации (*множественный*), является мерой адекватности регрессионной модели и определяется так же, как и в случае парной регрессии:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{2.4.3}$$

В матричной форме формулы для вычисления коэффициента детерминации можно записать следующим образом:

$$R^2 = \frac{\mathbf{b}^T\mathbf{X}^T\mathbf{Y}^T - n\bar{y}^2}{\mathbf{Y}^T\mathbf{Y} - n\bar{y}^2} = 1 - \frac{\mathbf{e}^T\mathbf{e}}{\mathbf{Y}^T\mathbf{Y}}. \tag{2.4.4}$$

Точно так же, как в случае парной регрессии, коэффициент детерминации характеризует долю вариации зависимой переменной, обусловленную регрессией. Чем ближе R^2 к единице, тем лучше регрессия описывает зависимость между зависимой и объясняющими переменными.

В случае множественной регрессии значение R^2 автоматически увеличивается при добавлении новых объясняющих переменных, хотя это не обязательно свидетельствует об улучшении качества регрессионной модели. Поэтому часто используют *скорректированный (исправленный) коэффициент детерминации*

$$\tilde{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2). \tag{2.4.5}$$

Из формулы 2.4.5 следуют свойства скорректированного коэффициента детерминации:

Свойство 1. $\tilde{R}^2 < R^2$ при $k > 1$ и $\tilde{R}^2 = R^2$ только если $R^2 = 1$.

Свойство 2. С ростом числа объясняющих переменных k скорректированный коэффициент детерминации $\tilde{R}^2$ растет медленнее, чем (обычный) коэффициент детерминации R^2.

Свойство 3. При добавлении новых объясняющих переменных, не оказывающих существенного влияния на зависимую переменную, $\tilde{R}^2$ может и уменьшаться.

Однако не следует абсолютизировать важность коэффициентов детерминации. Существует достаточно примеров неправильно специфицированных моделей, имеющих высокие коэффициенты детерминации. Поэтому коэффициент детерминации в настоящее время рассматривается лишь как один из ряда показателей, который нужно проанализировать, чтобы уточнить модель.

Оценка качества соответствия выборочного равнения регрессии наблюдаемым данным может производиться и с помощью *средней ошибки аппроксимации* $\bar{A}$ регрессии по формуле 1.7.4.

В многомерном случае важную роль играют *частные коэффициенты корреляции*. Дело в том, что коэффициенты парной корреляции могут давать ложное представление о характере и силе взаимосвязи между двумя переменными, так как они не учитывают влияние других переменных. Например, между двумя переменными может быть высокий положительный коэффициент корреляции не потому, что одна из них стимулирует изменение другой, а оттого, что обе эти переменные изменяются в одном направлении под влиянием других переменных, как учтенных в модели, так и, возможно, неучтенных.

Явление ложной корреляции хорошо известно в статистической литературе. Для оценки «истинной» взаимозависимости используются *коэффициенты частной корреляции* «очищенные» от влияния других факторов.

В общем случае выборочный коэффициент частной корреляции межу переменными x_i и x_j $(i \neq j)$, очищенный от влияния остальных $k-2$ объясняющих переменных, обозначается

$$r_{x_i x_j | x_1 \mathbf{K} x_{i-1} x_{x+1} \mathbf{K} x_{j-1} x_{j+1} \mathbf{K} x_k}.$$

Для вычисления коэффициентов частной корреляции между переменными в случае k объясняющих переменных составим корреляционную матрицу, состоящую из выборочных коэффициентов корреляции:

$$\mathbf{Q} = \begin{pmatrix} 1 & r_{x_1 x_2} & \mathbf{K} & r_{x_1 x_k} \\ r_{x_2 x_1} & 1 & \mathbf{K} & r_{x_2 x_k} \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ r_{x_k x_1} & r_{x_k x_2} & \mathbf{K} & 1 \end{pmatrix}, \qquad (2.4.6)$$

здесь $r_{x_i x_j}$ определяется по формулам 1.6.6, 1.7.8, причем $r_{x_i x_j} = r_{x_j x_i}$.

Выборочным коэффициентом частной корреляции (или просто — *частным коэффициентом корреляции*) между переменными x_i и x_j при фиксированных значениях остальных $k-2$ переменных называется выражение

$$r_{x_i x_j | x_1 \mathbf{K} x_{i-1} x_{x+1} \mathbf{K} x_{j-1} x_{j+1} \mathbf{K} x_k} = -\frac{A_{ij}}{\sqrt{A_{ii} A_{jj}}}, \qquad (2.4.7)$$

где через A_{ij} обозначены алгебраические дополнения элементов $r_{x_i x_j}$ матрицы выборочных коэффициентов корреляции $\mathbf{Q}$.

