ТЕМА. ПОСТРОЕНИЕ МНОГОФАКТОРНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Многофакторные регрессионные модели: понятие и этапы построения

Экономические явления, как правило, определяются большим числом одновременно и совокупно действующих факторов. В связи с этим часто возникает задача исследования зависимости переменной «Y» от нескольких объясняющих переменных «X_1, X_2,…, X_n».

Многофакторная (многомерная или множественная) регрессия позволяет построить и проверить связь между одной результирующей переменной и несколькими независимыми факторами, оказывающими на нее влияние.

Многофакторная (многомерная или множественная**) регрессия (***multiple regression model***)** – уравнение связи с несколькими независимыми объясняющими переменными. В общем виде его можно записать следующим образом:

$$y = \hat{f}(\vec{x}) = f(\vec{x}) + e,$$

где $\vec{x} = x_1, x_2, ..., x_n$ – независимые объясняющие переменные.

e – случайная величина, дополнительный остаточный член, который отражает влияние случайных ошибок, особенностей измерений и действий, оказывающих влияние на результирующую переменную, других объясняющих переменных, которые не были включены в уравнение. Ее также называют возмущением или остатком. Эту случайную составляющую можно считать случайной ошибкой прогноза *y* по заданному значению *x*.

Ошибка регрессии – это разность между фактическим и теоретическим значением:

$$e = f(\vec{x}) - \hat{f}(\vec{x}) = y - \hat{y}.$$

Примером такой связи можно рассматривать зависимость доходности финансовых активов от следующих факторов: темпов прироста валового внутреннего продукта (ВВП), уровня процентных ставок, уровня инфляции и усиления спекулятивного ожидания.

Разработка и построение любой модели для прогнозирования экономических процессов должны выполняться по следующим этапам (табл. 4.1). [3]

