МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИО ЭЛЕКТРОНИКИ» (ТУСУР)

Кафедра экономической математики, информатики и статистики

Составитель: Сидоренко М.Г.

ЭКОНОМЕТРИКА

Учебное пособие

Томск 2018

СОДЕРЖАНИЕ

Введение

В последнее время широкое распространение получило использование моделирования и количественного анализа в экономике. В результате выделилось и сформировалось одно из направлений экономических исследований – «эконометрика».

Формально «эконометрика» означает «измерения в экономике». Однако область исследований этой дисциплины гораздо шире. Эконометрика- это наука, в которой на базе реальных статистических данных строятся, анализируются и совершенствуются математические модели реальных экономических процессов. Эконометрика позволяет найти количественное подтверждение либо опровержение того или иного экономического закона или гипотезы. При этом одним из важнейших направлений эконометрики является построение прогнозов по различным экономическим показателям.

Эконометрика как научная дисциплина зародилась на основе слияния экономической теории, математики и статистики.

1. Парная линейная регрессия

1.1. Суть регрессионного анализа

Поведение и значение любого экономического показателя зависят практически от бесконечного количества факторов, и все их учесть нереально. Однако обычно лишь ограниченное количество факторов действительно существенно воздействуют на исследуемый экономический показатель. Доля влияния остальных факторов столь незначительна, что их игнорирование не может привести к существенным отклонениям в поведении исследуемого объекта. В естественных науках большей частью имеют дело со строгими (функциональными) зависимостями, при которых каждому значению одной переменной соответствует единственное значение другой. Однако в подавляющем большинстве случаев между экономическими переменными таких зависимостей нет. Поэтому в экономике говорят не о функциональных, а о *корреляционных,* либо *статистических,* зависимостях.

Если переменные обозначить X и Y, то зависимость вида:

$$M(Y|X) = f(x) \tag{1.1}$$

называется *функцией регрессии Y на X*. При этом X называется независимой (объясняющей) переменной (прегрессором), Y – зависимой (объясняемой) переменной. При рассмотрении двух случайных величин говорят о *парной регрессии*.

Зависимость нескольких переменных, выражаемая функцией

$$M(Y|\ x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m) \tag{1.2}$$

называют *множественной регрессией*.

Под *регрессией* понимается функциональная зависимость между объясняющими переменными и условным математическим ожиданием (средним значением) зависимой переменной, которая строится с целью предсказания (прогнозирования) этого среднего значения при фиксированных значениях первых.

Для отражения того факта, что реальные значения зависимой переменной не всегда совпадают с ее условными математическими ожиданиями и могут быть различными при одном и том же значении объясняющей переменной (наборе объясняющих переменных), фактическая зависимость должна быть дополнена некоторым слагаемым ε, которое, по существу, является случайной величиной и указывает на стохастическую суть зависимости. Из этого следует, что связи между зависимой и объясняющей (ими) переменными выражаются соотношениями

$$Y = M(Y|X) + \varepsilon \tag{1.3}$$

$$Y = M(Y| \ x_1, x_2, \ldots, x_m) + \varepsilon \tag{1.4}$$

называемыми *регрессионными моделями (уравнениями)*.

Среди причин обязательного присутствия в регрессионных моделях случайного фактора (отклонения) можно выделить следующие:

1.*Невключение в модель всех объясняющих переменных.* Любая регрессионная (в частности, эконометрическая) модель является упрощением реальной ситуации. Последняя всегда представляет собой сложнейшее переплетение различных факторов, многие из которых в модели не учитываются, что порождает отклонение реальных значений зависимой переменной от ее модельных значений. Например, спрос *(Q)* на товар определяется его ценой *(P),* ценой *(P_s)* на товары-заменители, ценой (P_c) на дополняющие товары, доходом *(Г)* потребителей, их количеством *(N),* вкусами *(T),* ожиданиями *(W)* и т. д. Безусловно, перечислить все объясняющие переменные здесь практически невозможно. Например, мы не учли такие факторы, как традиции, национальные или религиозные особенности, географическое положение региона, погода и многие другие, влияние которых приведет к некоторым отклонениям реальных наблюдений от модельных, которые можно выразить через случайный член е: $Q = f(P, P_s, P_c, I, N, T, W, \varepsilon)$. Проблема еще и в том, что никогда заранее не известно, какие факторы при создавшихся условиях действительно являются определяющими, а какими можно пренебречь. Например, в ряде случаев учесть непосредственно какой-то фактор нельзя в силу невозможности получения по нему статистических данных.

2.*Неправильный выбор функциональной формы модели.* Из-за слабой изученности исследуемого процесса либо из-за его переменчивости может быть неверно подобрана функция, его моделирующая. Это, безусловно, скажется на отклонении модели от реальности, что отразится на величине случайного члена. Кроме того, неверным может быть подбор объясняющих переменных.

3. *Агрегирование переменных.* Во многих моделях рассматриваются зависимости между факторами, которые сами представляют сложную комбинацию других, более простых переменных. Например, при рассмотрении в качестве зависимой переменной совокупного спроса проводится анализ зависимости, в которой объясняемая переменная является сложной композицией индивидуальных спросов, оказывающих на нее определенное влияние помимо факторов, учитываемых в модели. Это может оказаться причиной отклонения реальных значений от модельных.

4. *Ошибки измерений.* Какой бы качественной ни была модель, ошибки измерений переменных отразятся на несоответствии модельных значений эмпирическим данным, что также отразится на величине случайного члена.

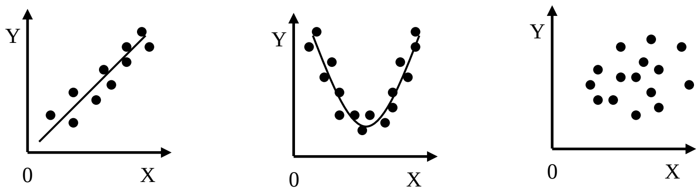
5. *Ограниченность статистических данных.* Зачастую строятся модели, выражаемые непрерывными функциями. Но для этого используется набор данных, имеющих дискретную структуру. Это несоответствие находит свое выражение в случайном отклонении.

6. *Непредсказуемость человеческого фактора.* Эта причина может «испортить» самую качественную модель. Действительно, при правильном выборе формы модели, скрупулезном подборе объясняющих переменных все равно невозможно спрогнозировать поведение каждого индивидуума.

Решение задачи построения качественного уравнения регрессии, соответствующего эмпирическим данным и целям исследования, является достаточно сложным и многоступенчатым процессом. Его можно разбить на три этана:

1) выбор формулы уравнения регрессии;
2) определение параметров выбранного уравнения;
3) анализ качества уравнения и поверка адекватности уравнения эмпирическим данным, совершенствование уравнения.

Выбор формулы связи переменных называется *спецификацией* уравнения регрессии. В случае парной регрессии выбор формулы обычно осуществляется по графическому изображению реальных статистических данных в виде точек в декартовой системе координат, которое называется *корреляционным полем (диаграммой рассеивания)* (рис.)



Если на первых двух графиках относительно четко определяется форма связи (для первого – линейная, для второго – квадратичная), то для третьего явная взаимосвязь между переменными отсутствует. В случае множественной регрессии определение подходящего вида регрессии является наиболее сложной задачей.

1.2. Парная линейная регрессия

Если функция регрессии линейна, то говорят о *линейной регрессии.* Модель линейной регрессии (линейное уравнение) является наиболее распространенным (и простым) видом зависимости между экономическими переменными. Кроме того, построенное линейное уравнение может служить начальной точкой эконометрического анализа. *Линейная регрессия (теоретическое линейное уравнение регрессии)* представляет собой линейную функцию между условным математическим ожиданием $M(Y|X = x_i)$ зависимой переменной *Y* и одной объясняющей переменной *X* (x_i - значения независимой переменной в *i*-ом наблюдении, $i = 1,2,\ldots,n$).

$$M(Y|X = x_i) = \beta_0 + \beta_1 x_i \tag{1.5}$$

Для отражения того факта, что каждое индивидуальное значение y_i отклоняется от соответствующего условного математического ожидания, необходимо ввести в последнее соотношение случайное слагаемое ε_i .

$$y_i = M(Y|X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1.6}$$

Это соотношение называется *теоретической линейной регрессионной моделью;* β_0 и β_1 — *теоретическими параметрами (теоретическими коэффициентами) регрессии;* ε_i — *случайным отклонением.*

Следовательно, индивидуальные значения y_i представляются в виде суммы двух компонент — систематической $(\beta_0 + \beta_1 x_i)$ и случайной ε_i , причина появления которой достаточно подробно рассмотрена ранее. В общем виде теоретическую линейную регрессионную модель будем представлять в виде:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1.7}$$

Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных *X* и *Y* генеральной совокупности, что практически невозможно.

Таким образом, задачи линейного регрессионного анализа состоят в том, чтобы по имеющимся статистическим данным (x_i, y_i), $i = 1,2,\ldots,n$ для переменных X и *Y:*

а) получить наилучшие оценки неизвестных параметров β_0 и β_1 ;

б) проверить статистические гипотезы о параметрах модели;

в) проверить, достаточно ли хорошо модель согласуется со статистическими данными (адекватность модели данным наблюдений).

Следовательно, по выборке ограниченного объема мы сможем построить так называемое *эмпирическое уравнение регрессии*

$$\hat{y}_i = b_0 + b_1 x_i \tag{1.8}$$

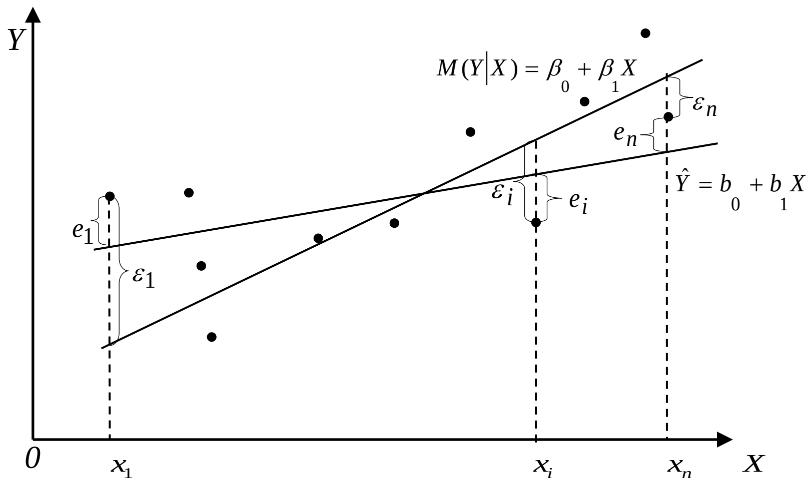
где $\hat{y}_i$ — оценка условного математического ожидания $M(Y|X = x_i)$; b_0 и b_1 -

оценки неизвестных параметров β_0 и β_1, называемые *эмпирическими коэффициентами регрессии.* Следовательно, в конкретном случае:

$$y_i = b_0 + b_1 x_i + e_i \tag{1.9}$$

где *отклонение* e_i — оценка теоретического случайного отклонения ε_i .

В силу несовпадения статистической базы для генеральной совокупности и выборки оценки b_0 и b_1 практически всегда отличаются от истинных значений коэффициентов β_0 и β_1, что приводит к несовпадению эмпирической и теоретической линий регрессии. Различные выборки из одной и той же генеральной совокупности обычно приводят к определению отличающихся друг от друга оценок. Возможное соотношение между теоретическим и эмпирическими уравнениями регрессии схематично изображено на рисунке.



Задача состоит в том, чтобы по конкретной выборке (x_i, y_i), $i = 1,2,\ldots,n$ найти оценки b_0 и b_1 неизвестных параметров β_0 и β_1 так, чтобы построенная линия регрессии являлась бы наилучшей в определенном смысле среди всех других прямых. Например, коэффициенты b_0 и b_1 эмпирического уравнения регрессии могут быть оценены исходя из условий минимизации одной из следующих сумм:

1. $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$, однако эта сумма не может быть мерой качества найденных оценок в силу того, что существует бесчисленное количество прямых, для которых $\sum_{i=1}^{n} e_i = 0$.

2. $\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} \lfloor y_i - \hat{y}_i \rfloor = \sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|$. Этот метод называется методом наименьшей суммы.

3. $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$. Это самый распростаренный и теоретически обоснованный метод, который получил название метода наименьших квадратов (МНК). Кроме того, он является наиболее простым с вычислительной точки зрения.

Найдем оценки b_0 и b_1, используя метод наименьших квадратов. При этом минимизируется следующая функция:

$$Q(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \qquad (1.10)$$

Эта функция является квадратичной функцией двух параметров b_0 и b_1. Условием существования минимума функции двух переменных является равенство нулю ее частных производных:

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\sum (y_i - b_0 - b_1 x_i) x_i = 0 \end{cases} \Rightarrow \begin{cases} n b_0 + b_1 \sum x_i = \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

Разделив оба уравнения системы на n, получим:

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y} \\ b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy} \end{cases} \Rightarrow \begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases}, \text{ где } \begin{array}{l} \bar{x} = \dfrac{1}{n}\sum x_i, \overline{x^2} = \dfrac{1}{n}\sum x_i^2 \\ \bar{y} = \dfrac{1}{n}\sum y_i \end{array} \qquad (1.11)$$

Из формул статистики очевидно, что:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$

Тогда: $b_1 = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y} \cdot \frac{S_y}{S_x} = r_{xy} \cdot \frac{S_y}{S_x}$, (1.12)

где r_{xy}-выборочный коэффициент корреляции, S_x, S_y - стандартные отклонения.

1.3.Проверка качества уравнения регрессии

Регрессионный анализ позволяет определить оценки коэффициентов регрессии. Но они не позволяют сдеалть вывод, насколько точно эмпирическое уравнение регресси соответствует уравнению для всей генеральной совокупности, насколько близки оценки b_0 и b_1 коэффициентов к своим теоретическим прототипам β_0 и β_1, как близко оцененное значение $\hat{y}_i$ к условному математическому ожиданию $M(Y|X = x_i)$, насколько надежны найденные оценки. Для ответа на эти вопросы необходимы дополнительные исследования.

Как следует из соотношения (1.6), значения y_i зависят от значений x_i и случайных отклонений ε_i. Следовательно, переменная Y является случайной величиной, напрямую связанной с ε_i. Таким образом, пока не будет определенности в вероятностном поведении ε_i, нет уверенности в качестве оценок.

Доказано, что для получения по МНК наилучших результатов необходимо, чтобы выполнялся ряд предпосылок относительно случайного отклонения (предпосылки Гаусса-Маркова):

1. Математическое ожидание случайного отклонения ε_i равно нулю: $M(\varepsilon_i) = 0$ для всех наблюдений. Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную. В каждом конкретном наблюдении может быть ибо положительным, либо отрицательным, но он не должен иметь систематического смещения. Выполнимость $M(\varepsilon_i) = 0$ влечет выполнимость $M(Y|X = x_i) = \beta_0 + \beta_1 x_i$

2. Дисперсия случайных отклонений ε_i постоянна: $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ для любых наблюдений i и j. Данное условие подразумевает, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть либо большим, либо меньшим, не должно быть некой причины, вызывающей большую ошибку (отклонение).
3. Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$. Выполнимость данной предпосылки предполагает, что отсутствует систематическая связь между любыми случайными отклонениями.
4. Случайное отклонение должно быть независимо от объясняющих переменных.
5. Модель является линейной относительно параметров.

Теорема Гаусса-Маркова. Если предпосылки 1-5 выполнены, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки являются несмещенными, т.е. $M(b_0) = \beta_0$, $M(b_1) = \beta_1$. Это вытекает из того, что $M(\varepsilon_i) = 0$, и говорит об отсутствии систематической ошибки в определении положения линии регрессии.
2. Оценки состоятельны, так как дисперсия оценок параметров при возрастании числа наблюдений n стремится к нулю: $D(b_0) \xrightarrow[n\to\infty]{} 0$, $D(b_1) \xrightarrow[n\to\infty]{} 0$. Другими словами, при увеличении объема выборки надежность оценок увеличивается (b_0 близко к β_0, а b_1 близко к β_1.
3. Оценки эффективны, т.е. они имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин y_i.

Если предпосылки 2 и 3 нарушены, то свойства несмещенности и состоятельности сохраняются, но свойство эффективности – нет.

Анализ точности определения оценок коэффициентов регрессии

В силу случайного отбора элементов в выборку являются также оценки b_0 и b_1 коэффициентов β_0 и β_1 теоретического уравнения регрессии. Их математические ожидания при выполнении предпосылок об отклонениях ε_i равны соответственно $M(b_0) = \beta_0$, $M(b_1) = \beta_1$. При этом оцени тем надежнее, чем меньше их разброс вокруг β_0 и β_1, т.е. чем меньше дисперсии $D(b_0)$ и $D(b_1)$ оценок, рассчитываемые по формулам:

$$D(b_0) = \frac{\sigma^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2}, \quad D(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{1.13}$$

Из соотношений (2.13) очевидны следующие выводы:

- Дисперсии b_0 и b_1 прямо пропорциональны дисперсии случайного отклонения σ^2. Следовательно, чем больше фактор случайности, тем менее точными будут оценки.
- Чем больше число наблюдений n, тем меньше дисперсия оценок. Это очевидно, так как чем большим числом данных мы располагаем, тем вероятнее получение более точных оценок.
- Чем больше дисперсия объясняющей переменной, тем меньше дисперсия оценок коэффициентов. Другими словами, чем шире область изменений объясняющей переменной, тем точнее будут оценки.