Значения коэффициентов частной корреляции, как и обычных выборочных коэффициентов парной корреляции, лежат в интервале $[-1,1]$. Можно сказать, что равенство нулю коэффициента частной корреляции означает отсутствие прямого (линейного) влияния одной переменной на другую.

При анализе модели множественной линейной регрессии часто необходимо вычислить коэффициент частной корреляции между зависимой переменной y и объясняющей переменной x_i, измеряющий влияние на y влияние только одного фактора x_i и «очищенный» от влияния остальных факторов. Рассмотрим расширенную матрицу, состоящую из выборочный парных коэффициентов корреляции:

$$\mathbf{\tilde{Q}} = \left(\begin{array}{c|cccc} 1 & r_{yx_1} & r_{yx_2} & \mathbf{K} & r_{yx_k} \\ \hline r_{yx_1} & 1 & r_{x_1x_2} & \mathbf{K} & r_{x_1x_k} \\ r_{yx_2} & r_{x_2x_1} & 1 & \mathbf{K} & r_{x_2x_k} \\ \mathbb{M} & \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ r_{yx_k} & r_{x_kx_1} & r_{x_kx_2} & \mathbf{K} & 1 \end{array}\right) = \left(\begin{array}{c|ccc} 1 & r_{yx_1} & r_{yx_2} & \mathbf{K} \; r_{yx_k} \\ \hline r_{yx_1} & & & \\ r_{yx_2} & & \mathbf{Q} & \\ \mathbb{M} & & & \\ r_{yx_k} & & & \end{array}\right). \qquad (2.4.8)$$

Выборочным коэффициентом частной корреляции между зависимой переменной y и объясняющей переменной x_j при фиксированных значениях остальных $k-2$ переменных называется выражение

$$r_{yx_i | x_1 \mathbf{K} x_{i-1} x_{x+1} \mathbf{K} x_k} = -\frac{A_{yi}}{\sqrt{A_{yy} A_{ii}}}, \qquad (2.4.9)$$

где A_{yi} — алгебраическое дополнение к элементу r_{yx_i} матрицы $\mathbf{\tilde{Q}}$, A_{yy} — алгебраическое дополнение к элементу $\tilde{q}_{11}$ (т.е. $A_{yy} = |\mathbf{Q}|$), A_{ii} — алгебраическое дополнение к элементу $\tilde{q}_{ii}$ (заметим, что $\tilde{q}_{ii}$ это единица, стоящая на пересечении i-ой строки и i-го столбца).

В случае двух переменных ($k=2$) коэффициенты частной корреляции между y и объясняющими переменными вычисляются по формуле:

$$r_{yx_1|x_2}=\frac{r_{yx_1}-r_{yx_2}r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}},\ r_{yx_2|x_1}=\frac{r_{yx_2}-r_{yx_1}r_{x_2x_1}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_2x_1}^2)}},\qquad(2.4.10)$$

где $r_{yx_1}, r_{yx_2}, r_{x_1x_2}$ «обычные» коэффициенты парной корреляции.

Между частными коэффициентами корреляции и коэффициентом детерминации существует тесная связь, которая заключается в равенстве:

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

Для того чтобы сравнить влияние на зависимую переменную различных объясняющих переменных, особенно когда эти переменные имеют различные единицы измерения, используют также *стандартизованные коэффициенты регрессии* b_i'

$$b_i'=b_i\frac{s_{x_i}}{s_y},\ i=\overline{1,k}\qquad(2.4.11)$$

и *коэффициенты эластичности* $Э_i$ (частные коэффициенты эластичности)

$$Э_i=b_i\frac{\overline{x}_i}{\overline{y}},\ i=\overline{1,k}.\qquad(2.4.12)$$

Стандартизованный коэффициент регрессии b_i' показывает, на сколько величин s_y изменится в среднем зависимая переменная y при увеличении только i-ой переменной на s_{x_i} (ср. с 1.7.7). а частный коэффициент эластичности $Э_i$ — на сколько процентов (от средней) изменится в среднем y при увеличении только переменной x_i на 1% и неизменных значениях остальных переменных.

§ 2.5. Проверка статистической значимости в множественной линейной регрессии.

Правило проверки статистической значимости оценок a и b_i основывается на статистических свойствах оценок МНК (§ 2.3) и проверке статистических гипотез $H_0: a=0, H_1: a\neq 0$ и $H_0: b_i=0, H_1: b_i\neq 0$. Невозможность отклонения какой-либо из гипотез означает статистическую незначимость соответствующего коэффициента и наоборот, отклонение какой-либо из гипотез означает, что соответствующий коэффициент статистически значим.