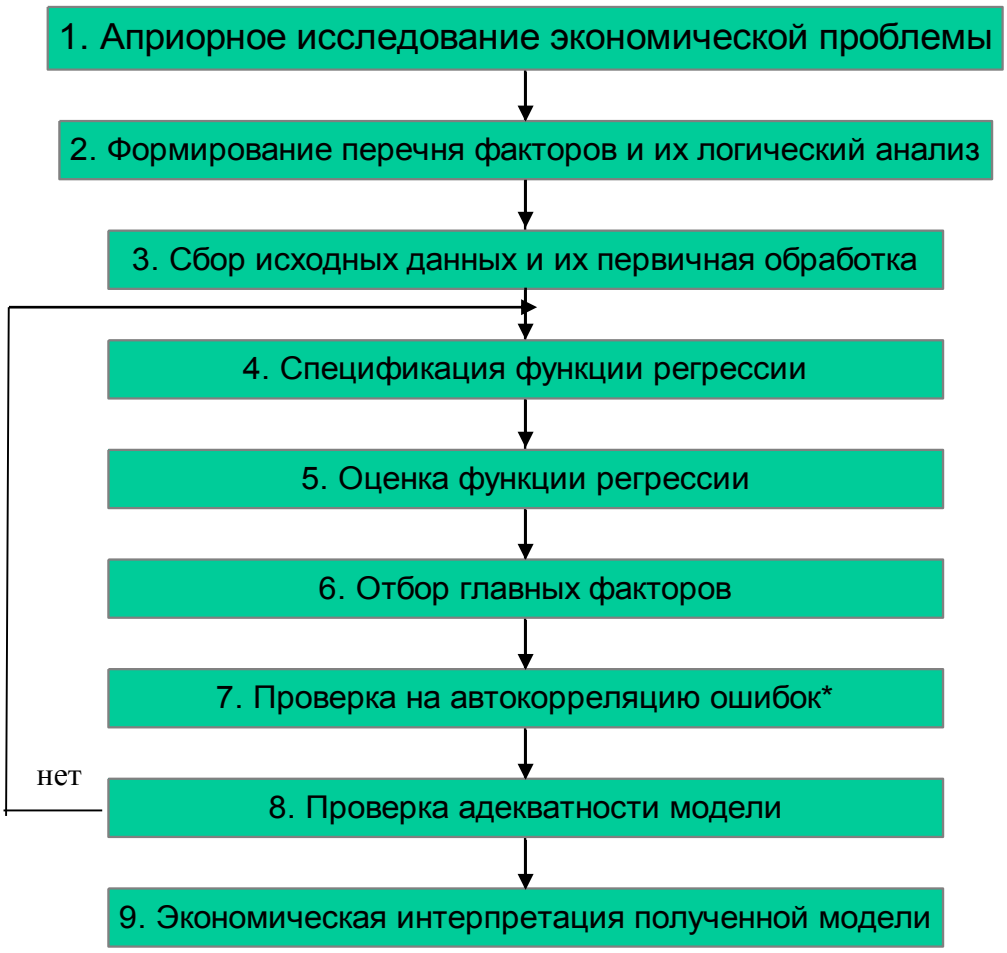

Рис. 4.1. Этапы построения многофакторной регрессионной модели

Рассмотрим подробнее **содержание этапов.**

На *первом этапе* в соответствии с целью работы конкретизируются явления, процессы, зависимость между которыми подлежит оценке. Формулируются гипотезы о зависимости экономических явлений.

На *втором этапе* определяют количество желаемых факторов из теоретических соображений, производят классификацию переменных на результирующую и объясняющие. Теоретически регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости.

На третьем – производят сбор данных. Определяют минимальный объем выборки.

На *четвертом этапе* формулируется гипотеза о форме связи (линейная или нелинейная, простая или множественная), т.е. проводится спецификация модели.

На *пятом этапе* определяют числовые значения параметров регрессии и показатели, характеризующие точность регрессионного анализа.

На *шестом* – отбираются основные факторы. Факторы, включаемые в модель должны удовлетворять следующим условиям:

1. Они должны быть количественно измеримы;
2. Факторы не должны содержать тенденции;
3. Факторы не должны быть интеркоррелированы (или находится в точной функциональной связи).

Седьмой этап включает про на автокорреляцию, (будет рассмотрена в отдельной теме)).

Восьмой этап включает проверку значимости показателя детерминации и оценку качества подбора теоретического уравнения регрессии. Оценивается ошибка регрессии, стандартная ошибка регрессии и средняя ошибка аппроксимации.

Если модель не адекватна, разработка модели должна начаться снова с четвертого этапа.

На последнем *девятом этапе* результаты сравниваются с гипотезами, предложенными на первом этапе исследования, и оценивается их правдоподобие. Проводится экономическая интерпретация модели. Модель готова для использования при построении прогнозов.

Рассмотрим подробнее некоторые из этапов построения регрессионной модели применительно к многофакторной регрессии.

Спецификация многофакторной регрессионной модели

Значение зависимой переменной *y* часто складывается под влиянием факторов, описываемых как количественными, так и неколичественными показателями. Для отражения влияния неколичественного показателя на результативный признак используют так называемые, фиктивные переменны.

Фиктивные переменные – это искусственно созданные переменные, для перевода качественных показателей в количественные. Например, район, пол, вид груза, марка автомобиля и т.д.

В качестве фиктивных переменных обычно используют дихотомические (бинарные) переменные, которые принимают всего два значения, например, «0» и «1».

Например, надо изучить зависимость заработной платы работников не только от количественных факторов x_1, x_2, $x_3 \ldots x_n$, но и от качественного признака Z_1 (например, фактор «пол работника»).

В этом случае регрессионная модель заработной платы будет выглядеть следующим образом:

$$\hat{y}=a+в_1x_{i1}+в_2x_{i2}+в_3x_{i3}+в_4x_{i4}+в_5x_{i5}+в_6x_{i6}+\ldots+в_nx_{in}+\beta_1Z_{i1}$$

где Z_{i1} = $\begin{cases}1, \text{ если } i\text{-й работник мужского пола;}\\0, \text{ если } i\text{-й работник женского пола.}\end{cases}$

Таким образом, можно считать, что средняя заработная плата у мужчин на $\beta_1 * 1 = \beta$ выше, чем у женщин, при неизменных значениях других параметров модели.

Если рассматриваемый качественный признак имеет несколько (k) уровней (градаций), то можно было бы ввести в регрессионную модель дискретную переменную, принимающую такое же количество значений. Однако обычно так не поступают из-за трудности содержательной интерпретации соответствующих коэффициентов регрессии, а вводят (k – 1) бинарных переменных.