В силу того, что случайные отклонения ε_i по выборке определены быть не могут, при анализе надежности оценок коэффициентов регрессии они заменяются отклонениями $e_i = y_i - b_0 - b_1 x_i$ значений y_i переменной Y от оцененной линии регрессии. Дисперсия случайных отклонений $D(\varepsilon_i) = \sigma^2$ заменяется ее несмещенной оценкой

$$S^2 = \frac{1}{n-2}\sum (y_i - b_0 - b_1 x_i)^2 = \frac{\sum e_i^2}{n-2} \tag{1.14}$$

Тогда:

$$D(b_0) \approx S_{b0}^2 = \frac{S^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2} = \overline{x^2} S_{b1}^2 \tag{1.15}$$

$$D(b_1) \approx S_{b1}^2 = \frac{S^2}{\sum (x_i - \bar{x})^2} \tag{1.16}$$

$S^2 = \dfrac{\sum e_i^2}{n-2}$ - необъясненная дисперсия (мера разброса зависимой переменной вокруг линии регрессии). $S = \sqrt{\dfrac{\sum e_i^2}{n-2}}$ - стандартная ошибка оценки (стандартная ошибка регрессии).

$S_{b0}=\sqrt{S_{b0}^2}$ и $S_{b1}=\sqrt{S_{b1}^2}$ - стандартные отклонения случайных величин b_0 и b_1, называемые стандартными ошибками коэффициентов регрессии.

Проверка гипотез относительно коэффициентов линейного уравнения регрессии

При проведении статистического анализа перед исследователем зачастую возникает необходимость сравнения эмпирических коэффициентов регрессии b_0 и b_1 с некоторыми теоретически ожидаемыми значениями β_0 и β_1 этих коэффициентов. Данный анализ осуществляется при помощи статистической проверкой гипотез.

Для проверки гипотезы $b_1=\beta_1$ используется статистика:

$$t=\frac{b_1-\beta_1}{S_{b1}} \tag{1.17}$$

которая при справедливости $b_1=\beta_1$ имеет распределение Стьюдента с числом степеней свободы $v=n-2$, где n - объем выборки. Следовательно, гипотеза $b_1=\beta_1$ отклоняется на основании данного критерия, если:

$$\left|T_{набл}\right|=\left|\frac{b_1-\beta_1}{S_{b1}}\right|\geq t_{\frac{\alpha}{2},n-2} \tag{1.18}$$

где α - требуемый уровень значимости. При невыполнении (2.18) считается, что нет оснований для отклонения $b_1=\beta_1$.

Наиболее важной на начальном этапе статистического анализа построенной модели является задача установления значимости коэффициентов. Эта задача решается при помощи отношений, называемых t-статистикой:

$$t=\frac{b_1}{S_{b1}} \quad \text{и} \quad t=\frac{b_0}{S_{b0}} \tag{1.19}$$

В случае, если $|t|>t_{\frac{\alpha}{2},n-2}$, то статистическая значимость соответствующего коэффициента регрессии подтверждается. Значения $t_{\frac{\alpha}{2},n-2}$ в зависимости от уровня значимости и числа степеней свободы v ($v=n-2$) приведены в приложении 1.

При оценке значимости коэффициента линейной регрессии на начальном этапе можно использовать следующее «грубое» правило (почти всегда работает при $n>10$), позволяющее не прибегать к таблицам:

- Если $|t| \le 1$, то коэффициент не может быть признан значимым, так как доверительная вероятность составит менее чем 0,7.
- Если $1 < |t| \le 2$, то найденная оценка может рассматриваться как относительно (слабо) значимая. Доверительная вероятность в этом случае лежит между 0,7 и 0,95.
- Если $2 < |t| \le 3$, то доверительная вероятность составляет от 0,95 до 0,99.
- Если $|t| > 3$, то это почти гарантия значимости коэффициента.

Интервальные оценки коэффициентов линейного уравнения регрессии

Доверительные интервалы коэффициентов b_0 и b_1, которые с надежностью $(1-\alpha)$ накрывают определяемые параметры β_0 и β_1, определяются по формулам:

$$\left(b_0 - t_{\frac{\alpha}{2},n-2} S(b_0);\ b_0 + t_{\frac{\alpha}{2},n-2} S(b_0)\right) \quad \text{и} \quad \left(b_1 - t_{\frac{\alpha}{2},n-2} S(b_1);\ b_1 + t_{\frac{\alpha}{2},n-2} S(b_1)\right) \qquad (1.20)$$

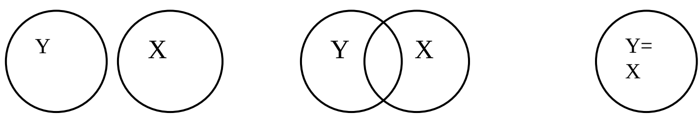
Доверительные интервалы для зависимой переменной

Интервал, определяющий границы, за пределами которых могут оказаться не более $100\alpha\%$ точек наблюдений при $X = x_p$, рассчитывается следующим образом:

$$\left(b_0 + b_1 x_p \pm t_{\frac{\alpha}{2},n-2} S\sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}}\right) \qquad (1.21)$$

Проверка общего качества уравнения регрессии.

Общее качество уравнения регрессии оценивается по тому, как хорошо эмпирическое уравнение регрессии согласуется со статистическими данными. Иными словами, насколько широко рассеяны точки наблюдений относительно линии регрессии. Очевидно, что если все точки лежат на построенной прямой, то регрессия Y на X «идеально» объясняет поведение зависимой переменной. Возможные соотношения между двумя переменными имеют наглядную графическую интерпретацию в виде так называемой диаграммы Венна.



На рисунке сначала X никак не влияет на Y. На следующем рисунке влияние X усиливается. В последнем случае значения Y целиком определяются значениями X.

Суммарно мерой общего качества уравнения регрессии является *коэффициент детерминации* R^2. В случае парной регрессии коэффициент детерминации будет совпадать с квадратом коэффициента корреляции. В общем случае коэффициент детерминации рассчитывается по формуле:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \qquad (1.22)$$

Справедливо соотношение $0 \le R^2 \le 1$. Чем теснее линейная связь между X и Y, тем ближе коэффициент детерминации R^2 к единице.

Пример. Для анализа зависимости объема потребления Y (у.е.) домохозяйства от располагаемого дохода X (у.е.) отобрана выборка n=12 (помесячно в течение года) необходимо провести регрессионный анализ.

Данные и расчеты представлены в таблице:

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$	e_i	e_i^2
1	107	102	11449	10914	10104	103,63	-1,63	2,66
2	109	105	11881	11445	11025	105,49	-0,49	0,24
3	110	108	12100	11880	11664	106,43	1,57	2,46
4	113	110	12769	12430	12100	109,23	0,77	0,59
5	120	115	14400	13800	13225	115,77	-0,77	0,59
6	122	117	14884	14274	13689	117,63	-0,63	0,40
7	123	119	15129	14637	14161	118,57	0,43	0,18
8	128	125	16384	16000	15625	123,24	1,76	3,10
9	136	132	18496	17952	17424	130,71	1,29	1,66
10	140	130	19600	18200	16900	134,45	-4,45	19,8
11	145	141	21025	20445	19881	139,11	1,89	3,57
12	150	144	22500	21600	20736	143,78	0,22	0,05
Сумма	1503	1448	190617	183577	176834	-	0	35,3
Среднее	125,25	120,67	15884,75	15298,08	14736,17	-	-	-

Согласно $\begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{15298{,}08 - 125{,}25 \cdot 120{,}67}{15884{,}75 - (125{,}25)^2} = 0{,}9339 \\ b_0 = \bar{y} - b_1 \bar{x} = 120{,}67 - 0{,}9339 \cdot 125{,}25 = 3{,}699 \end{cases}$ --.

Таким образом, уравнение парной регрессии имеет вид: $\hat{Y} = 3{,}699 + 0{,}9339X$. По этому уравнению рассчитаем $\hat{y}_i$, а также $e_i = y_i - \hat{y}_i$.

В нашем примере коэффициент b_1 может трактоваться как предельная склонность к потреблению. Фактически он показывает, на какую величину изменится объем потребления, если располагаемый доход возрастает на одну единицу. Свободный член b_0 определяет прогнозируемое значение Y при величине располагаемого дохода X, равной нулю (т.е. автономное потребление).

Рассчитаем другие показатели.

$$S^2 = \frac{\sum e_i^2}{n-2} = \frac{35,3}{12-2} = 3,53; \quad S = \sqrt{S^2} = 1,88$$

$$S_{b1}^2 = \frac{S^2}{\sum (x_i - \bar{x})^2} = \frac{3,53}{2366.25} = 0,0015; \; S_{b1} = \sqrt{S_{b1}^2} = 0,039$$

$$S_{b0}^2 = \overline{x^2} S_{b1}^2 = 15884,75 \cdot 0,0015 = 23,83; \quad S_{b0} = \sqrt{S_{b0}^2} = 4,88$$

Проверим статистическую значимость коэффициентов b_0 и b_1:

$$t_1 = \frac{b_1}{S_{b1}} = \frac{0,9339}{0,039} = 23,946 \quad \text{и} \quad t_0 = \frac{b_0}{S_{b0}} = \frac{3,699}{4,88} = 0,76.$$

Критическое значение при уровне значимости $\alpha = 0,05$ равно (см. приложение 1) $t_{крит} = t_{\frac{\alpha}{2}, n-2} = t_{0,025;10} = 2,228$. Так как $|t_1| = 23,946 > 2,228$, то это подтверждает статистическую значимость коэффициента регрессии b_1. Аналогично для другого коэффициента: так как $|t_0| = 0,76 < 2,228$, то гипотеза о статистической значимости коэффициента b_0 отклоняется. Это означает, что в данном случае свободным членом уравнения регрессии можно пренебречь, рассматривая регрессию как $Y = b_1 X$.

Определим доверительные интервалы коэффициентов регрессии (по формуле 2.20), которые с надежность 95% $(\alpha = 0.05)$ будут следующими:

для b_0 $(3,699 - 2,228 \cdot 4,88;\; 3,699 + 2,228 \cdot 4,88) = (-7,173;\; 14,572)$

для b_1 $(0,9339 - 2,228 \cdot 0,039;\; 0,9339 + 2,228 \cdot 0,039) = (0,8470;\; 1,021)$

Рассчитаем границы интервала, в котором будет сосредоточено 95% возможных объемов потребления при неограниченно большом числе наблюдений и уровне дохода $X = 160$ (формула 2.21):

$$3,699 + 0,9339 \cdot 160 \pm 2,228 \cdot 1,88 \cdot \sqrt{1 + \frac{1}{12} + \frac{(125,25 - 160)^2}{2102,1875}}.$$

Таким образом, этот интервал имеет вид: (147,4898; 158,7082).

Рассчитаем коэффициент детерминации: (2.22):

$$R^2 = 1 - \frac{35,3}{2108,6668} = 0,983.$$

Столь высокое значение коэффициента детерминации свидетельствует о высоком общем качестве построенного уравнения регрессии.

2. Множественная линейная регрессия

2.1. Определение параметров уравнения регрессии

На любой экономический показатель чаще всего оказывает влияние не один, а несколько факторов. В этом случае вместо парной регрессии $M(Y|X) = f(x)$ рассматривается *множественная регрессия*

$$M(Y|x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m) \quad (2.1)$$

Задача оценки статистической взаимосвязи переменных Y и $X_1, X_2, \ldots, X_m$ формулируется аналогично случаю парной регрессии. Уравнение множественной регрессии может быть представлено в виде:

$$Y = f(\beta, X) + \varepsilon, \quad (2.2)$$

где $X = (X_1, X_2, \ldots, X_m)$ - вектор независимых (объясняющих) переменных; β - вектор параметров (подлежащих определению); ε - случайная ошибка (отклонение); Y - зависимая (объясняемая) переменная.

Рассмотрим самую употребляемую и наиболее простую из моделей множественной регрессии – модель множественной линейной регрессии.

Теоретическое линейное уравнение регрессии имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon \quad (2.3)$$

или для индивидуальных наблюдений $i, i = 1,2,\ldots,n$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i. \quad (2.4)$$

Здесь $\beta = (\beta_0, \beta_1, \ldots, \beta_m)$ - вектор размерности $(m+1)$ неизвестных параметров. $\beta_j, j = 1,2,\ldots,m$ называется j-тым теоретическим коэффициентом регрессии (частичным коэффициентом регрессии). Он характеризует чувствительность величины Y к изменению величины X_j, т.е. отражает влияние на условное математическое ожидание $M(Y|x_1, x_2, \ldots, x_m)$ зависимой переменной Y объясняющей переменной X_j при условии, что все другие объясняющие переменные модели остаются постоянными. β_0 - свободный член, определяющий Y в случае, когда все объясняющие переменные X_j равны нулю.

После выбора линейной функции в качестве модели зависимости необходимо оценить параметры регрессии. Пусть имеется n наблюдений вектора объясняющих переменных $X = (X_1, X_2, \ldots, X_m)$ и зависимой переменной Y:

$$(x_{i1}, x_{i2}, \ldots, x_{im}, y_i), i = 1,2, \ldots, n.$$

Для того, чтобы однозначно можно было бы решить задачу отыскания параметров $\beta_0, \beta_1, \ldots, \beta_m$ (т.е. найти некоторый наилучший вектор β), должно выполняться неравенство $n \geq m + 1$. Если это неравенство не будет выполняться, то существует бесконечно много различных векторов параметров, при которых линейная формула связи между X и Y будет абсолютно точно соответствовать имеющимся наблюдениям.

Например, для однозначного определения оценок параметров уравнения регрессии $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ достаточно иметь выборку из трех наблюдений $(x_{i1}, x_{i2}, x_{i3}, y_i), i = 1,2,3$. В этом случае найденные значения параметров $\beta_0, \beta_1, \beta_2$ определяют такую плоскость $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ в трехмерном пространстве, которая пройдет именно через три точки. С другой стороны, добавление в выборку к имеющимся трем наблюдениям еще одного приведет к тому, что четвертая точка $(x_{41}, x_{42}, x_{43}, y_4)$ практически наверняка будет лежать вне построенной плоскости, что потребует определенной переоценки параметров.

Число $\nu = n - m - 1$ называется *числом степеней свободы*. Если число степеней свободы невелико, то статистическая надежность оцениваемой формулы невысока. Например, вероятность верного вывода (получения более точных оценок) по трем наблюдениям существенно ниже, чем по тридцати. Считается, что при оценивании множественной линейной регрессии для обеспечения статистической надежности требуется, чтобы число наблюдений по крайней мере в три раза превосходило число оцениваемых параметров.

Самым распространенным методом оценки параметров уравнения множественной регрессии является метод наименьших квадратов.

Предпосылки МНК:

1. Математическое ожидание случайного отклонения ε_i равно нулю для всех наблюдений: $M(\varepsilon_i)=0, i=1,2,\ldots,n$.
2. Постоянство дисперсии отклонений (гомоскедастичность): $D(\varepsilon_i)=D(\varepsilon_j)=\sigma^2$ для любых наблюдений *i* и *j*.
3. Отсутствие автокорреляции: случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i\neq j$.
4. Случайное отклонение должно быть независимо от объясняющих переменных.
5. Модель является линейной относительно параметров.
6. Отсутствие мультиколлинеарности: между объясняющими переменными отсутствует строгая (сильная) линейная зависимость.
7. Ошибки $\varepsilon_i, i=1,2,\ldots,n$ имеют нормальное распределение. Выполнимость данной предпосылки важна для проверки статистических гипотез и построения интервальных оценок.

Как и в случае парной регрессии, истинные значения параметров β_j по выборке получить невозможно. В этом случае вместо теоретического уравнения регрессии оценивается эмпирическое уравнение регрессии:

$$Y=b_0+b_1X_1+b_2X_2+\ldots+b_mX_m+e \qquad (2.5)$$

Здесь $b_0,b_1,\ldots,b_m$ - оценки теоретических значений $\beta_0,\beta_1,\ldots,\beta_m$ коэффициентов регрессии (эмпирические коэффициенты регрессии); e - оценка отклонения ε. Для индивидуальных наблюдений имеем:

$$y_i=b_0+b_1x_{i1}+b_2x_{i2}+\ldots+b_mx_{im}+e_i \qquad (2.6)$$

При выполнении предпосылок МНК относительно ошибок ε_i оценки $b_0,b_1,\ldots,b_m$ параметров $\beta_0,\beta_1,\ldots,\beta_m$ множественной линейной регрессии по МНК являются несмещенными, эффективными и состоятельными.

На основании (2.6): $e_i=y_i-b_0-b_1x_{i1}-\ldots-b_mx_{im}$. (2.7)

Тогда по методу наименьших квадратов для нахождения оценок $b_0,b_1,\ldots,b_m$ минимизируется следующая функция:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + \sum_{j=1}^{m} b_j x_{ij}))^2 \ . \qquad (2.8)$$

Необходимым условием минимизации функции Q является равенство нулю всех ее частных производных по b_j, т.е.:

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - (b_0 + \sum_{j=1}^{m} b_j x_{ij})) \\ \dfrac{\partial Q}{\partial b_j} = -2\sum_{i=1}^{n} (y_i - (b_0 + \sum_{j=1}^{m} b_j x_{ij}))x_{ij},\ j = 1,2,\ldots,m \end{cases} . \qquad (2.9)$$

Приравнивая их к нулю, получаем систему $m+1$ линейных уравнений с $m+1$ неизвестными. Такая система обычно имеет единственное решение и называется системой нормальных уравнений. Ее решение в явном виде наиболее наглядно представимо в векторно-матричной форме.