Проверка статистических гипотез осуществляется при некотором уровне значимости. В практических эконометрических исследованиях, как и в случае парной регрессии, наиболее часто используются 5% и 1% уровни значимости. Выбор того или иного уровня значимости определяется исследователем (см. § 1.8 о соотношении уровней значимости).

2.5.1. Правило проверки значимости коэффициента b_i:

Статистика $t_{b_i} = \dfrac{b_i}{m_{b_i}}$ при выполнении гипотезы $H_0 : b_i = 0$ распределена по закону Стьюдента с $n-k-1$ степенями свободы.

По таблице распределения Стьюдента с $n-k-1$ степенями свободы по заданному уровню значимости определяется значение $t_{табл}$ как критическая точка, соответствующая двусторонней области. Тогда:

1) Если $|t_{b_i}| \geq t_{табл}$, то гипотезу $H_0 : b_i = 0$ следует отклонить и, следовательно, признать коэффициент b_i статистически **значимым**;

2) Если $|t_{b_i}| < t_{табл}$, то гипотезу $H_0 : b_i = 0$ следует принять и, следовательно, признать коэффициент b_i статистически **незначимым**.

2.5.2. Правило проверки значимости коэффициента a:

Статистика $t_a = \dfrac{a}{m_a}$ при выполнении гипотезы $H_0 : a = 0$ распределена по закону Стьюдента с $n-k-1$ степенями свободы.

По таблице распределения Стьюдента с $n-k-1$ степенями свободы по заданному уровню значимости определяется значение $t_{табл}$ как критическая точка, соответствующая двусторонней области. Тогда:

1) Если $|t_a| \geq t_{табл}$, то гипотезу $H_0 : a = 0$ следует отклонить и, следовательно, признать коэффициент a статистически **значимым**;

2) Если $|t_a| < t_{табл}$, то гипотезу $H_0 : a = 0$ следует принять и, следовательно, признать коэффициент a статистически **незначимым**.

Статистическая значимость множественной регрессии в целом оценивается с помощью F критерия Фишера:

2.5.3. Правило проверки значимости линейной регрессии в целом (гипотезы $H_0 : b_1 = b_2 = \mathbf{K} = b_k = 0$) с использованием F статистики:

Если выполнены предположения регрессионного анализа, то при выполнении гипотезы $H_0 : b_1 = b_2 = \mathbf{K} = b_k = 0$ (что означает отсутствие взаимосвязи между факторами $x_1, x_2, \mathbf{K}, x_k$ и показателем y, а так же статистическую незначимость построенной множественной регрессии), то статистика $F = \dfrac{R^2}{1-R^2}\dfrac{n-k-1}{k}$ распределена по закону Фишера с числом степеней свободы числителя k и числом степеней свободы знаменателя $n-k-1$.

По таблице распределения Фишера-Снедекора при заданном уровне значимости определяется значение $F_{табл}$ как критическая точка

при числе степеней свободы числителя равном k и числе степеней свободы знаменателя равном $n-k-1$. Тогда:

1) Если $F \geq F_{табл}$, то гипотезу $H_0 : b_1 = b_2 = \mathbf{K} = b_k = 0$ следует отклонить и, следовательно, признать построенное уравнение линейной регрессии статистически **значимым**;

2) Если $F < F_{табл}$, то гипотезу $H_0 : b_1 = b_2 = \mathbf{K} = b_k = 0$ следует принять и, следовательно, признать построенное уравнение статистически **незначимым**.

Таким образом, принятие нулевой гипотезы равнозначно статистической незначимости коэффициента множественной детерминации R^2.

В отличие от случая парной регрессии, когда проверка значимости коэффициента b и проверка значимости уравнения в целом с помощью F критерия были равносильны, для множественной регрессии ситуация более сложная. Если объясняющие переменные достаточно сильно коррелируют, то t тест для каждой переменной может оказаться незначимым, в то время как F тест для уравнения в целом может быть значимым.

Что касается проверки статистической значимости коэффициентов частной корреляции $r_{x_i x_j | x_1 \mathbf{K} x_{i-1} x_{x+1} \mathbf{K} x_{j-1} x_{j+1} \mathbf{K} x_k}$, то ряд авторов указывает на то, что этот коэффициент, рассчитанный по выборке объема n, имеет такое же распределение, как и выборочной коэффициент корреляции $r_{x_i x_j}$, вычисленный по $n-k+2$ наблюдениям. Поэтому значимость коэффициента частной корреляции оценивают так же, как и «обычного» коэффициента корреляции, полагая количество наблюдений равным $n-k+2$.