Например, при исследовании зависимости заработной платы «y» от уровня образования Z можно рассматривать k = 3: Z_{i1} = 1 при наличии начального образования; Z_{i2} = 2 при наличии среднего образования; Z_{i3} = 3 при наличии высшего образования. Однако в рассматриваемом примере можно использовать всего две бинарные переменные [20]:

где Z_{i21} = $\begin{cases}1, \text{ если i-й работник имеет высшее образование;}\\0 \text{ - во всех остальных случаях.}\end{cases}$

где Z_{i22} = $\begin{cases}1, \text{ если i-й работник имеет среднее образование;}\\0 \quad \text{– во всех остальных случаях.}\end{cases}$

Для построения многофакторной регрессионной модели необходимо знать минимальный объем выборки, который зависит от числа факторов, включаемых в модель с учетом свободного члена. Для получения статистически значимой модели на один фактор требуется объем наблюдений, равный 5-8 наблюдениям.

Определить минимальный объем выборки для получения статистически значимой модели можно по формуле:

$$N_{min}=5(m+n),$$

где m – число факторов, включаемых в модель,

n – число свободных членов в уравнении.

Выбор вида математической функции для построения уравнения регрессии может быть осуществлен 3 методами [3]:

1. *Графическим* – используется только для однофакторных регрессий, когда вид уравнения удобно определять с помощью графика;

2. *Аналитическим* – основан на изучении материальной природы связи исследуемых признаков. Как пример, можно рассматривать метод наименьших квадратов;

3. *Экспериментальным* –позволяет на основе эксперимента задать уравнение регрессии.

Экспериментальный метод плох для изучения экономических процессов и им не пользуются. Графический метод удобен только для однофакторных регрессий. Следовательно, для многофакторной регрессии необходимо применять аналитический метод. В самом простом случае – это *метод наименьших квадратов (МНК)*.

Многофакторные регрессии можно разделить на линейную и нелинейные. В таблице 4.1 представлены некоторые виды функций, наиболее часто используемых на практике.

Для построения уравнений линейной регрессии и нелинейных, но приводимых к линейным, необходимо оценить ее параметры, это можно сделать, например, с помощью метода наименьших квадратов (МНК). При этом строится система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии.

Таблица 4.1

Наиболее распространенные виды функций и их преобразование

Название	Вид функции	Вид линейной функции	Метод преобразования нелинейной функции в линейную
Линейная	$\hat{y} = a_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_n x_n + e,$	–	–
Показательная	$\hat{y} = a_0 b_1^{x_1} b_2^{x_2} b_3^{x_3} ... b_n^{x_n} e.$	Ln Y = Ln a + x_1 Ln b_1 + x_2 Ln b_2 + + x_3 Ln b_3 + …+ x_n Ln b_n или $\hat{Y}$ = A + B_1 x_1 + B_2 x_2+ B_3 x_3+ … + B_n x_n	Линеаризация
Степенная	$\hat{y} = a_0 x_1^{b_1} x_2^{b_2} x_3^{b_3} ... x_n^{b_n} e.$	Ln Y = Ln a + b_1 Ln X_1 + + b_2 Ln X_2+ b_3 Ln X_3+ … + b_n Ln X_n или $\hat{Y}\hat{Y}$ = A + b_1 X_1 + b_2 X_2+ b_3 X_3+ … + b_n X_n	
Экспоненциальная	$\hat{y} = e^{a_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_n x_n + e}$	Ln Y = a + b_1x_1+ b_2x_2 + b_3x_3+ …+ b_nx_n Или $\hat{Y}$= a + b_1x_1+ b_2x_2 + b_3x_3+ …+ b_nx_n	
Гиперболическая	$\hat{y} = \frac{1}{a_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_n x_n} + e.$	$\frac{1}{y}$ = a + b_1x_1+ b_2x_2+ b_3x_3+ …+ + b_nx_n или $\hat{Y}$= a + b_1x_1+ b_2x_2 + b_3x_3+ …+ b_nx_n	Замена переменной

*Примечание 1***.** Если регрессия имеет мультипликативный вид, когда переменные и коэффициенты перемножаются, то ошибка е тоже домножается. Если регрессия имеет аддитивный вид, когда переменная и коэффициенты складываются, то ошибка е прибавляется.

*Примечание 2***.** В экспоненциальной регрессии в основании е – это экспонента, а в степени – ошибка регрессии.