2.2. Расчет коэффициентов множественной линейной регрессии

Данные наблюдений и соответствующие коэффициенты в матричной форме выглядят следующим образом:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}, \quad B = \begin{pmatrix} b_0 \\ b_1 \\ \ldots \\ b_m \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}.$$

Здесь $Y - n$-мерный вектор-столбец наблюдений зависимой переменной X; X - матрица размерности $n \times (m+1)$, в которой i-тая строка $(i = 1,2,\ldots,n)$ представляет наблюдение вектора значений независимых переменных $X_1, X_2, \ldots, X_m$; единица соответствует переменной при свободном члене b_0; B - вектор-столбец размерности $(m+1)$ параметров уравнения регрессии; e - вектор-столбец размерности n отклонений выборочных (реальных) значений y_i зависимой переменной Y от значений $\hat{y}_i$, получаемых по уравнению регрессии

$$\widehat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im}. \qquad (2.10)$$

Функция $Q = \sum_{i=1}^{n} e_i^2$ в матричной форме представима как произведение вектор-строки $e^T = (e_1, e_2, \ldots, e_n)$ на вектор-столбец e. Вектор-столбец e может быть в свою очередь представлен в следующем виде:

$$e = Y - XB. \qquad (2.11)$$

Отсюда:

$$Q = e^T e = (Y - XB)^T (Y - XB) = Y^T Y - B^T X^T Y - Y^T XB + B^T X^T XB =$$
$$= Y^T Y - 2B^T X^T Y + B^T X^T XB$$

Здесь e^T, B^T, X^T, Y^T - векторы и матрицы, транспонированные к e, B, X, Y соответственно. При выводе формулы использовались следующие известные соотношения линейной алгебры:

$$(Y - XB)^T = Y^T - (XB)^T; (XB)^T = B^T X^T; B^T X^T Y = Y^T XB$$

Необходимым условием экстремума функции Q является равенство нулю ее частных производных $\frac{\partial Q}{\partial b_j}$ по всем параметрам $b_j, j = 1,2,\ldots,m$.

Вектор-столбец $\frac{\partial Q}{\partial B}$ частных производных в матричном виде имеет следующий вид:

$$\frac{\partial Q}{\partial B} = -2X^T Y + 2(X^T X)B. \qquad (2.12)$$

Рассмотрим более подробно нахождение $\frac{\partial Q}{\partial B}$. Очевидно, что

$$\frac{\partial Q}{\partial B} = \frac{\partial (Y^T Y)}{\partial B} - 2\frac{\partial (B^T X^T Y)}{\partial B} + \frac{\partial (B^T X^T XB)}{\partial B}.$$

$Y^T Y$ от B не зависит, следовательно $\frac{\partial (Y^T Y)}{\partial B} = 0$.

Обозначим вектор-столбец $X^T Y$ размерности $(m+1)$ через R. Тогда $B^T X^T Y = B^T R = \sum_{j=0}^{m} b_j r_{j+1}$, где r_{j+1} - соответствующий элемент вектора R. Поэтому $\frac{\partial (B^T R)}{\partial B} = R = X^T Y$.

Обозначим матрицу $X^T X$ размерности $(m+1)\times(m+1)$ через Z. Тогда

$$S = B^T Z B = \left((b_0, b_1, \dots, b_m) \cdot \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1,m+1} \\ z_{21} & z_{22} & \dots & z_{2,m+1} \\ \dots & \dots & \dots & \dots \\ z_{m+1,1} & z_{m+1,2} & \dots & z_{m+1,m+1} \end{pmatrix} \right) \cdot \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_m \end{pmatrix} =$$

$$= \left(\sum_{i=0}^{m} b_i z_{i+1,1}, \sum_{i=0}^{m} b_i z_{i+1,2}, \dots, \sum_{i=0}^{m} b_i z_{i+1,m+1} \right) \cdot \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_m \end{pmatrix} = \sum_{j=0}^{m} b_j \sum_{i=0}^{m} b_i z_{i+1,j+1} =$$

$$= \sum_{j=0}^{m} \sum_{i=0}^{m} b_j b_i z_{i+1,j+1}.$$

Следовательно, частная производная $\frac{\partial S}{\partial b_j} = 2\sum_{i=0}^{m} b_i z_{i+1,j+1}$.

В результате имеем $\frac{\partial S}{\partial B} = 2(X^T X)B$.

Следовательно, формула (2.12) справедлива. Приравняв $\frac{\partial Q}{\partial B}$ к нулю, получаем:

$$-2X^T Y + 2(X^T X)B = 0 \Rightarrow$$

$$X^T Y = (X^T X)B \Rightarrow \tag{2.13}$$

$$B = (X^T X)^{-1} X^T Y \tag{2.14}$$

Здесь $(X^T X)^{-1}$ - матрица, обратная к $X^T X$.

Полученные общие соотношения справедливы для уравнений регрессии с произвольным количеством m объясняющих переменных. Проанализируем полученные результаты для случаев:

1. $m=1$. Для парной регрессии $Y=b_0+b_1X+e$ имеем:

$$Y=\begin{pmatrix} y_1\\ y_2\\ \dots\\ y_n \end{pmatrix},\quad X=\begin{pmatrix} 1 & x_1\\ 1 & x_2\\ \dots & \dots\\ 1 & x_n \end{pmatrix},\quad B=\begin{pmatrix} b_0\\ b_1 \end{pmatrix},\quad e=\begin{pmatrix} e_1\\ e_2\\ \dots\\ e_n \end{pmatrix}.$$

Из (2.11) следует: $e=Y-XB$, т.е. $\begin{pmatrix} e_1\\ e_2\\ \dots\\ e_n \end{pmatrix}=\begin{pmatrix} y_1\\ y_2\\ \dots\\ y_n \end{pmatrix}-\begin{pmatrix} 1 & x_1\\ 1 & x_2\\ \dots & \dots\\ 1 & x_n \end{pmatrix}\cdot\begin{pmatrix} b_0\\ b_1 \end{pmatrix}.$

$$Z=(X^TX)^{-1}=\begin{pmatrix} \dfrac{\sum x_i^2}{n\sum x_i^2-(\sum x_i)^2} & -\dfrac{\sum x_i}{n\sum x_i^2-(\sum x_i)^2}\\ -\dfrac{\sum x_i}{n\sum x_i^2-(\sum x_i)^2} & \dfrac{n}{n\sum x_i^2-(\sum x_i)^2} \end{pmatrix}$$

$$R=X^TY=\begin{pmatrix} 1 & 1 & \dots & 1\\ x_1 & x_2 & \dots & x_n \end{pmatrix}\cdot\begin{pmatrix} y_1\\ y_2\\ \dots\\ y_n \end{pmatrix}=\begin{pmatrix} \sum y_i\\ \sum x_iy_i \end{pmatrix}.$$

Из (2.14) следует $B=(X^TX)^{-1}X^TY=$

$$=\begin{pmatrix} \dfrac{\sum x_i^2\sum y_i}{n\sum x_i^2-(\sum x_i)^2}-\dfrac{\sum x_i\sum x_iy_i}{n\sum x_i^2-(\sum x_i)^2}\\ -\dfrac{\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}+\dfrac{n\sum x_iy_i}{n\sum x_i^2-(\sum x_i)^2} \end{pmatrix}=$$

$$=\begin{pmatrix} \dfrac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} \\ \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \end{pmatrix} = \begin{pmatrix} \bar{y} - b_1\bar{x} \\ b_1 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \qquad (2.15)$$

2. $m = 2$. Рассуждая аналогично, можно вывести формулы определения коэффициентов регрессии для уравнения с двумя объясняющими переменными. Соотношение (2.13) в этом случае имеет вид системы трех линейных уравнений с тремя неизвестными b_0, b_1, b_2:

$$\begin{cases} \sum y_i = nb_0 + b_1 \sum x_{i1} + b_2 \sum x_{i2} \\ \sum x_{i1} y_i = b_0 \sum x_{i1} + b_1 \sum x_{i1}^2 + b_2 \sum x_{i1} x_{i2} \\ \sum x_{i2} y_i = b_0 \sum x_{i2} + b_1 \sum x_{i1} x_{i2} + b_2 \sum x_{i2}^2 \end{cases} . \qquad (2.16)$$

Решение данной системы имеет вид:

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$b_1 = \frac{\sum(x_{i1}-\bar{x}_1)(y_i-\bar{y})\cdot\sum(x_{i2}-\bar{x}_2)^2 - \sum(x_{i2}-\bar{x}_2)(y_i-\bar{y})\cdot\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum(x_{i1}-\bar{x}_1)^2 \sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}$$

$$b_2 = \frac{\sum(x_{i2}-\bar{x}_2)(y_i-\bar{y})\cdot\sum(x_{i1}-\bar{x}_1)^2 - \sum(x_{i1}-\bar{x}_1)(y_i-\bar{y})\cdot\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum(x_{i1}-\bar{x}_1)^2 \sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}$$

(2.17)

2.3. Анализ качества эмпирического уравнения множественной линейной регрессии

Построение эмпирического уравнения регрессии является начальным этапом эконометрического анализа. Первое же построенное по выборке уравнение регрессии очень редко является удовлетворительным по тем или иным характеристикам. Поэтому следующей важнейшей оценкой является проверка качества уравнения регрессии. В эконометрике принята устоявшаяся схема такой проверки, которая проводится по следующим направлениям:

- проверка статистической значимости коэффициентов уравнения регрессии
- проверка общего качества уравнения регрессии
- проверка свойств данных, выполнимость которых предполагалась при оценивании уравнения (проверка выполнимости предпосылок МНК).

Прежде чем проводить анализ качества уравнения регрессии, необходимо определить дисперсии и стандартные ошибки коэффициентов, а также интервальные оценки коэффициентов.

Выборочные дисперсии эмпирических коэффициентов регрессии можно определить следующим образом:

$$S_{bj}^2 = S^2 z_{jj}^{/} = \frac{\sum e_i^2}{n-m-1} z_{jj}^{/},\ j=1,2,\ldots,m. \qquad (2.18)$$

Здесь $z_{jj}^{/}$ - j-тый диагональный элемент матрицы $Z^{-1}=(X^TX)^{-1}$.

При этом:

$$S^2 = \frac{\sum e_i^2}{n-m-1}, \qquad (2.19)$$

где m - количество объясняющих переменных модели. Иногда в формуле (2.19) знаменатель представляют в виде $n-m-1=n-k$, подразумевая под k число параметров модели (подлежащих определению коэффициентов регрессии).

В частности, для уравнения $\hat{Y}=b_0+b_1X_1+b_2X_2$ с двумя объясняющими переменными используются следующие формулы:

$$S_{b0}^2 = \left[\frac{1}{n} + \frac{\bar{x}_1^2\sum(x_{i2}-\bar{x}_2)^2 + \bar{x}_2^2\sum(x_{i1}-\bar{x}_1)^2 - 2\bar{x}_1\bar{x}_2\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum(x_{i1}-\bar{x}_1)^2\sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}\right]\cdot S^2,$$

$$S_{b1}^2 = \frac{\sum(x_{i2}-\bar{x}_2)^2}{\sum(x_{i1}-\bar{x}_1)^2\sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}\cdot S^2 \Rightarrow$$

$$S_{b1}^2 = \frac{S^2}{\sum(x_{i1}-\bar{x}_1)^2\cdot(1-r_{12}^2)},$$

$$S_{b2}^2 = \frac{\sum(x_{i1}-\bar{x}_1)^2}{\sum(x_{i1}-\bar{x}_1)^2\sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}\cdot S^2 \Rightarrow$$

$$S_{b2}^2 = \frac{S^2}{\sum(x_{i2} - \bar{x}_2)^2 \cdot (1 - r_{12}^2)},$$

$$S_{b0} = \sqrt{S_{b0}^2}, \quad S_{b1} = \sqrt{S_{b1}^2}, \quad S_{b2} = \sqrt{S_{b2}^2} \qquad (2.20)$$

Здесь r_{12} - выборочный коэффициент корреляции между объясняющими переменными X_1 и X_2; S_{bj} - стандартная ошибка коэффициента регрессии; S - стандартная ошибка регрессии (несмещенная оценка).

По аналогии с парной регрессией после определения точечных оценок b_j коэффициентов β_j ($j = 1, 2, \ldots, m$) теоретического уравнения регрессии могут быть рассчитаны интервальные оценки указанных коэффициентов. Доверительный интервал, накрывающий с надежностью $(1-\alpha)$ неизвестное значение параметра β_j, определяется:

$$\left(b_j - t_{\frac{\alpha}{2}, n-m-1} S(b_j);\; b_j + t_{\frac{\alpha}{2}, n-m-1} S(b_j) \right) \qquad (2.21)$$

Проверка статистической значимости коэффициентов уравнения регрессии.

Как и в случае парной регрессии, статистическая значимость коэффициентов множественной линейной регрессии с m объясняющими переменными проверяется на основе t-статистики:

$$t = \frac{b_j}{S_{bj}}, \qquad (2.22)$$

имеющей в данном случае распределение Стьюдента с числом степеней свободы $\nu = n - m - 1$. При требуемом уровне значимости наблюдаемое значение t-статистики сравнивается с критической точной $t_{\frac{\alpha}{2}, n-m-1}$ распределения Стьюдента.

В случае, если $|t| > t_{\frac{\alpha}{2}, n-m-1}$, то статистическая значимость соответствующего коэффициента регрессии подтверждается. Это означает, что фактор X_j линейно связан с зависимой переменной Y. Если же установлен факт незначимости коэффициента b_j, то рекомендуется исключить из уравнения переменную X_j. Это не приведет к существенной потере качества модели, но сделает ее более конкретной.

При оценке значимости коэффициентов линейной регрессии на начальном этапе также можно использовать «грубое» правило, позволяющее не прибегать к таблицам и рассмотренное в главе 1.3.

Проверка общего качества уравнения регрессии.

Для этой цели, как и в случае парной регрессии, используется *коэффициент детерминации* R^2:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \tag{2.23}$$

Справедливо соотношение $0 \le R^2 \le 1$. Чем ближе этот коэффициент к единице, тем больше уравнение регрессии объясняет поведение Y.

Для множественной регрессии коэффициент детерминации является неубывающей функцией числа объясняющих переменных. Добавление новой объясняющей переменной никогда не уменьшает значение R^2, так как каждая последующая переменная может лишь дополнить, но никак не сократить информацию, объясняющую поведение зависимой переменной.

Иногда при расчете коэффициента детерминации для получения несмещенных оценок в числителе и знаменателе вычитаемой из единицы дроби делается поправка на число степеней свободы, т.е. вводится так называемый *скорректированный (исправленный) коэффициент детерминации*:

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 /(n-m-1)}{\sum (y_i - \bar{y})^2 /(n-1)} \tag{2.24}$$

Соотношение (2.24) может быть представлено в следующем виде:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1} \tag{2.25}$$

Из (2.25) очевидно, что $\bar{R}^2 < R^2$ для $m > 1$. С ростом значения m скорректированный коэффициент детерминации растет медленнее, чем обычный. Очевидно, что $\bar{R}^2 = R^2$ только при $R^2 = 1$. $\bar{R}^2$ может принимать отрицательные значения.

Доказано, что $\bar{R}^2$ увеличивается при добавлении новой объясняющей переменной тогда и только тогда, когда t-статистика для этой переменной по модулю больше единицы. Поэтому добавление в модель новых объясняющих переменных осуществляется до тех пор, пока растет скорректированный коэффициент детерминации.

Рекомендуется после проверки общего качества уравнения регрессии провести анализ его статистической значимости. Для этого используется F -статистика:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} \tag{2.26}$$

Из формулы (2.26) очевидно, что показатели F и R^2 равны или не равен нулю одновременно. Если $F=0$, то $R^2=0$, следовательно, величина Y линейно не зависит от $X_1, X_2, \ldots, X_m$.

Расчетное значение F сравнивается с критическим $F_{кр}$. $F_{кр}$, исходя из требуемого уровня значимости α и числами степеней свободы $v_1 = m$ и $v_2 = n-m-1$, определяется на основе распределения Фишера (приложение 2). Если $F > F_{кр}$, то R^2 статистически значим.

Проверка выполнимости предпосылок МНК. Статистика Дарбина-Уотсона.

Статистическая значимость коэффициентов регрессии и близкое к единице значение коэффициента детерминации R^2 не гарантируют высокое качество уравнения регрессии. Поэтому следующим этапом проверки качества уравнения регрессии является проверка выполнимости предпосылок МНК. Причины и последствия невыполнимости этих предпосылок, методы корректировки регрессионных моделей будут рассмотрены в последующих главах. В данном параграфе рассмотрим популярную в регрессионном анализе статистику Дарбина-Уотсона.

При статистическом анализе уравнения регрессии на начальном этапе часто проверяют выполнимость одной предпосылки: условия статистической независимости отклонений между собой. При этом проверяется некоррелированность соседних величин $e_i, i = 1,2,\ldots,n$ (2.7).