§ 2.6. Доверительные интервалы

2.6.1. Доверительные интервалы для параметров регрессии

Так же, как и в случае парной регрессии (см. § 1.9), учитывая статистические свойства оценок МНК, можно построить доверительные интервалы для параметров a, $b_1, \mathbf{K}, b_k$ с заданным уровнем доверия, в качестве которого на практике обычно выбирают вероятность 0,95 (соответствующую уровню значимости 5%).

По таблицам распределения Стьюдента для заданного уровня значимости определяется критическое значение $t_{табл}$, соответствующее $n-k-1$ степеням свободы, тогда

$$(a - m_a t_{табл} ;\ a + m_a t_{табл}) \qquad (2.6.1)$$

есть доверительный интервал для a с заданным уровнем доверия, здесь m_a — стандартная ошибка коэффициента a (см. 2.3.4).

Аналогично, для коэффициента b_i:

$$(b_i - m_{b_i} t_{табл} ;\ b_i + m_{b_i} t_{табл}) \qquad (2.6.2)$$

есть доверительный интервал для b_i с заданным уровнем доверия, здесь m_{b_i} — стандартная ошибка коэффициента b_i (см. 2.3.5).

2.6.2. Доверительный интервал прогноза

Пусть $\mathbf{x}_p = (1, x_{1p}, x_{2p}, \mathbf{K}, x_{kp})$ — вектор, составленный из значений объясняющих переменных, для которых вычисляется прогноз зависимой переменной y.

Нетрудно вычислить точечный прогноз по уравнению множественной регрессии:

$$\hat{y}_p = a + b_1 x_{1p} + b_2 x_{2p} + \mathbf{K} + b_k x_{kp} = \mathbf{x}_p \mathbf{b}. \qquad (2.6.3)$$

Стандартная ошибка прогноза индивидуального значения зависимой переменной вычисляется по формуле

$$m_{\hat{y}_p} = S_{ост} \sqrt{1 + \mathbf{x}_p (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{x}_p^T}. \qquad (2.6.4)$$

Аналогично формуле 1.9.4 доверительный интервал прогноза индивидуального значения вычисляется по формуле

$$\hat{y}_p - m_{\hat{y}_p} t_{табл} < y_p < \hat{y}_p + m_{\hat{y}_p} t_{табл}. \qquad (2.6.5)$$

§ 2.7. Мультиколлинеарность

В практических исследованиях нередко встречается ситуация, когда t-статистики большинства оценок малы, что свидетельствует о незначимости соответствующих объясняющих переменных, т.е. построенная выборочная регрессия является «плохой»:. Но, в то же время, уравнение регрессии в целом может быть статистически значимо, т.е. значение F-статистики может быть достаточно большим. Одной из возможных причин этого является наличие высокой корреляции между объясняющими переменными — т.н. мультиколлинеарность.

Под *мультиколлинеарностью* понимается высокая взаимная коррелированность объясняющих переменных. Говорят, что мультиколлинеарность может проявляться в следующих формах:

1. *Функциональная (полная, совершенная) форма* — в этом случае, по крайней мере, между двумя объясняющими переменными существует линейная зависимость (т.е. один из столбцов матрицы $\mathbf{X}$ является линейной комбинацией остальных столбцов). В этом случае определитель матрицы $\mathbf{X}^T\mathbf{X}$ равен нулю, поэтому невозможно найти решение системы нормальных уравнений 2.2.3 и получить оценки параметров регрессии. Следует заметить, что на практике полная коллинеарность встречается достаточно редко.

2. *Стохастическая (неполная, несовершенная) форма* — случай, когда хотя бы между двумя объясняющими переменными имеется достаточно высокая степень корреляции. Определитель мат-

рицы $\mathbf{X}^T\mathbf{X}$ при этом хоть и отличен от нуля, но очень мал, т.е. матрица близка к вырожденной. Такой случай гораздо чаще встречается на практике и именно его обычно имеют в виду, говоря о *мультколлинеарности*.

В случае мультиколлинеарности оценки МНК формально существуют, но обладают «плохими» свойствами, так как вектор оценок $\mathbf{b}$ и дисперсии его компонент обратно пропорциональны величине определителя $|\mathbf{X}^T\mathbf{X}|$. В результате стандартные ошибки коэффициентов регрессии получаются достаточно большими и оценка их значимости по t-критерию может не иметь смысла. Несмотря на то, что свойства несмещенности и эффективности оценок остаются в силе, мультиколлинеарность в любом случае затрудняет разделение влияния объясняющих факторов на поведение зависимой переменной и делает оценки коэффициентов регрессии ненадежными.