Оценка параметров многофакторной регрессионной модели

Для определения значений параметров уравнения множественной регрессии используют числовую информацию, которая рассматривается как выборочная. Поэтому рассчитанные на ее основе величины называют оценками параметров, подчеркивая тем самым их возможную неточность из-за неполноты информации. Оценки параметров могут меняться от выборки к выборке, поэтому они рассматриваются как случайные величины.

Так как найденные параметры являются лишь выборочными оценками неизвестных параметров по генеральной совокупности, то возникает вопрос об их качестве. Характеристиками качества полученных оценок параметров регрессии являются их несмещенность, эффективность и состоятельность.

Оценка параметра является ***несмещенной,*** если ее математическое ожидание равно оцениваемому параметру, а математическое ожидание остатков равно нулю. Например, математическое ожидание оценки коэффициента регрессии a_j равно его значению в генеральной совокупности α_j:

$$M\, a_j = \alpha_j.\ M(e) = 0.$$

Поскольку оценки являются случайными переменными, их значения лишь по случайному совпадению могут в точности равняться характеристикам генеральной совокупности. Обычно будет присутствовать определенная ошибка, которая может быть большой и малой, положительной и отрицательной в зависимости от чисто случайных составляющих величин «*x*» в выборке.

Необходимо чтобы математическое ожидание оценки равнялось бы соответствующей характеристике генеральной совокупности. Если это так, то оценка называется несмещенной. Если это не так, то оценка называется смещенной, и разница между ее математическим ожиданием и соответствующей характеристикой генеральной совокупности называется смещением.

Оценка параметра является ***эффективной***, если она имеет наименьшую дисперсию среди всех возможных оценок данного параметра по выборкам одного и того же объема.

Желательно получить оценку с максимально возможной вероятностью, близкой к значению теоретической характеристики, что означает получить функцию плотности вероятности, как можно

более «сжатую» вокруг истинного значения. Один из способов выразить это требование – сказать, что необходимо получить сколь возможно малую дисперсию.

Эффективная оценка – это та, у которой дисперсия минимальна. Она минимальна в том случае, если наблюдения имеют равные веса.

Оценка параметра является ***состоятельной***, если с увеличением числа наблюдений она стремится к значению параметра в генеральной совокупности.

Состоятельной называется такая оценка, которая дает точное значение для большой выборки независимо от входящих в нее конкретных наблюдений [43, 57].

Простейшим методом оценки параметров множественной регрессии является МНК. МНК-оценки будут несмещенными, эффективными и состоятельными при выполнении определенных требований, называемых предпосылками МНК *(условия Гаусса-Маркова)*:

1. Математическое ожидание случайного остатка равно нулю:

$$M(e_i) = 0.$$

2. Случайные остатки не зависят друг от друга (не автокоррелированны):

$$r_{e_i e_j} = 0, i \neq j.$$

3. Случайные остатки не зависят от значений независимых переменных, входящих в модель регрессии: $r_{e_i x_j} = 0$

4. Случайные остатки распределены по нормальному закону распределения.

5. Дисперсия случайных остатков одинакова для различных i и j:

$$\sigma_{\varepsilon_i}^2 = \sigma_{\varepsilon_j}^2 = \sigma_{\varepsilon}^2$$

Ситуация, когда дисперсия каждого отклонения одинакова для всех значений x называется **гомоскедастичностью** (рис. 4.2).