Для анализа коррелированности отклонений используют статистику Дарбина-Уотсона:

$$DW = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2} \tag{2.27}$$

Критические значения d_1 и d_2 определяются на основе специальных таблиц (приложение 3) для требуемого уровня значимости α, числа наблюдений n и количества объясняющих переменных m.

Выводы об отсутствии автокорреляции остатков осуществляются по следующей схеме:

Не обращаясь к таблицам, можно пользоваться «грубым» правилом и считать, что автокорреляция остатков отсутствует, если $1{,}5 < DW < 2{,}5$. Для более надежного вывода целесообразно обращаться к табличным значениям.

Пример. Анализируется объем S сбережений домохозяйства за 10 лет. Предполагается, что его размер s_t в текущем году t зависит от величины y_t располагаемого дохода Y и от величины z_t реальной процентной ставки Z. Статистические данные представлены в таблице:

Год	80	81	82	83	84	85	86	87	88	89	90
Y, тыс.у.е.	100	110	140	150	160	160	180	200	230	250	260
Z, %	2	2	3	2	3	4	4	3	4	5	5
S, тыс.у.е.	20	25	30	30	35	38	40	38	44	50	55

Средние значения исходных данных равны: $\bar{y} = 176{,}3636$, $\bar{z} = 3{,}3636$, $\bar{s} = 36{,}8182$.

Представим требующиеся для построения модели множественной регрессии и проведения дальнейшего анализа промежуточные вычисления в таблице:

Год	$(y_i - \bar{y})^2$	$(z_i - \bar{z})^2$	$(s_i - \bar{s})^2$	$(y_i - \bar{y}) \cdot (z_i - \bar{z})$	$(y_i - \bar{y}) \cdot (s_i - \bar{s})$	$(z_i - \bar{z}) \cdot (s_i - \bar{s})$
80	5831,4050	1,8595	282,8512	104,1322	1284,2975	22,9339
81	4404,1322	1,8595	139,6694	90,4959	784,2975	16,1157
82	1322,3140	0,1322	46,4876	13,2231	247,9339	2,4793
83	695,0413	1,8595	46,4876	35,9504	179,7521	9,2975
84	267,7686	0,1322	3,3058	5,9504	29,7521	0,6612
85	267,7686	0,4050	1,3967	-10,4132	-19,3388	0,7521
86	13,2231	0,4050	10,1240	2,3140	11,5702	2,0248
87	558,6777	0,1322	1,3967	-8,5950	27,9339	-0,4298
88	2876,8595	0,4050	51,5785	34,1322	385,2066	4,5702

89	5422,3140	2,6777	173,7603	120,4959	970,6612	21,5702
90	6995,0413	2,6777	330,5785	136,8595	1520,6612	29,7521
$\sum$	28654,5455	12,5455	1087,6364	524,5455	5422,7273	109,7273

Расчет коэффициентов уравнения регрессии производится по формулам (2.17):

$$b_0 = 36{,}8182 - 0{,}124189 \cdot 176{,}3636 - 3{,}553796 \cdot 3{,}3636 = 2{,}962233$$

$$b_1 = \frac{5422{,}7273 \cdot 12{,}5455 - 109{,}7273 \cdot 524{,}5455}{28654{,}5455 \cdot 12{,}5455 - (524{,}5455)^2} = \frac{10473{,}8639}{84337{,}619} = 0{,}124189$$

$$b_2 = \frac{109{,}7273 \cdot 28654{,}5455 - 5422{,}7273 \cdot 524{,}5455}{28654{,}5455 \cdot 12{,}5455 - (524{,}5455)^2} = \frac{299718{,}7075}{84337{,}619} = 3{,}553796$$

Таким образом, эмпирическое уравнение регрессии имеет вид:

$$s_t = 2{,}962233 + 0{,}124189\, y_t + 3{,}553796\, z_t$$

Подставляя соответствующие значения y_t и z_t в эмпирическое уравнение регрессии, получаем $\hat{s}_t$. Расчет отклонений e_i реальных значений от модельных представлен в таблице:

Год	S	$\hat{S}$	e_i	e_i^2	$e_i - e_{i-1}$	$(e_i - e_{i-1})^2$
80	20	22,489	-2,48873	6,19375	-	-
81	25	23,731	1,26939	1,61134	3,75811	14,1234
82	30	31,01	-1,01008	1,02026	-2,27947	5,19597
83	30	28,698	1,30183	1,69475	2,31191	5,34491
84	35	33,494	1,50614	2,26845	0,20431	0,04174
85	38	37,048	0,95234	0,90696	-0,55380	0,30669
86	40	39,531	0,46856	0,21955	-0,48378	0,23404
87	38	38,461	-0,46142	0,21291	-0,92998	0,86487
88	44	45,741	-1,74089	3,03069	-1,27947	1,63703
89	50	51,778	-1,77846	3,16293	-0,03758	0,00141
90	55	53,02	1,97965	3,91900	3,75811	14,1234
сумма	405	405	≈ 0	24,24060	4,46837	41,8734
	36,8182	36,8182				

Рассчитаем дисперсию регрессии по формуле (2.19):

$$S^2 = \frac{\sum e_i^2}{n-m-1} = \frac{24{,}2406}{11-2-1} = 3{,}03.$$

Определим дисперсии и стандартные ошибки коэффициентов по (2.20):

$$S_{b0}^2 = \left[\frac{1}{11} + \frac{(176{,}3636)^2 \cdot 12{,}5455 + (3{,}3636)^2 \cdot 28654{,}5455 - 2 \cdot 176{,}3636 \cdot 3{,}3636 \cdot 524{,}5455}{28654{,}5455 \cdot 12{,}5455 - (524{,}5455)^2}\right] \cdot 3{,}03$$

$$S_{b0} = \sqrt{S_{b0}^2} = \sqrt{3{,}5832} = 1{,}8929$$

$$S_{b1}^2 = \frac{12{,}5455}{28654{,}5455 \cdot 12{,}5455 - (524{,}5455)^2} \cdot 3{,}03 = 0{,}00054$$

$$S_{b1} = \sqrt{S_{b1}^2} = \sqrt{0{,}00054} = 0{,}0212$$

$$S_{b2}^2 = \frac{28654{,}5455}{28654{,}5455 \cdot 12{,}5455 - (524{,}5455)^2} \cdot 3{,}03 = 1{,}0294$$

$$S_{b2} = \sqrt{S_{b2}^2} = \sqrt{1{,}0294} = 1{,}0146$$

Рассчитаем по формуле (2.22) соответствующие t-статистики:

$t_{b0} = 1{,}565$, $t_{b1} = 5{,}858$, $t_{b2} = 3{,}503$.

Проверим статистическую значимость коэффициентов на основе распределения Стьюдента. По таблице, приведенной в приложении 1, определим критические значения с уровнем значимости $\alpha = 0{,}05$: $t_{кр} = t_{\frac{\alpha}{2};n-m-1} = t_{0{,}025;\ 8} = 2{,}306$. Таким образом, $\left|t_{b0}\right| < t_{кр}$, $\left|t_{b1}\right| > t_{кр}$, $\left|t_{b2}\right| > t_{кр}$.

По формуле (2.21) определим 95%-е интервальные оценки коэффициентов:

для β_0: (2,962233-2,306*1,8929; 2,962233+2,306*1,8929), т.е. (-1,4028; 7,3273)

для β_1: (0,124189-2,306*0,0212; 0,124189+2,306*0,0212), т.е. (0,0753; 0,1731)

для β_2: (3,553796-2,306*1,0146; 3,553796+2,306*1,0146), т.е. (1,2141; 5,8935)

Рассчитаем коэффициент детерминации по формуле (2.23):

$$R^2 = 1 - \frac{24{,}2406}{1087{,}6364} = 0{,}9777.$$

Анализ статистической значимости коэффициента детерминации осуществляется на основе F -статистики (2.26):

$$F = \frac{0,9777}{1-0,9777} \cdot \frac{11-2-1}{2} = 175,3722$$

Определим по приложению 2 критическую точку распределения Фишера: $F_{кр} = F_{0,05;2;8} = 4,46$ с 95%-ой вероятностью. Очевидно, что 175,3722>4,46, следовательно, коэффициент детерминации статистически значим, т.е. совокупное влияние переменных Y и Z на переменную S существенно.

На основе проведенных рассуждений и вычислений можно заключить, что построенное уравнение регрессии объясняет 97,77% разброса зависимой переменной S. Однако для уверенности и обоснованности (чтобы исключить автокорреляцию) проведем исследование с помощью статистики Дарбина-Уотсона.

Рассчитаем статистику по формуле (2.27):

$$DW = \frac{41,8734}{24,2406} = 1,7274 .$$

По приложению 3 определим критические точки для уровня значимости 0,05 и числа наблюдений 11: $d_1 = 0,658; d_2 = 1,604$.

Таким образом, $1,604 < DW < 2,396$, т.е. ($d_2 < DW < 4 - d_2$), следовательно, имеются основания считать, что автокорреляция отсутствует. Это является одним из подтверждений высокого качества модели.

По всем статистическим показателям модель может быть признана удовлетворительной.

3. Нелинейная регрессия

Во многих практических случаях моделирование экономических зависимостей линейными уравнениями дает вполне удовлетворительный результат и может использоваться для анализа и прогнозирования. Однако многие экономические зависимости не являются линейными по своей сути, и поэтому их моделирование линейными уравнениями регрессии не даст положительного результата.

Построение и анализ нелинейных моделей имеют свою специфику. Рассмотрим нелинейные модели, допускающими сведение их к линейным. Такие модели называют *линейные относительно параметров модели*. Будем рассматривать модели парной регрессии с целью простоты изложения и графической иллюстрации.

3.1. Логарифмические (лог-линейные) модели

Рассмотрим модель парной регрессии. Пусть некоторая экономическая зависимость моделируется формулой:

$$Y = AX^{\beta} \tag{3.1}$$

где A, β - параметры модели (константы, подлежащие определению).

Это функция может отражать зависимость спроса Y на благо от его цены X (тогда $\beta < 0$); зависимость спроса Y на благо от дохода X (тогда $\beta > 0$); зависимость объема выпуска Y от использования ресурса X (тогда $0 < \beta < 1$).

Для анализа функций вида (3.1) используется логарифмирование по экспоненте. Прологарифмировав обе части, имеем:

$$\ln Y = \ln A + \beta \ln X \tag{3.2}$$

Заменим $\ln A$ на β_0:

$$\ln Y = \beta_0 + \beta \ln X \tag{3.3}$$

С целью статистической оценки коэффициентов добавим в модель случайную погрешность ε, в результате получим *двойную логарифмическую модель* (и зависимая переменная, и объясняющая переменная заданы в логарифмическом виде):

$$\ln Y = \beta_0 + \beta \ln X + \varepsilon \tag{3.4}$$

Введем замены $Y^* = \ln Y$ и $X^* = \ln X$, получаем:

$$Y^* = \beta_0 + \beta X^* + \varepsilon \tag{3.5}$$

Модель (3.5) является линейной моделью, подробно рассмотренной ранее. Коэффициент β определяет *эластичность* переменной Y по переменной X, т.е. процентное изменение Y для данного процентного изменения X. Действительно, продифференцировав обе части (3.4), получаем:

$$\frac{1}{Y} \cdot \frac{dY}{dX} = \beta \cdot \frac{1}{X} \Rightarrow \beta = \frac{dY/Y}{dX/X} = \frac{\text{относительное изменение} Y}{\text{относительное изменение } X} = \text{эластичность}$$

В случае парной регрессии обоснованность использования логарифмической модели проверить достаточно просто. Для этого определяются точки $(\ln x_i; \ln y_i)$, $i = 1, 2, \ldots, n$, которые затем наносятся на корреляционное поле. Если их расположение соответствует прямой линии, то использование данной модели обоснованно.

При большем числе переменных:

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \ldots + \beta_m \ln X_m + \varepsilon \tag{3.6}$$

Коэффициенты $\beta_j,\ j = 1,2,\ldots,m$ являются эластичностями переменной Y по переменной X_j.

3.2. Полулогарифмические модели

Полулогарифмическими моделями являются модели вида (в случае парной регрессии):

$$\ln Y = \beta_0 + \beta X + \varepsilon \tag{3.7}$$

$$Y = \beta_0 + \beta \ln X + \varepsilon \tag{3.8}$$

Лог-линейная модель.

Рассмотрим известную в банковском анализе зависимость:

$$Y_t = Y_0 (1 + r)^t \tag{3.9}$$

где Y_0 - первоначальный вклад в банке; r - процентная ставка; Y_t - вклад в банке в момент времени t.

Прологарифмировав обе части (3.9): $\ln Y_t = \ln Y_0 + t \cdot \ln(1 + r)$ и введя обозначение $\ln Y_0 = \beta_0$, $\ln(1 + r) = \beta$, а также введя случайное слагаемое ε_t, получаем модель вида (3.7):

$$\ln Y_t = \beta_0 + \beta t + \varepsilon_t$$

Полулогарифмическая модель (3.7) легко сводится к линейной модели путем замены $Y^* = \ln Y$, т.е. $Y^* = \beta_0 + \beta t + \varepsilon_t$.

Коэффициент β в (3.7) имеет смысл темпа прироста переменной Y по переменной X. Действительно, продифференцировав обе части (3.7), получаем:

$$\frac{1}{Y} \cdot \frac{dY}{dX} = \beta \Rightarrow \beta = \frac{dY / Y}{dX} = \frac{\textit{относительное изменение}\, Y}{\textit{абсолютное изменение}\, X}$$

Линейно-логарифмическая модель

Линейно-логарифмическая модель (3.8) сводится к линейной модели путем замены: $X^* = \ln X$, т.е.:

$$Y = \beta_0 + \beta X^* + \varepsilon \tag{3.10}$$

Коэффициент β в (3.8) имеет определяет изменение переменной Y вследствие единичного относительного прироста X. Действительно, продифференцировав обе части (3.8), получаем:

$$\frac{dY}{dX} = \beta \cdot \frac{1}{X} \Rightarrow \beta = \frac{dY}{dX / X} = \frac{абсолютное\,изменение\,Y}{относительное\,изменение\,X}$$

3.3. Обратная модель

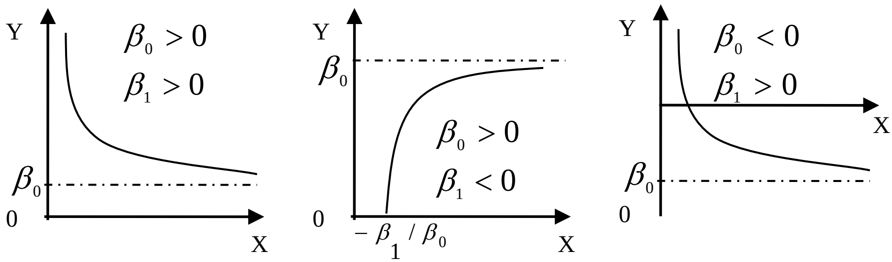
Обратной моделью называется модель вида:

$$Y = \beta_0 + \beta_1 \cdot \frac{1}{X} + \varepsilon \tag{3.11}$$

Она сводится к линейной путем замены: $X^* = \dfrac{1}{X}$. Таким образом:

$$Y = \beta_0 + \beta_1 \cdot X^* + \varepsilon \tag{3.12}$$

В зависимости от знаков β_0, β_1 возможны случаи:



а б в

График (а) может отражать зависимость между объемом выпуска X и средними фиксированными издержками Y. График (б) может отражать зависимость между доходом X и спросом на блага Y. График (в) может отражать зависимость между уровнем безработицы X в процентах и процентным изменение заработной платы Y (кривая Филлипса).

3.4. Показательная модель

Показательная функция имеет вид:

$$Y = \beta_0 e^{\beta X} \tag{3.13}$$

и сводится к лог-линейной модели путем логарифмирования:

$$\ln Y = \ln \beta_0 + \beta X \tag{3.14}$$

Сведение лог-линейной модели к линейной рассмотрено в главе 3.2.

3.5. Выбор формы модели

Многообразие и сложность экономических процессов предопределяет многообразие моделей, используемых для экономического анализа. Это существенно усложняет процесс нахождения максимально адекватной формулы зависимости. Для случая парной регрессии подбор модели обычно осуществляется на основе расположения наблюдаемых точек на корреляционном поле. Однако нередки ситуации, когда расположение точек приблизительно соответствует нескольким функциям и необходимо из них выявить наилучшую.

На практике неизвестно, какая модель является верной, и зачастую подбирают такую модель, которая наиболее точно соответствует реальным данным. Признаками «хорошей» модели являются:

1. Скупость (простота). Модель должна быть максимально простой. Данное свойство определяется тем фактом, что модель не отражает действительность идеально, а является ее упрощением. Поэтому из двух моделей, приблизительно одинаково отражающих реальность, предпочтение отдается модели, содержащей меньшее число объясняющих переменных.
2. Единственность. Для любого набора статистических данных определяемые коэффициенты должны вычисляться однозначно.
3. Максимально соответствие. Уравнение тем лучше, чем большую часть разброса зависимой переменной оно может объяснить. Поэтому стремятся построить уравнение с максимально возможным скорректированным коэффициентом детерминации $\bar{R}^2$.
4. Согласованность с теорией. Никакое уравнение не может быть признано качественным, если оно не соответствует известным теоретическим предпосылкам.
5. Прогнозные качества. Модель может быть признана качественной, если полученные на ее основе прогнозы подтверждаются реальностью. Другим критерием прогнозных качеств оцененной модели регрессии может служить следующее отношение:

$$V = \frac{S}{\bar{y}}, \tag{3.15}$$

где $S=\sqrt{\dfrac{\sum e_i^2}{n-m-1}}$ - стандартная ошибка регрессии, $\bar{y}$ - среднее значение зависимой переменной. Если V мала (она определяет относительную ошибку прогноза в процентах) и отсутствует автокорреляция остатков, то прогнозные качества модели высоки.