Итак, если коротко сформулировать основные последствия мультиколлинеарности, то можно выделить следующие:

1. Большие дисперсии (стандартные ошибки) оценок. Это затрудняет нахождение истинных значений определяемых величин и расширяет интервальные оценки, ухудшая их точность.

2. Уменьшаются t-статистики коэффициентов, что может привести к неоправданному выводу о существенности влияния соответствующей объясняющей переменной на зависимую переменную.

3. Оценки коэффициентов по МНК и их стандартные ошибки становятся очень чувствительными к небольшим изменениям данных, т.е. они становятся неустойчивыми.

4. Затрудняется определение вклада каждой из объясняющей переменных в объясняемую уравнением регрессии дисперсию зависимой переменной.

5. Возможно получение неверного знака у коэффициента регрессии.

Мультиколлинеарность может возникать в силу различных причин. Например, несколько переменных могут иметь общую тенденцию изменения во времени. Часто выделяют несколько наиболее характерных *признаков*, по которым может быть установлено наличие мультиколлинеарности. К их числу обычно относят следующие:

Признак 1. Небольшое изменение исходных данных (например, добавление новых наблюдений) приводит к существенному изменению оценок коэффициентов регрессии.

Признак 2. Оценки регрессии имеют большие стандартные ошибки, малую значимость (статистически незначимы), в то время как в целом модель является значимой, т.е. значения коэффициента детерминации R^2 и F-статистики достаточно велики.

Признак 3. Высокие коэффициенты (частной) корреляции. На практике обычно анализируют корреляционную матрицу между объяс-

няющими переменными. Если существуют пары переменных, имеющие высокие коэффициенты корреляции (обычно больше 0,8), то говорят о мультиколлинеарности. Наличие высокого (обычно больше 0,6) множественного коэффициента детерминации между одной объясняющей переменной и некоторой группой других переменных также свидетельствует о мультиколлинеарности.

Признак 4. Оценки коэффициентов имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения.

К сожалению, *не существует абсолютно точных количественных критериев*, позволяющих установить наличие или отсутствие мультиколлинеарности. О наличии мультиколлинеарности также может свидетельствовать близость определителя $|\mathbf{X}^T\mathbf{X}|$ к нулю.

Нет однозначного ответа и на вопрос что же делать в случае, если по все признакам мультиколлинеарность имеется. Некоторые авторы полагают, что ответ на этот вопрос зависит от целей эконометрического исследования и в ряде случаев мультиколлинеарность не является таким уж серьезным злом, чтобы прилагать серьезные усилия по ее выявлению и устранению. Если основная задача модели — прогноз будущих значений зависимой переменной, то при достаточно большом коэффициенте детерминации R^2 (не меньше 0,9) наличие мультиколлинеарности зачастую не сказывается на прогнозных качествах модели. Если же целью исследования является определение степени влияния каждой из объясняющих переменных на зависимую переменную, то наличие мультиколлинеарности, приводящее к увеличению стандартных ошибок, скорее всего, исказит истинные зависимости между переменными. В этой ситуации мультиколлинеарность представляется серьезной проблемой.

Несмотря на то, что единого метода устранения (или уменьшения) мультиколлинеарности, пригодного для любого случая, не существует можно предложить несколько подходов:

1. *Исключение переменных из модели.* Например, из двух объясняющих переменных, имеющих высокий коэффициент корреляции (больше 0,8) одну переменную обычно исключают из модели. Для того чтобы избежать ошибок спецификации (т.е. не исказить содержательный смысл модели) в первую очередь выбирают исключаемую переменную исходя из экономических (теоретических) соображений. Если ни одной из переменных нельзя отдать предпочтение, то оставляют ту переменную, которая имеет наибольший коэффициент (частной) корреляции с зависимой переменной. В прикладных эконометрических моделях желательно не исключать объясняющие переменные до тех пор, пока коллинеарность не станет серьезной проблемой.

2. *Переход к новым переменным.* От исходных объясняющих переменных, связанных между собой достаточно тесно корреляционной

зависимостью, можно перейти к новым переменным, представляющим собой линейные комбинации исходных. Новые переменные должны быть некоррелированными (или, по крайней мере, слабокоррелированными). Новые переменные могут быть предложены исходя из экономической теории или из формальных математических соображений. Так, например, в качестве таких переменных иногда берут *главные компоненты* вектора исходных объясняющих переменных.