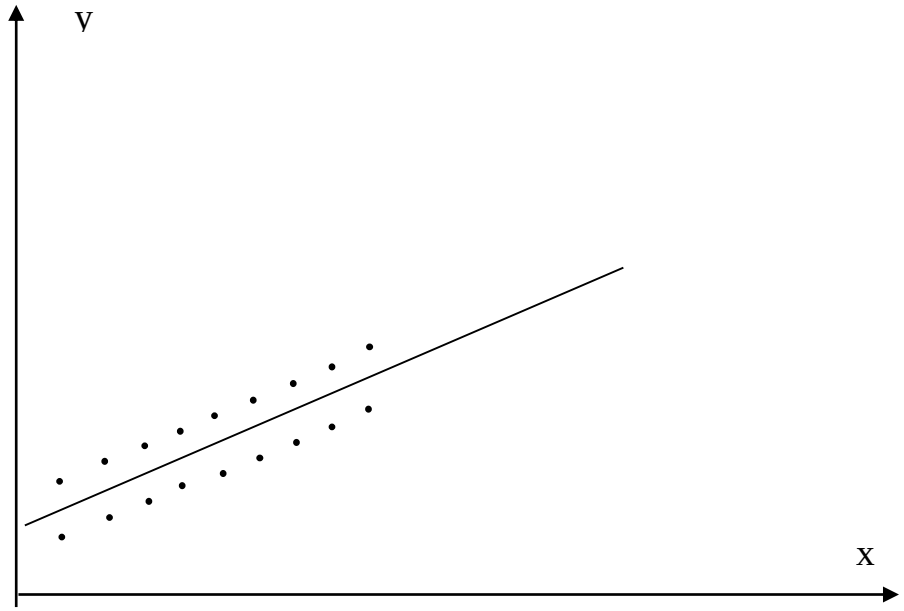

Рис. 4.2. Гомоскедастичность остатков

Ситуация, когда дисперсия каждого отклонения неодинакова для всех значений *x* называется **гетероскедастичностью** (рис. 4.3).

$$\sigma_{ei}^2 \neq \sigma_{ej}^2 \neq \sigma^2, j \neq i..$$

Наличие гетероскедастичности приводит в отдельных случаях к тому, что оценки коэффициентов регрессии оказываются смещенными. Кроме того, гетероскедастичность будет сказываться на уменьшении эффективности оценок [3, 43, 57].

При малом объеме выборки оценить нарушение гомоскедастичности можно с помощью параметрического **теста Гольдфельда-Квандта**:

1. Упорядочение n наблюдений по мере возрастания переменной x;
2. Исключение из рассмотрения C центральных наблюдений;

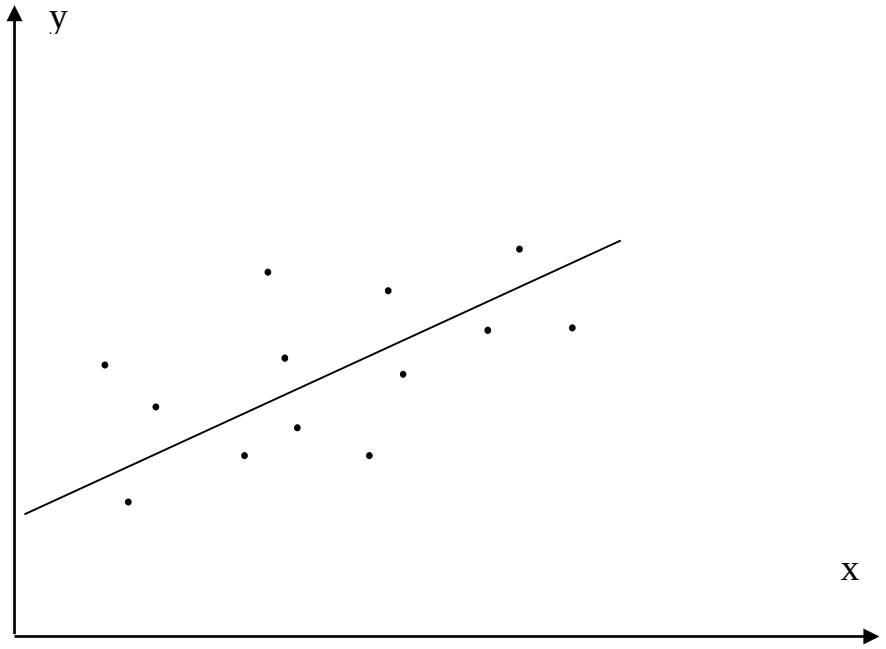

Рис. 4.3. Гетероскедастичность остатков

3. Разделение совокупности (n-C) наблюдений на 2 группы (где в одной будут малые значения x, а в другой – большие) и определение для них уравнений регрессий;
4. Нахождение остаточной суммы квадратов отклонений для этих групп и определение их соотношения.

Если распределение случайных остатков не соответствует предпосылкам, то проводят корректировку исходной модели [3].