Пример. **А**нализируется индекс потребительских цен Y по объему денежной массы X на основании приведенных в таблице данных. Необходимо построить логарифмическую модель.

Год	Y	X	Год	Y	X
81	65	110	89	95	235
82	68	125	90	100	240
83	72,5	132	91	106,5	245
84	77,5	137	92	112	250
85	82	160	93	115,5	275
86	85,5	177	94	118,5	285
87	88,5	192	95	120	295
88	91	215	96	120,5	320
			97	121	344

Решение:

Логарифмическая модель имеет вид: $Y = AX^{\beta}$. Данная модель сводится к линейной следующим образом: $\ln Y = b_0 + b\ln X$ (глава 3.1). Для определения коэффициентов в этой модели определим логарифмы переменных Y и X, $(\ln X)^2$, $(\ln X)\cdot(\ln Y)$ и представим их в таблице.

Год	Y	X	$\ln Y$	$\ln X$	$(\ln X)^2$	$(\ln X)\cdot(\ln Y)$
81	65	110	4,1744	4,7005	22,0947	19,6218
82	68	125	4,2195	4,8283	23,3125	20,3730
83	72,5	132	4,2836	4,8828	23,8417	20,9160
84	77,5	137	4,3503	4,9200	24,2064	21,4035
85	82	160	4,4067	5,0752	25,7577	22,3649
86	85,5	177	4,4485	5,1761	26,7920	23,0259
87	88,5	192	4,4830	5,2575	27,6413	23,5694
88	91	215	4,5109	5,3706	28,8433	24,2262
89	95	235	4,5539	5,4596	29,8072	24,8625
90	100	240	4,6052	5,4806	30,0370	25,2393

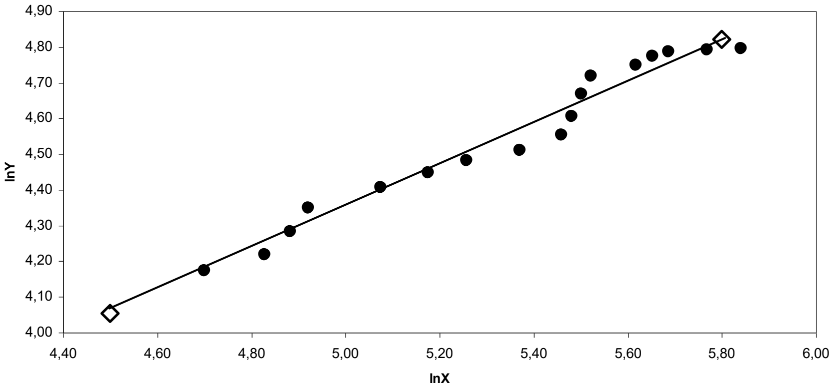
91	106,5	245	4,6681	5,5013	30,2643	25,6806
92	112	250	4,7185	5,5215	30,4870	26,0532
93	115,5	275	4,7493	5,6168	31,5484	26,6759
94	118,5	285	4,7749	5,6525	31,9508	26,9901
95	120	295	4,7875	5,6870	32,3420	27,2265
96	120,5	320	4,7916	5,7683	33,2733	27,6394
97	121	344	4,7958	5,8406	34,1126	28,0103
Сумма	1639	3737	77,3217	90,7392	486,3122	413,8784
Среднее	96,4118	219,8235	4,5483	5,3376	28,6066	24,3458

Затем, по аналогии с примером, приведенным в главе 1, рассчитываются коэффициенты для этой модели следующим образом:

$$\begin{cases} b = \dfrac{\overline{(\ln X)\cdot\ln(Y)} - \overline{\ln X}\cdot\overline{\ln Y}}{\overline{(\ln X)^2} - (\overline{\ln X})^2} = \dfrac{24{,}3458 - 5{,}3376\cdot 4{,}5483}{28{,}6066 - (5{,}3376)^2} = \dfrac{0{,}0688}{0{,}1166} = 0{,}5901 \\ b_0 = \overline{\ln Y} - b\cdot\overline{\ln X} = 4{,}5483 - 0{,}5901\cdot 5{,}3376 = 1{,}3986 \end{cases}$$

Следовательно, модель имеет вид: $\ln Y = 1{,}3986 + 0{,}5901\cdot\ln X$. Если свести данную модель к виду $Y = AX^{\beta}$, то получим: $Y = 4{,}0495\cdot X^{0{,}5901}$ (т.к. $\ln A = b_0 = 1{,}3986$, следовательно, $A = e^{b_0} = 4{,}0495$).

Представим графически корреляционное поле для переменных $\ln Y$ и $\ln X$, а также график рассчитанной модели $\ln Y = 1{,}3986 + 0{,}5901\cdot\ln X$.



После определения коэффициентов модели необходимо проверить значимость коэффициентов уравнения регрессии, определить их интервальные оценки и рассчитать коэффициент детерминации. Расчет проводится аналогично примеру в главе 1 для модели вида $\ln Y = 1{,}3986 + 0{,}5901 \cdot \ln X$.

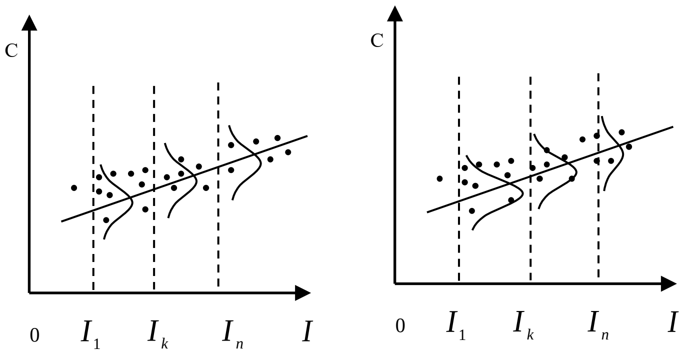
4. Гетероскедастичность

4.1. Суть гетероскедастичности

При практическом проведении регрессионного анализа при помощи метода МНК следует обратить серьезное внимание на проблемы, связанные с выполнимостью свойств случайных отклонений моделей. Свойства оценок коэффициентов регрессии напрямую зависят от свойств случайного члена в уравнении регрессии. Для получения качественных оценок необходимо следить за выполнимостью предпосылок МНК, так как при их нарушении МНК может давать оценки с плохими статистическими свойствами. Одной из ключевых предпосылок МНК является условие постоянства дисперсий случайных отклонений. Выполнимость данной предпосылки называется *гомоскедастичностью* (постоянством дисперсии отклонений), невыполнимость данной предпосылки называется *гетероскедастичностью* (непостоянством дисперсии отклонений).

Случайные отклонения принимают произвольные значения некоторых вероятностных распределений. Но несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть большим либо меньшим, положительным либо отрицательным, не должно быть причины, вызывающей большие отклонения при одних наблюдениях и меньшую при других.

На рисунке приведены два примера линейной регрессии - зависимости потребления C от дохода I : $C = \beta_0 + \beta_1 I + \varepsilon$.



а б

В обоих случаях с ростом дохода растет среднее значение потребления. Но на рисунке (а) дисперсия остается одной и той же дл различных уровней дохода, а на рисунке (б) дисперсия потребления не остается постоянной, а увеличивается с ростом дохода. Фактически это означает, что во втором случае субъекты с большим доходом в среднем потребляют больше, чем субъекты с меньшим доходом, и, кроме того, разброс в их потреблении более существенен для большего уровня дохода. Люди с большим доходом имеют больший простор для его распределения. Реалистичность данной ситуации не вызывает сомнений.

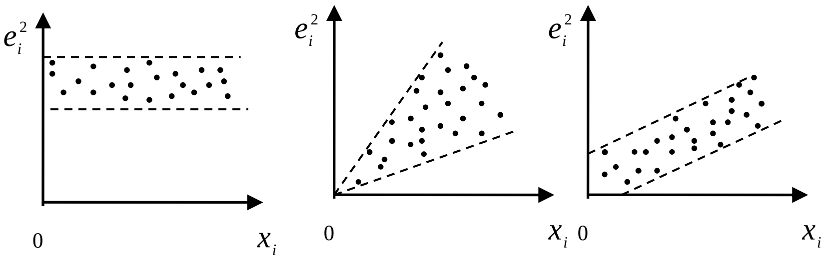
Гетероскедастичность приводит к тому, что выводы, полученные на основе t- и F- статистик, а также интервальные оценки будут ненадежными.

4.2.Обнаружение гетероскедастичности

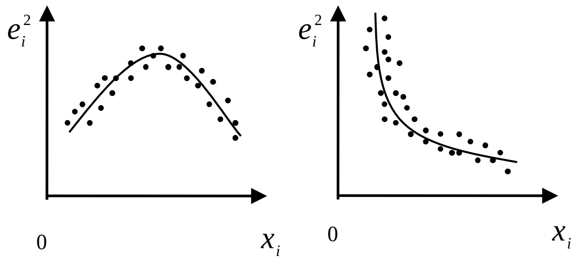
Обнаружение гетероскедастичности является довольно сложной задачей. В настоящее время существует ряд методов, позволяющих определить наличие гетероскедастичности.

1. Графический анализ остатков.

В этом случае по оси абсцисс откладываются значения объясняющей переменной X, а по оси ординат либо отклонения e_i, либо их квадраты e_i^2. Примеры таких графиков представлены на рисунке:



а б в



г д

На рисунке (а) все отклонения находятся внутри полуполосы постоянной ширины, параллельной оси абсцисс. Это говорит о независимости дисперсий e_i^2 от значений переменной X и их постоянстве, т.е. в этом случае выполняются условия гомоскедастичности.

На рисунках (б)-(д) наблюдаются некоторые систематические изменения в соотношениях между e_i^2 и X. Рисунок (б) соответствует примеру из главы 4.1. Рисунок (в) отражает линейную, рисунок (г) – квадратичную, рисунок (д) – гиперболическую зависимости между квадратами отклонений и значениями объясняющей переменной X. Другими словами, ситуации (б)-(д) отражают большую вероятность наличия гетероскедастичности для рассматриваемых статистических данных.

Графический анализ остатков является удобным в случае парной регрессии. При множественной регрессии графический анализ возможен для каждой из объясняющих переменных X_j, $j = 1,2,\dots,m$. Чаще вместо объясняющих переменных X_j по оси абсцисс откладывают значения $\hat{y}_i$, $i = 1,2,\dots,n$, получаемые из эмпирического уравнения регрессии.

2. Тест ранговой корреляции Спирмена

Значения x_i и e_i (абсолютные величины) ранжируются (упорядочиваются по величинам). Затем определяется коэффициент ранговой корреляции:

$$r_{x,e} = 1 - 6 \cdot \frac{\sum d_i^2}{n(n^2 - 1)} \qquad (4.1)$$

где d_i - разность между рангами x_i и e_i, $i = 1,2,\ldots,n$; n - число наблюдений.

Например, если x_{20} является 25-ым по величине среди всех наблюдений, а e_{20} является 32-м, то $d_{20} = 25 - 32 = -7$.

Затем рассчитывается статистика:

$$t = \frac{r_{x,e}\sqrt{(n-2)}}{\sqrt{1 - r_{x,e}^2}} \qquad (4.2)$$

Если значение, рассчитанное по формуле (4.2), превышает критическое $t_{kp} = t_{\frac{\alpha}{2}, n-2}$ (определяемое по приложению 1), то необходимо отклонить гипотезу об отсутствии гетероскедастичности. В противном случае гипотеза об отсутствии гетероскедастичности принимается.

Если в модели регрессии больше, чем одна объясняющая переменная, то проверка гипотезы может осуществляться с помощью t-статистики для каждой из них отдельно.

3. Тест Голдфелда-Квандта

В данном случае предполагается, что стандартное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению x_i переменной X в этом наблюдении, т.е. $\sigma_i^2 = \sigma^2 x_i^2$, $i = 1,2,\ldots,n$.

Тест Голдфелда-Квандта состоит в следующем:

1. Все n наблюдений упорядочиваются по величине X.
2. Вся упорядоченная выборка после этого разбивается на три подвыборки размерностей $k, (n-2k), k$ соответственно.
3. Оцениваются отдельные регрессии для первой подвыборки (k первых наблюдений) и для третьей подвыборки (k последних наблюдений). Для парной регрессии Голдфелд и Квандт предлагает следующие пропорции: $n = 30, k = 11$; $n = 60, k = 22$. Если предположение о пропорциональности дисперсий отклонений значениям X верно, то дисперсия регрессии по первой подвыборке

(рассчитываемая как $S_1 = \sum_{i=1}^{k} e_i^2$) будет существенно меньше дисперсии регрессии по третьей подвыборке (рассчитываемой как $S_3 = \sum_{i=n-k+1}^{n} e_i^2$).

4. Для сравнения соответствующих дисперсий строится соответствующая F -статистика:

$$F = \frac{S_3/(k-m-1)}{S_1/(k-m-1)} = \frac{S_3}{S_1} \tag{4.3}$$

Здесь $(k-m-1)$- число степеней свободы соответствующих выборочных дисперсий (m- количество объясняющих переменных в уравнении регрессии).

Построенная F -статистика имеет распределение Фишера с числом степеней свободы $v_1 = v_2 = n-m-1$.

5. Если $F_{набл} = \frac{S_3}{S_1} > F_{кр}$ (где $F_{кр} = F_{\alpha,v_1,v_2}$, определяется из приложения 2, α - выбранный уровень значимости), то гипотеза об отсутствии гетероскедастичности отклоняется.

Этот же тест может использоваться при предположении об обратной пропорциональности между σ_i и значениями объясняющей переменной. При этом статистика Фишера имеет вид:

$$F = \frac{S_1}{S_3} \tag{4.4}$$

Для множественной регрессии данный тест обычно проводится для той объясняющей переменной, которая в наибольшей степени связана с σ_i. При этом k должно быть больше, чем $(m+1)$. Если нет уверенности относительно выбора переменной X_j, то данный тест может осуществляться для каждой из объясняющих переменных.

4.3. Методы смягчения проблемы гетероскедастичности

Как отмечалось ранее, гетероскедастичность приводит к неэффективности оценок, что может привести к необоснованным выводам по качеству модели. Поэтому при установлении гетероскедастичности необходимо преобразовать модель с целью устранения данного недостатка. Вид преобразования зависит от того, известны или нет дисперсии σ_i^2 отклонений $\varepsilon_i, i=1,2,\dots,n$.

1. Дисперсии отклонений известны (метод взвешенных наименьших квадратов)

Данный метод применяется при известных для каждого наблюдения значениях σ_i^2. В этом случае можно устранить гетероскедастичность, разделив каждое наблюдаемое значение на соответствующее ему значение дисперсии. В этом суть метода взвешенных наименьших квадратов (ВМНК).

Для простоты рассмотрим взвешенный метод наименьших квадратов на примере парной регрессии:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.5}$$

Разделим обе части (4.5) на известное $\sigma_i = \sqrt{\sigma_i^2}$:

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i} \tag{4.6}$$

Обозначив: $\frac{y_i}{\sigma_i} = y_i^*$, $\frac{x_i}{\sigma_i} = x_i^*$, $\frac{\varepsilon_i}{\sigma_i} = v_i$, $\frac{1}{\sigma_i} = z_i$, получим уравнение регрессии без свободного члена, но с дополнительной объясняющей переменной Z и с «преобразованным» отклонением v, для которого выполняется условие гомоскедастичности:

$$y_i^* = \beta_0 z_i + \beta_1 x_i^* + v_i \tag{4.7}$$

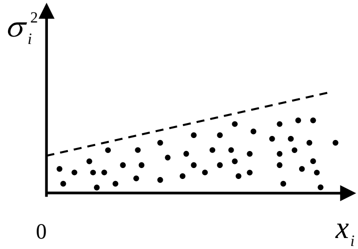
Таким образом, ВМНК включает в себя следующие этапы:

1. Значения каждой пары наблюдений (x_i, y_i) делят на известную величину σ_i. Тем самым наблюдениям с наименьшими дисперсиями придаются наибольшие «веса», а с максимальными дисперсиями – наименьшие «веса». Это увеличивает вероятность получения более точных оценок.
2. По методу наименьших квадратов для преобразованных значений $\left(\frac{1}{\sigma_i}, \frac{x_i}{\sigma_i}, \frac{y_i}{\sigma_i}\right)$ строится уравнение регрессии без свободного члена с гарантированными качествами оценок.

2. Дисперсии отклонений неизвестны

Для применения ВМНК необходимо знать фактические значения дисперсий σ_i^2. На практике такие значения известны очень редко. Следовательно, чтобы применить ВМНК, необходимо сделать реалистические предположения о значениях σ_i^2.