3. *Получение новых эмпирических данных.* Поскольку мультиколлинеарность непосредственно зависит от выборки, то получение новой выборки или дополнительных наблюдений (увеличение объема выборки) может ослабить эту проблему.

§ 2.8. Фиктивные переменные. Регрессионные модели с переменной структурой

Как правило, независимые переменные в регрессионных моделях могут «непрерывно» изменяться в некоторой области. Но теория не накладывает никаких ограничений на характер объясняющих переменных, в частности некоторые переменные могут принимать лишь дискретные значения.

На практике довольно часто возникает необходимость исследовать зависимость показателя от *качественного* признака, имеющего несколько значений (пол, наличие образования, вкусы потребителя, время года и т.п.). Так, например, при исследовании зависимости зарплаты от различных признаков требуется принять во внимание наличие у работника высшего или специального образования и т.д. В принципе можно оценивать соответствующие зависимости по отдельности внутри каждой категории, а затем изучат различия между ними, но введение дискретных переменных позволяет оценивать одно уравнение сразу по все категориям.

Качественные признаки обычно существенно влияют на структуру линейных связей между переменными, в этом случае говорят о *регрессионных уравнениях с переменной структурой.* Влияние качественного признака в регрессионных моделях отражается в виде *фиктивной* (*искусственной*) переменной, которая отражает различные состояния качественного фактора (в простейшем случае — противоположные состояния). Например, «фактор действует — фактор не действует», «курс валюты фиксированный — курс валюты плавающий», «сезон летний — сезон зимний», «есть высшее образование — нет высшего образования» и т. д. — в этом случае фиктивная переменная выражается в двоичной (бинарной) форме:

$$d = \begin{cases} 0, \text{ фактор не действует}, \\ 1, \text{ фактор действует}. \end{cases}.$$

В англоязычной литературе по эконометрике подобные переменные называются *dummy variables*, что обычно переводится как

«*фиктивные переменные*». Тем не менее, переменная d такая же «равноправная» переменная, как и любая другая «обычная» переменная. Ее «фиктивность» состоит только в том, что она количественным образом описывает качественный признак. Все статистические процедуры регрессионного анализа для модели с фиктивными переменными (оценка параметров регрессии, проверка значимости и т.д.) проводятся точно так же, как и в случае «обычных» количественных объясняющих переменных.

Например, пусть y — размер заработной платы, которая зависит от стажа работы x и наличия у работника высшего образования d, т.е. мы рассматриваем модель:

$$\hat{y}_i = a + b_1 x_i + g\, d_i, \qquad (2.8.1)$$

где $d_i = 0$ если i работник не имеет высшего образования, $d_i = 1$, если i работник имеет высшее образование. Таким образом, мы считаем, что средняя зарплата есть $a + b_1 x$ при отсутствии высшего образования и $a + b_1 x + g$ при его наличии. Величина g представляет собой среднее изменение зарплаты при переходе из одной категории в другую при неизменных значениях стажа работы. Тестируя гипотезу $H_0 : g = 0$, мы проверяем предположение о существенном влиянии фактора «наличие высшего образования» на размер заработной платы работника.

Коэффициент g в модели (2.8.1) иногда называется *дифференциальным коэффициентом свободного члена*, т. к. он показывает, на какую величину отличается свободный член модели при значении фиктивной переменной, равном единице, от свободного члена модели при нулевом (базовом) значении фиктивной переменной.

Качественные различия можно описывать с помощью переменных, принимающих любое количество произвольных значений, но в эконометрической практике почти всегда используют двоичные переменные, принимающие значение 0 или 1, поскольку в этом случае интерпретация уравнения выглядит наиболее просто. Если признак может принимать p различных значений (градаций) то можно было бы ввести фиктивную переменную, принимающую такое же количество значений. Однако на практике обычно так не поступают из-за трудности интерпретации соответствующих коэффициентов регрессии — вместо этого вводят $p-1$ бинарную переменную.

Так, в примере с зарплатой, если предположить, что образование может быть начальным, средним или высшим, то для учета фактора образования в регрессионную модель 2.8.1 вводят две бинарные переменные d_{i1} и d_{i2}:

$$\hat{y}_i = a + b_1 x_i + g_1 d_{i1} + g_2 d_{i2}$$

где

$$d_{i1} = \begin{cases} 1, & \text{если } i \text{ работник имеет высшее образование,} \\ 0, & \text{в остальных случаях,} \end{cases}$$

$$d_{i2} = \begin{cases} 1, & \text{если } i \text{ работник имеет среднее образование,} \\ 0, & \text{в остальных случаях.} \end{cases}$$

Очевидно, что третьей бинарной переменной не требуется: если i-й работник имеет начальное образование, то это соответствует паре значений $d_{i1} = 0$, $d_{i2} = 0$.