МНК основывается на принципе минимизации квадратов отклонений фактических значений результативно признака y от его выровненных значений $\hat{y}$, рассчитанных по уравнению регрессии, т.е.

$$\sum (y - \hat{y})^2 \to \min .$$

Другими словами, из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной, т.е.

$$\sum_i e_i^2 \to \min,$$

Введем обозначение:

$$S = \sum e_i^2 .$$

Тогда

$$S = \sum(y - a - b_1x_1 - b_2x_2 - b_3x_3 - \ldots - b_nx_n)^2 \rightarrow \min.$$

Для нахождения экстремума по каждому из неизвестных параметров a и b_i рассчитывается производная функция и полученное выражение приравнивается к нулю:

$$\begin{cases} \frac{dS}{da} = \sum(-2)(y - a - b_1x_1 - b_2x_2 - b_3x_3 - \ldots - b_nx_n) = 0; \\ \frac{dS}{db_1} = \sum(-2x_1)(y - a - b_1x_1 - b_2x_2 - b_3x_3 - \ldots - b_nx_n) = 0; \\ \frac{dS}{db_2} = \sum(-2x_2)(y - a - b_1x_1 - b_2x_2 - b_3x_3 - \ldots - b_nx_n) = 0; \\ \ldots \\ \frac{dS}{db_n} = \sum(-2x_n)(y - a - b_1x_1 - b_2x_2 - b_3x_3 - \ldots - b_nx_n) = 0; \end{cases}$$

После преобразований получаем следующую систему нормальных уравнений [57]:

$$\begin{cases} \sum y = na + b_1\sum x_1 + b_2\sum x_2 + \ldots + b_n\sum x_n, \\ \sum yx_1 = a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 + \ldots + b_n\sum x_nx_1, \\ \sum yx_2 = a\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2 + \ldots + b_n\sum x_2x_n. \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum yx_n = a\sum x_n + b_1\sum x_1x_n + b_2\sum x_2x_n + \ldots + b_n\sum x_n^2. \end{cases}$$

В многофакторной линейной регрессии *коэффициенты* b_i называются **коэффициентами многофакторной регрессии**.

Интерпретация коэффициентов многофакторной регрессии:

– коэффициент регрессии при количественной объясняющей переменной интерпретируется как среднее изменение результирующей переменной при единичном изменении самой объясняющей переменной и неизменных значениях остальных независимых переменных.

– коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение результирующей переменной при переходе от одной категории к другой при неизменных значениях остальных независимых переменных.

Оценка тесноты связи в модели многофакторной регрессии

Соотношение между одной результирующей переменной и несколькими независимыми объясняющими переменными, взятыми в целом, может быть найдено путем вычисления **коэффициента многофакторной корреляции.**

Коэффициент множественной корреляции можно найти по формуле:

$$R_{y,x_1,x_2,...,x_n} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_{общ}}},$$

где $\sigma^2_{ост}$ – остаточная дисперсия результирующей переменной;

$\sigma^2_{общ}$ – общая дисперсия результирующей переменной.

Они находятся по следующим формулам:

$$\sigma^2_{ост} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1},$$

$$\sigma^2_{общ} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n-1},$$

где n - объем выборочной совокупности.

Коэффициент множественной корреляции можно использовать не только для многофакторных регрессий, но и для однофакторных. Можно сказать, что коэффициент множественной корреляции является показателем адекватности модели.

Коэффициент детерминации служит для оценки точности регрессии, т.е. соответствия полученного уравнения регрессии и фактическим данным. Его можно найти по формуле:

$$R^2_{y,x_1,x_2,...,x_n} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_{общ}}$$

Коэффициент детерминации показывает долю вариации результативного признака, находящегося под воздействием факторных признаков, т.е. определяет, какая доля вариации признака «у» учтена в модели и обусловлена влиянием на него факторов включаемых в модель. Чем ближе R^2 к единице, тем выше качество модели.

Зависимость между одной результирующей переменной и одной независимой объясняющей переменной при других упомянутых переменных, которые предполагаются постоянными, может быть найдено путем вычисления **коэффициента частной корреляции**. 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 случае, если у нас только две переменные x_1 и x_2 в регрессии, коэффициент частной корреляции вычисляется по формуле:

$$r_{y,x_1,x_2} = \sqrt{\frac{\sigma^2_{(ост)y,x_1} - \sigma^2_{(ост)y,x_1,x_2}}{\sigma^2_{(ост)y,x_1}}}.$$

Он будет называться коэффициентом частной корреляции 1ого порядка. Для получения коэффициентов более высокого порядка необходимо рассматривать многофакторные регрессии, содержащие большее количество факторов. Например, если рассматривается модель, содержащая s количество факторов, то влияние фактора x_1 можно рассматривать через коэффициенты частной корреляции:

r_{y,x_1,x_2} – при постоянном действии фактора x_2;

r_{y,x_1,x_2,x_3} – при постоянном действии фактора x_2 и x_3 (коэффициент корреляции 2-ого порядка);

и т.д.,

$r_{y,x_1,x_2,x_3,\ldots,x_s}$ – при постоянном действии фактора x_2, $x_3,\ldots,x_s$ (коэффициент корреляции s-ого порядка).