а). Может оказаться целесообразным предположить, что дисперсии σ_i^2 отклонений ε_i пропорциональны значениям x_i, что отражено на рисунке:



В этом случае $\sigma_i^2 = \sigma^2 x_i$ (σ^2 - коэффициент пропорциональности). Тогда уравнение (4.5) преобразуется делением его левой и правой части на $\sqrt{x_i}$:

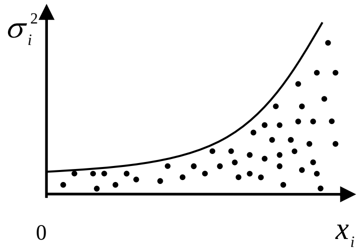
$$\frac{y_i}{\sqrt{x_i}} = \frac{\beta_0}{\sqrt{x_i}} + \beta_1 \frac{x_i}{\sqrt{x_i}} + \frac{\varepsilon_i}{\sqrt{x_i}} \Rightarrow \frac{y_i}{\sqrt{x_i}} = \beta_0 \frac{1}{\sqrt{x_i}} + \beta_1 \sqrt{x_i} + v_i \qquad (4.8)$$

При этом для случайных отклонений v_i выполняется условие гомоскедастичности. Оценив для (4.8) коэффициенты β_0 и β_1, затем возвращаются к исходному уравнению регрессии (4.5).

Если в уравнении регрессии присутствует несколько объясняющих переменных, можно поступить следующим образом. Вместо конкретной объясняющей переменной X_j используется значения, рассчитанные по эмпирическому уравнению регрессии: $\widehat{Y} = b_0 + b_1 X_1 + \ldots + b_m X_m$. В этом случае получают следующую регрессию:

$$\frac{y_i}{\sqrt{\widehat{y}_i}} = \beta_0 \frac{1}{\sqrt{\widehat{y}_i}} + \beta_1 \frac{x_{i1}}{\sqrt{\widehat{x}_i}} + \ldots + \beta_m \frac{x_{im}}{\sqrt{\widehat{x}_i}} + \frac{\varepsilon_i}{\sqrt{\widehat{x}_i}} \qquad (4.9)$$

б). Можно сделать предположение о том, что дисперсии σ_i^2 отклонений ε_i пропорциональны значениям x_i^2, что отражено на рисунке:



В этом случае необходимо преобразовать (4.5) делением на x_i к виду:

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 \frac{x_i}{x_i} + \frac{\varepsilon_i}{x_i} \Rightarrow \frac{y_i}{x_i} = \beta_0 \frac{1}{x_i} + \beta_1 + v_i . \qquad (4.10)$$

При этом для случайных отклонений $v_i = \dfrac{\varepsilon_i}{x_i}$ выполняется условие гомоскедастичности. Оценив для (4.10) коэффициенты β_0 и β_1, затем возвращаются к исходному уравнению регрессии (4.5).

Для применения описанных выше методов весьма значимы знания об истинных значениях дисперсий отклонений σ_i^2, либо предположения, какими эти дисперсии могут быть. На практике рекомендуется применять несколько методов определения гетероскедастичности и способов ее корректировки.

Пример. Исследуем зависимость между доходом (X) домохозяйства и его расходом (Y) на продукты питания. Выборочные данные по 40 домохозяйствам представлены в таблице:

X	Y
25,5	14,5
26,5	11,3
27,2	14,7
29,6	10,2
35,7	13,5
38,6	9,9
39,0	12,4

X	Y
42,5	14,9
44,2	11,6
44,8	21,5
45,5	10,8
45,5	13,8
48,3	16,0
49,5	18,2

X	Y
61,0	10,9
61,7	16,1
62,5	10,5
64,7	10,6
69,7	29,0
71,2	8,2
73,8	14,3

X	Y
79,2	19,8
81,5	21,2
82,4	29,0
82,8	17,3
83,0	23,5
85,9	22,0
86,4	18,3

39,3	8,6	52,3	19,1	74,7	21,8	86,9	13,7
40,0	10,3	55,7	16,3	75,8	26,1	88,3	14,5
41,9	13,9	59,0	17,5	76,9	20,0	89,0	27,3

Построим эмпирическое уравнение регрессии и проведем анализ модели на наличие гетероскедастичности.

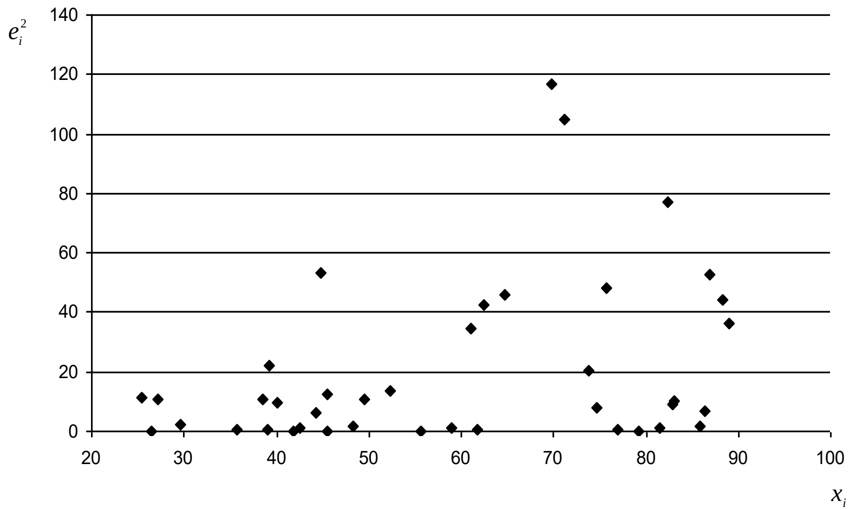
По (1.11) определим коэффициенты эмпирического уравнения регрессии: $b_0 = 7{,}04$, $b_1 = 0{,}16$. Следовательно, уравнение имеет вид: $\hat{y}_i = 7{,}04 + 0{,}16 \cdot x_i$.

Определим отклонения e_i (где $e_i = y_i - \hat{y}_i$), e_i^2, ранги X и e_i. Рассчитанные величины представим в таблице:

X	Y	$\hat{Y}$	e_i	e_i^2	Ранг X	Ранг e_i (абсол.вел.)	d_i	d_i^2
25,5	14,5	11,120	3,380	11,4244	1	25	-33	1089
26,5	11,3	11,280	0,020	0,0004	2	1	-17	289
27,2	14,7	11,392	3,308	10,9429	3	23	-30	900
29,6	10,2	11,776	-1,576	2,4838	4	15	-11	121
35,7	13,5	12,752	0,748	0,5595	5	7	-19	361
38,6	9,9	13,216	-3,316	10,9959	6	24	-4	16
39,0	12,4	13,280	-0,880	0,7744	7	9	-9	81
39,3	8,6	13,328	-4,728	22,3540	8	29	1	1
40,0	10,3	13,440	-3,140	9,8596	9	20	-2	4
41,9	13,9	13,744	0,156	0,0243	10	3	-11	121
42,5	14,9	13,840	1,060	1,1236	11	11	-15	225
44,2	11,6	14,112	-2,512	6,3101	12	16	-2	4
44,8	21,5	14,208	7,292	53,1733	13	37	-25	625
45,5	10,8	14,320	-3,520	12,3904	14	26	5	25
45,5	13,8	14,320	-0,520	0,2704	15	5	-4	16
48,3	16,0	14,768	1,232	1,5178	16	14	-13	169
49,5	18,2	14,960	3,240	10,4976	17	22	-15	225
52,3	19,1	15,408	3,692	13,6309	18	27	-17	289
55,7	16,3	15,952	0,348	0,1211	19	4	-3	9
59,0	17,5	16,480	1,020	1,0404	20	10	-5	25
61,0	10,9	16,800	-5,900	34,8100	21	30	15	225
61,7	16,1	16,912	-0,812	0,6593	22	8	5	25
62,5	10,5	17,040	-6,540	42,7716	23	32	18	324
64,7	10,6	17,392	-6,792	46,1313	24	34	21	441
69,7	29,0	18,192	10,808	116,8129	25	40	-15	225
71,2	8,2	18,432	-10,232	104,6938	26	39	25	625
73,8	14,3	18,848	-4,548	20,6843	27	28	19	361
74,7	21,8	18,992	2,808	7,8849	28	18	-2	4
75,8	26,1	19,168	6,932	48,0526	29	35	-8	64

76,9	20,0	19,344	0,656	0,4303	30	6	7	49
79,2	19,8	19,712	0,088	0,0077	31	2	11	121
81,5	21,2	20,080	1,120	1,2544	32	12	5	25
82,4	29,0	20,224	8,776	77,0182	33	38	-6	36
82,8	17,3	20,288	-2,988	8,9281	34	19	22	484
83,0	23,5	20,320	3,180	10,1124	35	21	4	16
85,9	22,0	20,784	1,216	1,4787	36	13	8	64
86,4	18,3	20,864	-2,564	6,5741	37	17	24	576
86,9	13,7	20,944	-7,244	52,4755	38	36	36	1296
88,3	14,5	21,168	-6,668	44,4622	39	33	35	1225
89,0	27,3	21,280	6,020	36,2404	40	31	4	16

Проанализируем графически остатки, представив зависимость e_i^2 от x_i:



Изучая график, можно обнаружить, что с увеличением x_i возрастает разброс значений e_i^2, что свидетельствует о наличии гетероскедастичности.

Применим для обнаружения гетероскедастичности тест ранговой корреляции Спирмена. Для этого рассчитаем по (4.1) коэффициент ранговой корреляции:

$$r_{x,e} = 1 - 6 \cdot \frac{\sum d_i^2}{n(n^2 - 1)} = 1 - 6 \cdot \frac{7595}{40(1600 - 1)} = 0.2875$$

Рассчитаем t-статистику:

$$t = \frac{r_{x,e}\sqrt{(n-2)}}{\sqrt{1-r_{x,e}^2}} = \frac{0,2875\sqrt{(40-2)}}{\sqrt{1-(0,2875)^2}} = 1,8504 .$$

Из приложения 1 определим критическое значение t-статистики для числа степеней свободы $v = n - 2 = 38$ и уровня значимости $\alpha = 0,10$: $t_{кр} = 1,303$. Так как рассчитанное значение t-статистики превышает критическое, определенное по приложению 1, то гипотеза об отсутствии гетероскедастичности отклоняется с уровнем значимости $\alpha = 0,10$.

Проверим гипотезу об отсутствии гетероскедастичности с помощью теста Голдфелда-Квандта. Для этого разобьем ряд на три подвыборки размерности 14, 12, 14.

Определим дисперсии отклонений для первой и третьей подвыборок:

$$S_1 = \sum_{i=1}^{14} e_i^2 = 142,4165 \text{ и } S_3 = \sum_{i=27}^{40} e_i^2 = 315,6039 .$$

Определим значение F -статистики. $F = \frac{S_3}{S_1} = \frac{315,6039}{142,4165} = 2,2161$.

Из приложения 2 определим критическое значение F -статистики для числа степеней свободы $v_1 = v_2 = n - m - 1 = 40 - 1 - 1 = 38$ и уровня значимости $\alpha = 0,10$: $F_{кр} = 1,51$. Так как рассчитанное значение F -статистики превышает критическое, определенное по приложению 2, то гипотеза об отсутствии гетероскедастичности отклоняется с уровнем значимости $\alpha = 0,10$.

Следовательно, по всем трем тестам гетероскедастичность в данной модели присутствует.

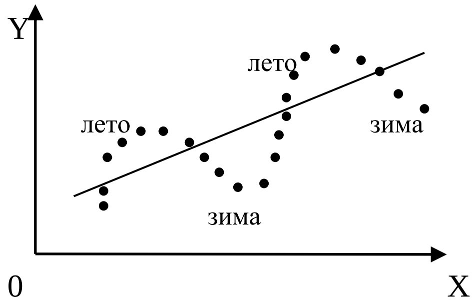
5. Автокорреляция

5.1. Суть и причины автокорреляции

Важной предпосылкой построения качественной регрессионной модели по МНК является независимость значений случайных отклонений ε_i от значений отклонений во всех других наблюдениях. Это гарантирует отсутствие коррелированности между любыми отклонениями и, в частности, между соседними отклонениями.

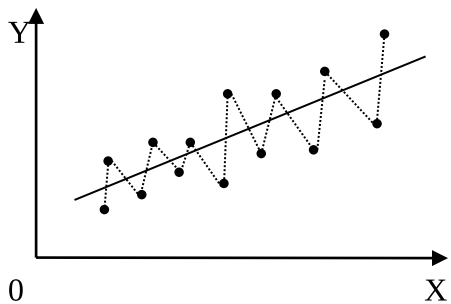
Автокорреляция (последовательная корреляция) определяется как корреляция между наблюдаемыми показателями, упорядоченными во времени (временные ряды) или в пространстве (перекрестные ряды). Автокорреляция остатков (отклонений) обычно встречается в регрессионном анализе при использовании данных временных рядов и очень редко при использовании перекрестных данных.

В экономических задачах значительно чаще встречается положительная автокорреляция, нежели отрицательная автокорреляция. В большинстве случаев *положительная автокорреляция* вызывается направленным постоянным воздействием некоторых неучтенных в модели факторов. Например, исследуем спрос Y на прохладительные напитки в зависимости от дохода X по месяцам. Фактические точки наблюдений и трендовая линейная модель представлены на рисунке:



Фактические точки наблюдений обычно будут превышать трендовую линию в летние периоды и будут ниже ее в зимние (что видно и из графика).

Отрицательная автокорреляция фактически означает, что за положительным отклонением следует отрицательное и наоборот. Такая ситуация может иметь место если ту де зависимость между спросом на прохладительные напитки Y и доходами X рассматривать по сезонным данным (зима-лето). Возможная схемы рассеивания точек при отрицательной автокорреляции может выглядеть следующим образом:



Среди основных причин, вызывающих автокорреляцию, можно выделит следующие:

1. Ошибки спецификации. Неучет в модели какой-либо важной объясняющей переменной либо неправильный выбор формы зависимости обычно приводит к системным отклонениям точек наблюдения от линии регрессии, что может обусловить автокорреляцию.
2. Инерция. Многие экономические показатели (инфляция, безработица, ВНП и т.д.) обладают определенной цикличностью, связанной с волнообразностью деловой активности. Поэтому изменение показателей происходит не мгновенно, а обладает определенной инертностью.
3. Эффект паутины. Во многих производственных и других сферах экономические показатели реагируют на изменение экономических условий с запаздыванием. (временным лагом). Например, предложение сельскохозяйственной продукции реагирует на изменение цены с запаздыванием, равным периоду созревания урожая. Большая цена сельскохозяйственной продукции в прошедшем году вызовет (скорее всего) ее перепроизводство в текущем году, а, следовательно, цена на нее снизится и т.д.
4. Сглаживание данных. Зачастую данные по некоторому продолжительному временному периоду получают усреднением данных по составляющим его интервалам. Это может привести к определенному сглаживанию колебаний, которые имелись внутри рассматриваемого периода, что в свою очередь может служить причиной автокорреляции.

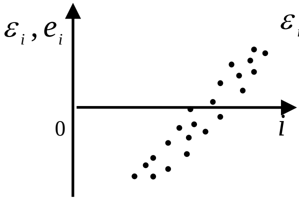
Последствия автокорреляции схожи с последствиями гетероскедастичности: выводы по t- и F- статистикам, определяющие значимость коэффициента регрессии и коэффициента детерминации, возможно, будут неверными.

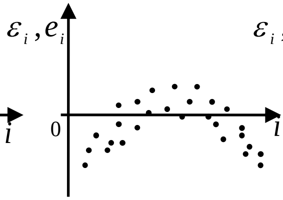
5.2. Обнаружение автокорреляции

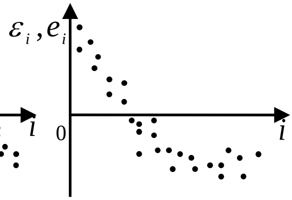
Существует несколько методов, позволяющих обнаружить автокорреляцию.

1. Графический метод

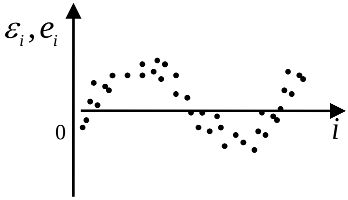
Существует несколько вариантов графического определения автокорреляции. Один из них увязывает отклонения e_i с моментами их получения $i = 1,2\ldots,n$. При этом по оси абсцисс откладывают либо время получения статистических данных, либо порядковый номер наблюдения, а по оси ординат – отклонения ε_i, либо оценки отклонений e_i.

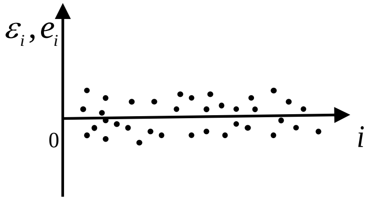






а б в

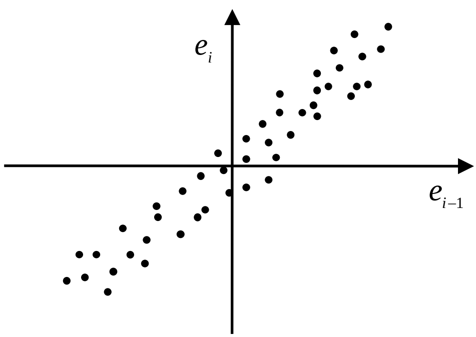




г д

Естественно предположить, что на рисунках (а)-(г) имеются определенные связи между отклонениями, т.е. автокорреляция имеет место. Отсутствие зависимости на рисунке (д) скоре всего свидетельствует об отсутствии автокорреляции.