Другим типичным примером является исследование сезонных колебаний. Если есть основания считать, что объем потребления y зависит от времени года, то для выявления сезонности можно ввести три бинарные переменные d_1, d_2, d_3:

$$d_{i1} = \begin{cases} 1, & \text{если месяц } i \text{ является зимним,} \\ 0, & \text{в остальных случаях,} \end{cases}$$

$$d_{i2} = \begin{cases} 1, & \text{если месяц } i \text{ является весенним,} \\ 0, & \text{в остальных случаях,} \end{cases} \qquad (2.8.2)$$

$$d_{i3} = \begin{cases} 1, & \text{если месяц } i \text{ является летним,} \\ 0, & \text{в остальных случаях.} \end{cases}$$

и оценивать зависимость

$$y_i = a + b_1 d_{i1} + b_2 d_{i2} + b_3 d_{i3} + e_i. \qquad (2.8.3)$$

Заметим, что вводить четвертую переменную, относящуюся к осени (в примере с зарплатой — третью переменную, соответствующую начальному образованию) *нельзя*, иначе выполнялось бы тождество $d_{i1} + d_{i2} + d_{i3} + d_{i4} = 1, i = \overline{1,n}$, что означало бы линейную зависимость факторов (явление мультиколлинеарности) и, как следствие, невозможность получения оценок МНК (см. § 2.7). Такая ситуация, когда сумма фиктивных переменных тождественно равна константе, называется *ловушкой фиктивных переменных* («dummy trap»). Поэтому, чтобы избежать такой ловушки, следуют правилу:

Правило введения фиктивных переменных. Если качественный признак имеет p альтернативных значений (градаций), то число вводимых бинарных фиктивных переменных должно быть равно $p-1$.

Фиктивные переменные, несмотря на внешнюю простоту, являются весьма мощным инструментов при исследовании влияния качественных признаков. В рассмотренных выше примерах влияние качественного признака сказывалось только на *свободном члене* уравнения регрессии. С помощью фиктивных переменных можно учесть и влияние качественного признака на *параметры при переменных* регрессионной модели.

В примере 2.8.3 с сезонными различиями можно ввести независимую переменную x — доход, используемый на потребление. В модели

$$y_i = a + b_1 x_i + e_i \qquad (2.8.4)$$

коэффициент b_1 называется «склонностью к потреблению». Естественно исследовать вопрос влияния сезона и на объем и на склонность к потреблению. Для этого можно рассмотреть модель

$$y_i = a + b_1 d_{i1} + b_2 d_{i2} + b_3 d_{i3} + (b_4 d_{i1}) x_i + (b_5 d_{i2}) x_i + (b_6 d_{i3}) x_i + b_7 x_i + e_i,$$

где склонность к потреблению зимой, весной, летом и осенью есть $b_4 + b_7$, $b_5 + b_7$, $b_6 + b_7$ и b_7 соответственно. Используя эту модель можно строить оценки и проверять гипотезы о влиянии сезонных факторов на склонность к потреблению.

Фиктивные переменные позволяют строить и оценивать так называемые *кусочно-линейные модели*, широко используемые для исследования влияния структурных изменений (например, новых налоговых правил, реформ и т.п.). Зависимость в этом случае может иметь вид

$$y_i = a + b_1 x_i + g_1 d_i + g_2 d_i x_i + e_i, \qquad (2.8.5)$$

где

$$d_i = \begin{cases} 1, & \text{до структурного изменения условий,} \\ 0, & \text{после структурного изменения условий.} \end{cases}$$

Коэффициенты g_1 и g_2 в уравнении (2.8.5) называются *дифференциальным свободным членом* и *дифференциальным угловым коэффициентом* соответственно. Фиктивная переменная d_i используется как в *аддитивном виде* ($g_1 d_i$), так и в *мультипликативном* ($g_2 d_i x_i$), что позволяет фактически разбивать рассматриваемую зависимость на две части, связанные с изменениями некоторого рассматриваемого в модели качественного фактора. Тестируя гипотезу $H_0 : g_2 = 0$, мы проверяем предположение о том, что фактически структурного изменения не произошло.