Аналогично можно рассматривать влияние фактора x_2:

r_{y,x_2,x_1} – при постоянном действии фактора x_1;

r_{y,x_2,x_1,x_3} – при постоянном действии фактора x_1 и x_3;

и т.д.,

$r_{y,x_2,x_1,x_3,\ldots,x_s}$ – при постоянном действии фактора x_1, x_3,…,x_s.

Аналогично можно рассматривать влияние и остальных факторов.

Сопоставление коэффициентов частной корреляции разного порядка по мере увеличения включаемых факторов в модель позволяет определиться с их количеством [3].

Например, при изучении зависимости себестоимости добычи газа от объема добычи коэффициент парной корреляции оказался равный -0,74. Модно говорить о довольно тесной обратной связи признаков. Частный коэффициент корреляции этой зависимости при постоянном влиянии уровня производительности труда равен 0,61, а значит показывает достаточную, но уже заметно менее тесную связь себестоимости и объема добычи. Закрепив на постоянном уровне и размер основных фондов, теснота связи рассматриваемых признаков получилась еще более низкой -0,48.

Отбор главных факторов. Мультиколлинеарность

После отбора факторов на основе знаний экономической теории, происходит их отсев на основе корреляционной матрицы. Наибольшие трудности в использовании аппарата множественной регрессии возникают при наличии мультиколлинеарности факторов, когда более, чем два фактора связаны между собой линейной зависимостью.

Включаемые в модель факторы должны объяснять вариацию результативной переменной. Показатель детерминации r^2 (ρ^2) показывает долю объясненной вариации у за счет рассматриваемых в регрессии n факторов. Влияние других неучтенных в модели факторов оценивается как $1-r^2$ с соответствующей остаточной дисперсией. При дополнительном включении в регрессию n+1 фактора показатель детерминации должен возрастать, а остаточная дисперсия уменьшаться. Если это не происходит, то включаемый в анализ фактор x_{n+1} не улучшает модель и является лишним фактором. Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и показателя детерминации, но и приводит к статистической незначимости параметров регрессии.

Мультиколлинеарность – попарная корреляционная зависимость между факторами. Она присутствует, если коэффициент парной корреляции $\left|r_{yx_i}\right| \geq 0.7,$.

Отрицательное воздействие мультиколлинеарности состоит в следующем:

– Усложняется процедура выбора главных факторов;

– Искажается смысл коэффициента множественной корреляции;

– Усложняются вычисления при построении самой модели;

– Снижается точность оценки параметров регрессии, искажается оценка дисперсии.

Поэтому обязательным условием при построении регрессионной модели является анализ факторов на мультиколлинеарность и её устранение.

Если бы факторы не коррелировали между собой, то определитель матрицы парных коэффициентов был бы равен единице. Например, для модели $y= a_0+a_1x_1+a_2x_2+e$

$$\begin{vmatrix} r_{x1x1} & r_{x1x2} \\ r_{x2x1} & r_{x2x2} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1.$$

И наоборот, если все факторы коррелированы, т.е. между ними существует линейная зависимость, то определитель этой матрицы будет равен 0 или на примере:

$$\begin{vmatrix} r_{x1x1} & r_{x1x2} \\ r_{x2x1} & r_{x2x2} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 0.$$

Значит можно сделать вывод, что чем ближе к 0 определитель такой матрицы, тем сильнее линейная зависимость между факторами [3].

Для устранения мультиколлинеарности используют *метод исключения переменных* [41]. Он заключается в том, что высоко коррелированные объясняющие переменные устраняются из регрессии, и она заново оценивается. Если $\left|r_{yx_i}\right| \geq 0.7,$ то одну из переменных можно исключить, но какую именно, решают исходя из управляемости факторов.

Если возникает ситуацию, когда оба фактора одновременно управляемы или нет, то решить вопрос об исключении того или иного фактора можно с помощью *процедуры отбора главных факторов.*

Процедура отбора главных факторов включает обязательно следующие этапы:

1. Производится анализ значения коэффициентов парной корреляции r_{ij} между факторами x_i и x_j.
2. Анализ тесноты взаимосвязи объясняющих факторов с результативной переменной.