Для случая (б) отклонения сначала являются отрицательными, затем положительными, затем снова отрицательными. Это свидетельствует о наличии между отклонениями определенной зависимости, более того, можно утверждать, что в этом случае имеет место положительная автокорреляция. Она становится более наглядной, если построить график зависимости e_i от e_{i-1}:



Подавляющее большинство точек на этом графике расположено в первой и третьей четвертях декартовой системы координат, подтверждая положительную зависимость между соседними отклонениями.

2. Критерий Дарбина-Уотсона

Этот критерий является наиболее известным для обнаружения автокорреляции. Метод определения автокорреляции на основе критерия Дарбина-Уотсона и пример был подробно рассмотрен в главе 2.3.

5.3. Методы устранения автокорреляции

Основной причиной наличия случайного члена в модели являются несовершенные знания о причинах и взаимосвязях, определяющих то или иное значение зависимой переменной. Поэтому свойства случайных отклонений, в том числе и автокорреляция, в первую очередь зависят от выбора формулы зависимости и состава объясняющих переменных. Так как автокорреляция чаще всего вызывается неправильной спецификацией модели, то необходимо прежде всего скорректировать саму модель. Возможно, автокорреляция вызвана отсутствием в модели некоторой объясняющей переменной. Следует попытаться определить данный фактор и учесть его в уравнении регрессии. Также можно попробовать изменить форму зависимости (например, линейную на лог-линейную, линейную на гиперболическую и т.д.).

Однако если все разумные процедуры изменения спецификации модели, на ваш взгляд, исчерпаны, а автокорреляция имеет место, то можно предположить, что она обусловлена какими-то внутренними свойствами ряда $\{e_i\}$. В этом случае можно воспользоваться авторегрессионным преобразованием. В линейной регрессионной модели либо в моделях, сводящихся к линейной, наиболее целесообразным и простым преобразованием является *авторегрессионная схема первого порядка* AR(1).

Для простоты изложения AR(1) рассмотрим модель парной линейной регрессии:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (5.1)$$

Тогда наблюдениям i и $(i-1)$ соответствуют формулы:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (5.2)$$

$$y_{i-1} = \beta_0 + \beta_1 x_{i-1} + \varepsilon_{i-1} \qquad (5.3)$$

Пусть случайные отклонения подвержены воздействию авторегрессии первого порядка:

$$\varepsilon_i = \rho\varepsilon_{i-1} + v_i \qquad (5.4)$$

где v_i, $i = 2,3\ldots,n$ - случайные отклонения, удовлетворяющие всем предпосылкам МНК, а коэффициент ρ известен.

Вычтем из (5.2) соотношение (5.3), умноженное на ρ:

$$y_i - \rho y_{i-1} = \beta_0(1-\rho) + \beta_1(x_i - \rho x_{i-1}) + (\varepsilon_i - \rho\varepsilon_{i-1}) \qquad (5.5)$$

Примем $y_i^* = y_i - \rho y_{i-1}$, $x_i^* = x_i - \rho x_{i-1}$, $\beta_0^* = \beta_0(1-\rho)$, и с учетом (5.4) получим:

$$y_i^* = \beta_0^* + \beta_1 x_i^* + v_i \qquad (5.6)$$

Так как по предположению коэффициент ρ известен , то очевидно, y_i^* , x_i^*, v_i вычисляются достаточно просто.

Однако способ вычисления y_i^* и x_i^* приводит к потере первого наблюдения. Число степеней свободы уменьшится на единицу, что при больших выборках не так существенно, но при малых выборках может привести к потере эффективности. Эта проблема обычно преодолевается с помощью *поправки Прайса-Винстена*:

$$\begin{aligned} x_1^* &= \sqrt{1-\rho^2} \cdot x_1 \\ y_1^* &= \sqrt{1-\rho^2} \cdot y_1 \end{aligned} \qquad (5.7)$$

Рассмотренное авторегрессионное преобразование может быть обобщено на произвольное число объясняющих переменных, т.е. использовано для уравнения множественной регрессии.

На практике значения коэффициента ρ обычно неизвестно и его необходимо оценивать. Существует несколько методов оценивания. Рассмотрим наиболее употребляемые.

1. Определение ρ на основе статистики Дарбина-Уотсона

Статистика Дарбина-Уотсона тесно связана с коэффициентом корреляции между соседними отклонениями через соотношение:

$$DW \approx 2(1 - r_{e_i, e_{i-1}}) \qquad (5.8)$$

Тогда в качестве оценки коэффициента ρ может быть взят коэффициент $r = r_{e_i, e_{i-1}}$. Из (5.8) имеем:

$$r \approx 1 - \frac{DW}{2} \qquad (5.9)$$

Этот метод оценивания рекомендуется применять при большом числе наблюдений. В этом случае оценка r параметра ρ будет достаточно точной.

2. Метод Хилдрета-Лу

По данному методу регрессия (5.5) оценивается для каждого возможного значения ρ из отрезка [-1;1] с любым шагом (например, 0,001; 0,01 и т.д.). Величина $\hat{\rho}$, дающая наименьшую стандартную ошибку регрессии, принимается в качестве оценки коэффициента ρ. И значения β_0^* и β_1 оцениваются из уравнения регрессии (5.5) именно с данным значением $\hat{\rho}$.

3. Метод первых разностей

В случае, когда автокорреляция отклонений очень велика, используется метод первых разностей.

При высокой положительной автокорреляции полагают, что $\rho = 1$, следовательно, уравнение (5.5) примет вид:

$$y_i - y_{i-1} = \beta_1(x_i - x_{i-1}) + (\varepsilon_i - \varepsilon_{i-1})$$

или

$$y_i - y_{i-1} = \beta_1(x_i - x_{i-1}) + v_i \tag{5.10}$$

Обозначив $\Delta y_i = y_i - y_{i-1}; \Delta x_i = x_i - x_{i-1}$, из (5.10) получаем:

$$\Delta y_i = \beta_1 \Delta x_i + v_i \tag{5.11}$$

Из уравнения (5.11) по МНК оценивается коэффициент β_1. Коэффициент β_0 в данном случае не определяется непосредственно. Но из МНК известно, что $\beta_0 = \bar{y} - \beta_1 \bar{x}$.

В случае $\rho = -1$, сложив (5.2) и (5.3) с учетом (5.4), получаем следующее уравнение регрессии:

$$y_i - y_{i-1} = 2\beta_0 + \beta_1(x_i - x_{i-1}) + v_i$$

или

$$\frac{y_i - y_{i-1}}{2} = \beta_0 + \beta_1 \frac{x_i - x_{i-1}}{2} + v_i \tag{5.12}$$

Недостатком этого метода является то, что он предполагает слишком большое упрощение ($\rho = \pm 1$), поэтому более предпочтительными являются приведенные выше методы.

6. Фиктивные переменные в регрессионных моделях

6.1. Необходимость использования фиктивных переменных

В регрессионных моделях в качестве объясняющих переменных часто приходится использовать не только количественные (определяемые численно), но и качественные переменные. Например, спрос на какое-либо благо может определяться как количественными переменными (цена данного блага, цена на заменители данного блага, доход потребителя и т.д.), так и качественными (вкусы потребителей, их ожидания, национальные и религиозные особенности и т.д.). качественные показатели в численном виде представить нельзя. Возникает проблема отражения в модели влияния таких переменных на исследуемую величину.

Обычно в моделях влияние качественного фактора выражается в виде фиктивной (искусственной) переменной, которая отражает два противоположных состояния качественного фактора. В этом случае фиктивная переменная может выражаться в двоичной форме:

$$D = \begin{cases} 0, & \text{фактор не действует} \\ 1, & \text{фактор действует} \end{cases}$$

Например, D=0, если потребитель не имеет высшего образования, D=1, если потребитель имеет высшее образование; D=0, если в обществе имеются инфляционные ожидания, D=1, если инфляционных ожиданий нет.

Переменная D называется *фиктивной (искусственной, двоичной) переменной (индикатором)*.

Регрессионные модели, содержащие лишь качественные объясняющие переменные, называются *моделями дисперсионного анализа* (ANOVA-моделями).

Например, пусть Y - начальная заработная плата.

$$D = \begin{cases} 0, & \text{если претендент не имееет высшего образования} \\ 1, & \text{если претендент имеет высшее образование} \end{cases}$$

Тогда зависимость можно выразить моделью парной регрессии:

$$Y = \beta_0 + \gamma D + \varepsilon \qquad (6.1)$$

Очевидно, $$M(Y|D=0) = \beta_0 + \gamma \cdot 0 = \beta_0$$
$$M(Y|D=1) = \beta_0 + \gamma \cdot 1 = \beta_0 + \gamma$$

При этом коэффициент β_0 определяет среднюю начальную заработную плату при отсутствии высшего образования. Коэффициент γ указывает, на какую величину отличаются средние начальные заработные платы при наличии и при отсутствии высшего образования у претендента. Проверяя статистическую значимость коэффициента γ с помощью t-статистики либо значимость коэффициента детерминации R^2 с помощью F-статистики, можно определить, влияет или нет наличие высшего образования на начальную заработную плату.

Модели, в которых объясняющие переменные носят как количественный, так и качественный характер, называются *моделями ковариационного анализа* (ANCOVA-моделями).

6.2. Модели ковариационного анализа

Существует несколько разновидностей таких моделей.

1. Модели ковариационного анализа при наличии у фиктивной переменной двух альтернатив

Сначала рассмотри простейшую модель с одной количественной и одной качественной переменной, имеющей два альтернативных состояния:

$$Y = \beta_0 + \beta_1 X + \gamma D + \varepsilon \tag{6.2}$$

Пусть, например, Y заработная плата сотрудника фирмы, X - стаж сотрудника, D - пол сотрудника, т.е.

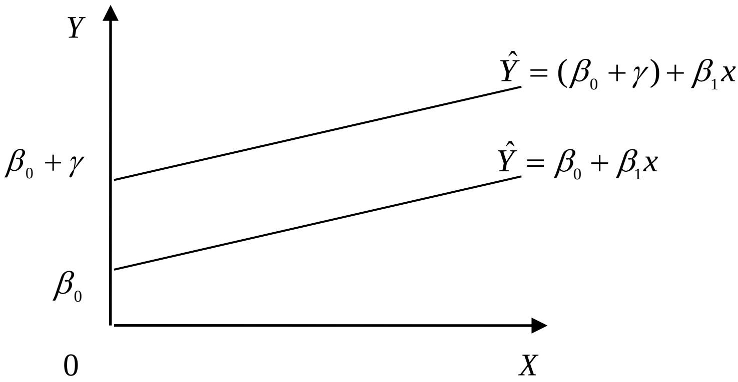
$$D = \begin{cases} 0, & \text{если сотрудник - женщина} \\ 1, & \text{если сотрудник - мужчина} \end{cases}$$

Тогда ожидаемое значение заработной платы сотрудников при x годах трудового стажа будет:

$$M(Y|x, D = 0) = \beta_0 + \beta_1 x$$

$$M(Y|x, D = 1) = \beta_0 + \beta_1 x + \gamma = (\beta_0 + \gamma) + \beta_1 x$$

Заработная плата в данном случае является линейной функцией от стажа работы. Графически данную ситуацию можно представить следующим образом:



Очевидно, что и для мужчин, и для женщин заработная плата меняется с одним и тем же коэффициентом пропорциональности β_1. Однако свободные члены отличаются на величину γ. Проверив с помощью t-статистики статистические значимости коэффициентов β_0 и $(\beta_0 + \gamma)$, можно определить, имеет ли место в фирме дискриминация по половому признаку. Если эти коэффициенты окажутся статистически значимыми, то, очевидно, дискриминация есть. Более того, при $\gamma > 0$ она будет в пользу мужчин, при $\gamma < 0$ - в пользу женщин.

При составлении моделей с фиктивными переменными необходимо руководствоваться следующим *правилом моделирования*: если качественная переменная имеет k альтернативных значений, то при моделировании используются только $(k-1)$ фиктивных переменных. Таким образом, если переменная имеет два альтернативных значения (например, пол), то в модель можно ввести только одну фиктивную переменную.

Если не следовать данному правилу, то при моделировании исследователь попадает в ситуацию совершенной мультиколлинеарности или так называемую *ловушку фиктивной переменной*.

Значения фиктивной переменной можно изменять на противоположные. Суть модели от этого не изменится. Например, в модели (6.2) можно положить, что:

$$D = \begin{cases} 0, & \text{если сотрудник - мужчина} \\ 1, & \text{если сотрудник - женщина} \end{cases}.$$

Однако при этом знак коэффициента γ изменится на противоположный.

Значения качественной переменной, для которого принимается $D = 0$, называется *базовым* или *сравнительным*. Выбор базового значения обычно диктуется целями исследования, но может быть и произвольным.

Коэффициент γ в модели (6.2) иногда называется *дифференциальным коэффициентом свободного члена*, так как он показывает, на какую величину отличается свободный член модели при значении фиктивной переменной, равном единице, от свободного члена модели при базовом значении фиктивной переменной.

2. Модели ковариационного анализа при наличии у качественных переменных более двух альтернатив

Рассмотрим модель с двумя объясняющими переменными, одна из которых – количественная, а другая - качественная. Причем качественная имеет три альтернативы. Например, в расходы на содержание ребенка могут быть связаны с доходами домохозяйства и возрастом ребенка: дошкольный, младший школьный и старший школьный. Так как качественная переменная имеет три альтернативы, то по общему правилу моделирования необходимо использовать две фиктивные переменные. Таким образом, модель может быть представлена в виде:

$$Y = \beta_0 + \beta_1 X + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon \qquad (6.3)$$

где Y - расходы, X - доходы домохозяйств.

$$D_1 = \begin{cases} 0, & \text{если дошкольник} \\ 1, & \text{в противопол ожном случае} \end{cases}$$

$$D_2 = \begin{cases} 0, & \text{если дошкольник или младший школьник} \\ 1, & \text{в противопол ожном случае} \end{cases}$$

Образуются следующие зависимости:

Средний расход на дошкольника:

$$M(Y|D_1 = 0, D_2 = 0) = \beta_0 + \beta_1 X \qquad (6.4)$$

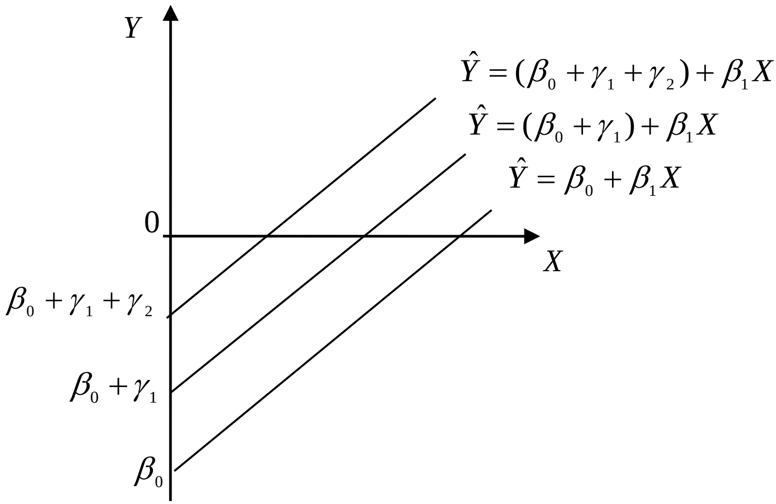
Средний расход на младшего школьника:

$$M(Y|D_1 = 1, D_2 = 0) = (\beta_0 + \gamma_1) + \beta_1 X \qquad (6.5)$$

Средний расход на старшего школьника:

$$M(Y|D_1 = 1, D_2 = 1) = (\beta_0 + \gamma_1 + \gamma_2) + \beta_1 X \qquad (6.6)$$

Здесь γ_1, γ_2 - дифференциальные свободные члены. Базовым значением качественной переменной является значение «дошкольник». Таким образом, получаются три регрессионные прямые, параллельные друг другу:



После вычисления коэффициентов уравнений регрессии (6.4)-(6.6) определяется статистическая значимость коэффициентов γ_1 и γ_2 на основе обычной t-статистики.

Если коэффициенты γ_1 и γ_2 оказываются статистически незначимыми, то можно сделать вывод, что возраст ребенка не оказывает существенного влияния на расходы по его содержанию.

3. Регрессия с одной количественной и двумя качественными переменными

Техника фиктивных переменных может быть распространена на произвольное число качественных факторов. Для простоты рассмотрим ситуацию с двумя качественными переменными.

Пусть Y - заработная плата сотрудников фирмы, X - стаж работы, D_1 - наличие высшего образования, D_2 - пол сотрудника:

$$D_1 = \begin{cases} 0, & \text{если сотрудник - женщина} \\ 1, & \text{если сотрудник - мужчина} \end{cases}$$

$$D_2 = \begin{cases} 0, & \text{если нет высшего образования} \\ 1, & \text{в противопол ожном случае} \end{cases}$$

Таким образом, получим следующую модель:

$$Y = \beta_0 + \beta_1 X + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon \tag{6.7}$$

Из этой модели выводятся следующие регрессионные зависимости:

Средняя зарплата женщины без высшего образования:

$$M(Y|D_1 = 0, D_2 = 0) = \beta_0 + \beta_1 X \quad (6.8)$$

Средняя зарплата женщины с высшим образованием:

$$M(Y|D_1 = 0, D_2 = 1) = (\beta_0 + \gamma_2) + \beta_1 X \quad (6.9)$$

Средняя зарплата мужчины без высшего образования:

$$M(Y|D_1 = 1, D_2 = 0) = (\beta_0 + \gamma_1) + \beta_1 X \quad (6.10)$$

Средняя зарплата мужчины с высшим образованием:

$$M(Y|D_1 = 1, D_2 = 1) = (\beta_0 + \gamma_1 + \gamma_2) + \beta_1 X \quad (6.11)$$

Очевидно, что все регрессии отличаются только свободными членами. Дальнейшее определение статистической значимости коэффициентов γ_1 и γ_2 позволяет убедиться, влияют ли образование и пол сотрудника на его заработную плату.