Выводы

- Модель множественной линейной регрессии является наиболее распространенным (и простым) уравнением зависимости в случае, когда на рассматриваемый показатель оказывает влияние несколько факторов. Метод наименьших квадратов дает наилучшие (в определенном смысле) оценки параметров регрессии. Решающее значение для правильного и обоснованного применения регрессионного анализа в эконометрических исследованиях имеет выполнение условий Гаусса–Маркова.

- Необходимым элементом эконометрического анализа является проверка статистической значимости полученных оценок коэффициентов, а также всего уравнения регрессии в целом. В качестве показателя качества регрессии может использоваться множественный коэффициент детерминации.
- Особое значение в случае множественной регрессии имеют коэффициенты частной корреляции, показывающие силу влияния фактора «очищенные» от влияния остальных факторов.
- При использовании множественной линейной регрессии для построения прогнозов необходимо учитывать доверительные интервалы прогноза и параметров регрессии.
- В случае множественной регрессии необходимо учитывать возможную мультиколлинеарность объясняющих переменных.
- Использование фиктивных переменных позволяет гибко учитывать влияние качественных факторов, в том числе сезонных и структурных изменений.

Вопросы для самопроверки

1. Опишите эконометрическую модель, приводящую к множественной линейной регрессии.
2. Какова эмпирическая основа построения эмпирической парной регрессии?
3. Что понимается под спецификацией модели множественной линейной регрессии?
4. Докажите справедливость формул вычисления МНК оценок параметров множественной линейной регрессии.
5. Дайте интерпретацию уравнению регрессии $\hat{y} = 3 + 2x_1 - 4x_2$, где y — доход (в млн. руб), x_1 — объем инвестиций в ИТ технологии (в сотнях тыс. долларов), x_2 — объем заработной платы (в сотнях тысяч руб.).
6. В чем состоят основные предположения регрессионного анализа в случае множественной регрессии?
7. Что является несмещенной оценкой дисперсии возмущений? Приведите формулу.
8. Как определяются стандартные ошибки коэффициентов регрессии?
9. Укажите статистики, распределенные по закону Стьюдента в множественной линейной регрессии.
10. Как строятся интервальные оценки коэффициентов регрессии и как они связаны с проверкой коэффициентов на статистическую значимость?
11. Каким образом можно оценить качество уравнения регрессии?
12. Чем отличается множественный коэффициент детерминации от скорректированного коэффициента детерминации?

13. Чем отличается выборочный коэффициент парной корреляции от коэффициента частной корреляции?
14. В чем суть статистической значимости коэффициентов регрессии? Сформулируйте правило проверки статистической значимости коэффициентов парной линейной регрессии.
15. В чем состоит идея проверки статистической значимости уравнения регрессии в целом? Сформулируйте правило проверки.
16. Объясните значение терминов коллинеарность и мультиколлинеарность.
17. В чем отличия между различными формами мультиколлинеарности?
18. Каковы основные последствия мультиколлинеарности?
19. Как можно обнаружить мультиколлинеарность?
20. Перечислите основные методы устранения мультиколлинеарности.
21. Что представляют собой фиктивные переменные?
22. Каковы основные причины использования фиктивных переменных в регрессионном анализе?
23. В чем смысл «ловушки фиктивных переменных»?
24. В чем состоит основное правило для определения количества вводимых фиктивных переменных?
25. Пусть для некоторого предприятия выборочная регрессионная модель имеет вид: $\hat{y} = 5 + 2x + 3d$, где $\hat{y}$ — заработная плата (в сотнях долларов), x — стаж работы (в десятилетиях), d — фиктивная переменная, отражающая пол сотрудника ($d = 0$ для женщин, $d = 1$ — для мужчин). Дайте интерпретацию коэффициентам этого уравнения.
26. Приведите примеры использования фиктивных переменных для учета сезонных особенностей.

Библиография

[1] Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. — М.: ЮНИТИ, 1998. — 650 с.

[2] Буре В.М.. Евсеев Е.А. Основы эконометрики: Учеб. Пособие. — СПб.: Изд-во С.-Петерб. ун-та, 2004.— 72 с.

[3] Валландер С.С. Заметки по эконометрике. — СПб.: Европ. ун-т, 2001. — 46 с.

[4] Доугерти К. Введение в эконометрику: учебник. 2-е изд. М.: ИНФРА-М, 2004.— 432 с.

[5] Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов.— М.: ЮНИТИ-ДАНА, 2004.— 311 с.

[6] Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. — М.: Дело, 2000. — 400 с.

[7] Эконометрика: Учебник / Под ред. И.И.Елисеевой. — М.: Финансы и статистика, 2001. — 344 с.