Прогнозирование по модели множественной регрессии

Прогнозирование по модели множественной линейной регрессии предполагает оценку ожидаемых значений зависимой переменной при заданных значениях независимых переменных, входящих в уравнение регрессии. Различают точечный и интервальный прогнозы.

Точечный прогноз – это расчетное значение зависимой переменной, полученное подстановкой в уравнение множественной линейной регрессии прогнозных (заданных исследователем) значений независимых переменных.

Если заданы значения $x_1^{пр}, x_2^{пр}, \ldots, x_n^{пр}$, то прогнозное значение зависимой переменной (точечный прогноз) будет равно:

$$\hat{y}_{пр} = a + b_1 x_1^{пр} + b_2 x_2^{пр} + \ldots + b_n x_n^{пр}.$$

Интервальный прогноз – это минимальное и максимальное значения зависимой переменной, в промежуток между которыми она попадает с заданной долей вероятности и при заданных значениях независимых переменных.

Интервальный прогноз для линейной функции вычисляется по формуле:

$$\hat{y}_t \pm t_\alpha S_y,$$

где t_α t_a – критическое значение критерия Стьюдента;

S_y – стандартная ошибка прогноза, вычисляемая по формуле:

$$S_y = \sqrt{\frac{\sum (y - \hat{y})^2}{n - l - 1}},$$

где n – объем выборки;

l – количество факторов в модели.

Рассмотрим их построение на примере уравнения однофакторной линейной регрессии $y=f(x)$. На рисунке 4.4 ветви гиперболы ограничивают доверительную область, и располагаются выше и ниже линии регрессии.

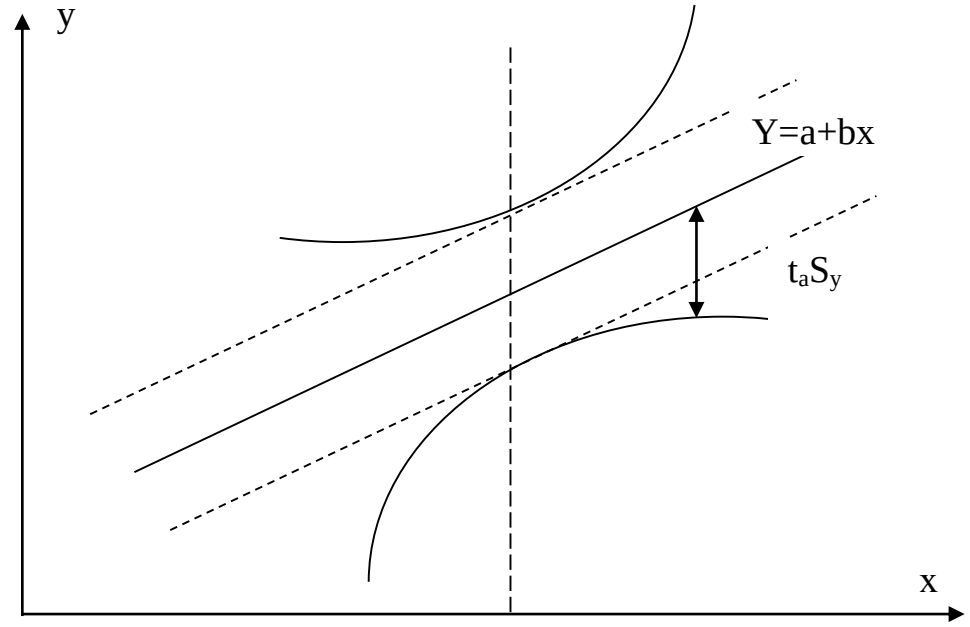

Рис. 4.4. Доверительные интервалы уравнения линейной регрессии

Прогнозирование по нелинейным моделям множественной регрессии осуществляется аналогично, предварительно линеаризовав указанные модели.

Оценить точность прогнозов можно с помощью ошибки прогноза, она будет показывать относительное отклонение расчетных (регрессионных) значений от фактических.

Ее можно найти по формуле:

$$\frac{y_{фактическое} - y_{теоретическое}}{y_{фактическое}} \cdot 100\% = \frac{y - \hat{y}}{y} \cdot 100\%$$