Пример. Рассмотрим зависимость между весом новорожденного Y (в граммах), X - количеством сигарет, выкуриваемых в день будущей матерью во время беременности и фиктивной переменной D, которая отражает факт того, является ребенок первенцем или нет. Пусть $D = 0$, если ребенок – первенец и $D = 1$, если ребенок не первенец. Рассмотри выборку из 20 значений:

наблюдение	Y	X	D
1	3520	10	1
2	3460	19	1
3	3000	16	1
4	3320	26	1
5	3540	4	1
6	3310	14	1
7	3360	21	1
8	3650	10	1
9	3150	22	1
10	3440	8	1

наблюдение	Y	X	D
11	3210	29	1
12	3290	15	1
13	3190	3	0
14	3060	12	0
15	3270	17	0
16	3170	14	0
17	3230	18	0
18	3700	11	0
19	3300	14	0
20	3460	9	0

Данная модель содержит одну количественную и одну качественную переменные. В общем виде запишем ее следующим образом: $Y = b_0 + b_1 X + b_2 D$. Коэффициенты b_0, b_1, b_2 определяются из формулы (2.17). Вспомогательная таблица для расчета коэффициентов имеет вид:

№	$(y_i-\bar{y})$	$(x_i-\bar{x})$	$(d_i-\bar{d})$	$(x_i-\bar{x})^2$	$(d_i-\bar{d})^2$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$	$(d_i-\bar{d})\cdot(y_i-\bar{y})$	$(x_i-\bar{x})\cdot(d_i-\bar{d})$
1	188,5	-4,6	188,5	21,16	0,16	-867,1	75,4	-1,84
2	128,5	4,4	128,5	19,36	0,16	565,4	51,4	1,76
3	-331,5	1,4	-331,5	1,96	0,16	-464,1	-132,6	0,56
4	-11,5	11,4	-11,5	129,96	0,16	-131,1	-4,6	4,56
5	208,5	-10,6	208,5	112,36	0,16	-2210,1	83,4	-4,24
6	-21,5	-0,6	-21,5	0,36	0,16	12,9	-8,6	-0,24
7	28,5	6,4	28,5	40,96	0,16	182,4	11,4	2,56
8	318,5	-4,6	318,5	21,16	0,16	-1465,1	127,4	-1,84
9	-181,5	7,4	-181,5	54,76	0,16	-1343,1	-72,6	2,96
10	108,5	-6,6	108,5	43,56	0,16	-716,1	43,4	-2,64
11	-121,5	14,4	-121,5	207,36	0,16	-1749,6	-48,6	5,76
12	-41,5	0,4	-41,5	0,16	0,16	-16,6	-16,6	0,16
13	-141,5	-11,6	-141,5	134,56	0,36	1641,4	84,9	6,96
14	-271,5	-2,6	-271,5	6,76	0,36	705,9	162,9	1,56
15	-61,5	2,4	-61,5	5,76	0,36	-147,6	36,9	-1,44
16	-161,5	-0,6	-161,5	0,36	0,36	96,9	96,9	0,36
17	-101,5	3,4	-101,5	11,56	0,36	-345,1	60,9	-2,04
18	368,5	-3,6	368,5	12,96	0,36	-1326,6	-221,1	2,16
19	-31,5	-0,6	-31,5	0,36	0,36	18,9	18,9	0,36
20	128,5	-5,6	128,5	31,36	0,36	-719,6	-77,1	3,36
$\sum$				856,8	4,8	-8278	272	18,8

При этом: $\bar{y} = 3331{,}5$, $\bar{x} = 14{,}6$, $\bar{d} = 0{,}6$.

$$b_0 = 3331{,}5 + 11{,}93 \cdot 14{,}6 - 103{,}39 \cdot 0{,}6 = 3443{,}64$$

$$b_1 = \frac{(-8278)\cdot 4{,}8 - 272\cdot 18{,}8}{856{,}8\cdot 4{,}8 - (18{,}8)^2} = -11{,}93$$

$$b_2 = \frac{272\cdot 856{,}8 - (-8278)\cdot 18{,}8}{856{,}8\cdot 4{,}8 - (18{,}8)^2} = 103{,}39$$

Таким образом, уравнение регрессии с учетом рассчитанных коэффициентов примет вид: $\hat{Y} = 3443{,}64 - 11{,}93X + 103{,}39D$

Затем аналогично примеру, приведенному в главе 2, рассчитывается статистическая значимость коэффициентов. Рассчитанное значение t - статистики для коэффициента b_2 при фиктивной переменной D составляет $t = 1{,}23$.

Из приложения 1 определим для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $v = n - m - 1 = 20 - 2 - 1 = 17$ критическое значение t - статистики: $t_{кр} = t_{\frac{\alpha}{2}, n-m-1} = t_{0{,}025{,}17} = 2{,}110$. Так как $t < t_{кр}$, то коэффициент b_2 при фиктивной переменной D является статистически незначимым с уровнем значимости 0,05.

Однако можно предположить, что это объясняется малым размером выборки (20 значений). Если рассмотреть большую выборку, то обнаружится статистическая значимость данного коэффициента.

6.3. Сравнение двух регрессий

В примерах, рассматриваемых до сих пор, предполагалось, что изменение значения качественного фактора влияет лишь на изменение свободного члена. Но это не всегда так. Изменение качественного фактора может привести также к изменению наклона прямой регрессии.

Обычно это характерно для временных рядов экономических данных при изменении институциональных условий, введении новых правовых или налоговых ограничений. Например, можно предположить, что до некоторого года в стране обменный курс валют был фиксированным, а затем плавающим. Или налог на ввозимые автомобили был одним, а затем он существенно изменился. В этом случае зависимость может быть выражена так:

$$Y_t = \beta_0 + \beta_1 X_t + \gamma_1 D_t + \gamma_2 D_t X_t + \varepsilon_t \qquad (6.12)$$

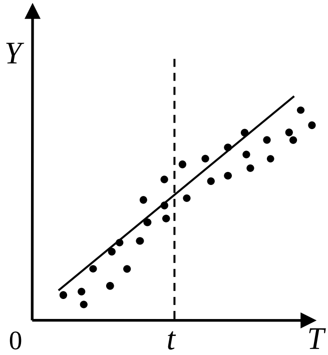
где $D_t = \begin{cases} 0, & \text{до изменения институциональных условий} \\ 1, & \text{после изменения институциональных условий} \end{cases}$

В этой ситуации ожидаемое значение зависимой переменной определяется следующим образом:

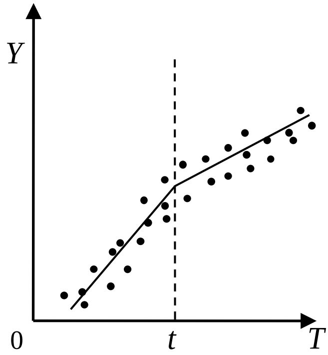
$$M(Y_t | D_t = 0) = \beta_0 + \beta_1 X_t \qquad (6.13)$$

$$M(Y_t | D_t = 1) = (\beta_0 + \gamma_1) + (\beta_1 + \gamma_2) X_t \qquad (6.14)$$

Коэффициенты γ_1 и γ_2 в уравнении (6.12) называются *дифференциальным свободным членом* и *дифференциальным угловым* *коэффициентом* соответственно. Фиктивная переменная D_t в уравнении (6.12) используется как в аддитивном виде $(\gamma_1 D_t)$, так и в мультипликативном $(\gamma_2 D_t X_t)$, что позволяет фактически разбивать рассматриваемую зависимость на две части, связанные с периодами изменения некоторого рассматриваемого в модели качественного фактора. Уравнение регрессии (6.12) достаточно хорошо моделирует ситуацию, изображенную на рисунке:



а



б

На рисунке (а) зависимость моделируется обыкновенной линейной регрессией. На рисунке (б) в модели учитываются изменения, произошедшие с некоторого времени t в характере расположения точек наблюдений. На данном примере хорошо видно, каким образом можно проанализировать, имеет ли смысл разбивать выборку на части и строить для каждой из них уравнение регрессии (т.е. фактически строить сложную регрессию с фиктивными переменными), либо можно ограничиться общей «обыкновенной» регрессией для всех точек наблюдений. Для этого можно использовать тест Чоу.

Суть теста Чоу заключается в следующем. Пусть выборка имеет объем n. Через S_0 обозначим сумму квадратов отклонений $\sum e_i^2$ значений y_i от общего уравнения регрессии. Пусть есть основание предполагать, что целесообразно общую выборку разбить на две подвыборки объемами n_1 и n_2 соответственно $(n_1 + n_2 = n)$ и построить для каждой из выборок уравнение регрессии. Обозначим через S_1 и S_2 суммы квадратов отклонений значений y_i каждой из подвыборок от соответствующих уравнений регрессий. Затем рассчитывается F - статистика, которая для теста Чоу имеет вид:

$$F = \frac{S_0 - S_1 - S_2}{S_1 + S_2} \cdot \frac{n - 2m - 2}{m + 1} \tag{6.15}$$

где m - число количественных объясняющих переменных в уравнении регрессии (одинаково для всех трех уравнений регрессии).

Из приложения 2 определяется $F_{кр}$ для числа степеней свободы $v_1 = m + 1$, $v_2 = n - 2m - 2$ и требуемого уровня значимости α. Если $F_{набл} < F_{кр}$ при выбранном уровне значимости, то нет смысла разбивать уравнение регрессии на части. В противном случае разбиение на подынтервалы целесообразно с точки зрения улучшения качества модели, что означает необходимость введения в уравнение регрессии соответствующе фиктивной переменной.

6.4. Использование фиктивных переменных в сезонном анализе

Многие экономические показатели напрямую связаны с сезонными колебаниями. Например, спрос на туристические путевки, охлажденную воду и мороженое существенно выше летом, чем зимой. Спрос на обогреватели, шубы выше зимой, чем летом. Некоторые показатели имеют существенные квартальные колебания и т.д.

Обычно сезонные колебания характерны для временных рядов. Устранение или нейтрализация сезонного фактора в таких моделях позволяет сконцентрироваться на других важных количественных и качественных характеристиках модели, в частности, на общем направлении развития модели, так называемом *тренде*. Такое устранение сезонного фактора называется *сезонной корректировкой*. Существует несколько методов сезонной корректировки, одним из которых является *метод фиктивных переменных*.

Пусть переменная Y определяется количественной переменной X, причем эта зависимость существенно разнится по кварталам. Тогда общую модель в этой ситуации можно представить в виде:

$$Y_t = \beta_0 + \beta_1 X_t + \gamma_1 D_{1t} + \gamma_2 D_{2t} + \gamma_3 D_{3t} + \varepsilon_t \tag{6.16}$$

где $D_{1t} = \begin{cases} 1, & \text{если рассматривается II квартал} \\ 0, & \text{в противном случае} \end{cases}$

$$D_{2t} = \begin{cases} 1, & \text{если рассматривается III квартал} \\ 0, & \text{в противном случае} \end{cases}$$

$$D_{3t} = \begin{cases} 1, & \text{если рассматривается IV квартал} \\ 0, & \text{в противном случае} \end{cases}$$

Заметим, что число кварталов равно четырем, следовательно, по правилу моделирования, число фиктивных переменных должно быть равно трем.

В нашем примере в качестве базы выбран I квартал. Если значения Y существенно различаются по кварталам (сезонам), то в уравнении (6.16) коэффициенты при фиктивных переменных окажутся статистически значимыми. Тогда ожидаемое значение Y по кварталам определяется следующими соотношениями:

$M(Y_t | D_{1t} = 0, D_{2t} = 0, D_{3t} = 0) = \beta_0 + \beta_1 X_t$ - для I квартала

$M(Y_t | D_{1t} = 1, D_{2t} = 0, D_{3t} = 0) = (\beta_0 + \gamma_1) + \beta_1 X_t$ - для II квартала

$M(Y_t | D_{1t} = 0, D_{2t} = 1, D_{3t} = 0) = (\beta_0 + \gamma_2) + \beta_1 X_t$ - для III квартала

$M(Y_t | D_{1t} = 0, D_{2t} = 0, D_{3t} = 1) = (\beta_0 + \gamma_3) + \beta_1 X_t$ - для IV квартала

В модели (6.16) рассматриваются такие ситуации, при которых квартальные различия отражаются лишь в различии свободных членов моделей. Если же различия затрагивают и изменения коэффициента пропорциональности, то этот факт может быть отражен в следующей модели:

$$Y_t = \beta_0 + \beta_1 X_t + \gamma_1 D_{1t} + \gamma_2 D_{2t} + \gamma_3 D_{3t} + \gamma_4 D_{1t} X_t + \gamma_5 D_{2t} X_t + \gamma_6 D_{3t} X_t + \varepsilon_t \tag{6.17}$$

Выбор правильной формы модели регрессии является в данном случае достаточно серьезной проблемой, так как вполне вероятны ошибки спецификации. Наиболее рациональной практической стратегией выбора модели является следующая схема. Вначале рассматривается модель (6.17). определяется статистическая значимость коэффициентов. Если дифференциальные угловые коэффициенты оказываются статистически незначимыми, то переходят к модели (6.16). Если в этой модели дифференциальные свободные члены оказываются статистически незначимыми, то делают вывод, что квартальные (сезонные) изменения несущественны для рассматриваемой зависимости.

Пример. Данные по расходам потребителей на газ и электричество в США в постоянных ценах 1972 года с I квартала 1977 года по IV квартал 1982 года (млрд. долл.) представлены в таблице.

	I квартал	II квартал	III квартал	IV квартал
1977 год	7,33	4,70	5,10	5,46
1978 год	7,65	4,92	5,15	5,55
1979 год	7,96	5,01	5,05	5,59
1980 год	7,74	5,10	5,67	5,92
1981 год	8,04	5,27	5,51	6,04
1982 год	8,26	5,51	5,41	5,83

Для модели вида (6.16) с соответствующими переменными D_{1t}, D_{2t}, D_{3t} по данным, приведенным в таблице, в результате расчетов по формуле (2.14) получена модель вида:

$$\hat{Y}_t = 7{,}50 + 0{,}030t - 2{,}78D_1 - 2{,}58D_2 - 2{,}19D_3$$

Из этого результата выводим отдельные уравнения для каждого квартала:

для Iквартала: $\hat{Y}_t = 7{,}50 + 0{,}030t$

для II квартала: $\hat{Y}_t = 4{,}72 + 0{,}030t$

для III квартала: $\hat{Y}_t = 4{,}92 + 0{,}030t$

для IV квартала: $\hat{Y}_t = 5{,}31 + 0{,}030t$

7. Динамические модели

7.1. Временные ряды. Лаги в экономических моделях

При анализе многих экономических показателей (особенно в макроэкономике) часто используются ежегодные, ежеквартальные, ежемесячные, ежедневные данные. Для рационального анализа необходимо систематизировать моменты получения соответствующих статистических данных.

В этом случае следует упорядочить данные по времени их получения и построить так называемые *временные ряды*.

Пусть исследуется показатель Y. Его значение в текущий момент (период) времени t обозначают y_t; значения Y в последующие моменты обозначаются $y_{t+1}, y_{t+2}, \ldots, y_{t+k}, \ldots$; значения Y в предыдущие моменты времени обозначаются $y_{t-1}, y_{t-2}, \ldots, y_{t-k}, \ldots$.

Нетрудно понять, что при изучении зависимостей между такими показателями либо при анализе их развития во времени в качестве объясняющих переменных используются не только текущие значения переменных, но и некоторые предыдущие по времени значения, а также само время T . Модели такого типа называются *динамическими*.

В свою очередь переменные, влияние которых характеризуется определенным запаздыванием, называются *лаговыми переменными*.

Обычно динамические модели подразделяются на два класса:

1. *модели с лагами* (модели с распределенными лагами) – содержат в качестве лаговых переменных лишь независимые (объясняющие) переменные. Примером является модель:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_k x_{t-k} + \varepsilon_t \tag{7.1}$$

2. *авторегрессионные модели* – модели, уравнения которых в качестве лаговых объясняющих переменных включают значения зависимых переменных.

$$y_t = \alpha + \beta_0 x_t + \gamma y_{t-1} + \varepsilon_t \tag{7.2}$$

Во многих случаях воздействие одних экономических факторов на другие осуществляется не мгновенно, а с некоторым временным запаздыванием – лагом. Причин наличия лагов в экономике достаточно много, среди них можно выделить следующие:

- *психологические причины* – обычно выражаются через инерцию в поведении людей. Например, люди тратят свой доход не мгновенно, а постепенно. Привычка к определенному образу жизни приводит к тому, что люди приобретают те же блага в течение некоторого времени даже после падения реального дохода.
- *технологические причины*. Например, изобретение персональных компьютеров не привело к мгновенному вытеснению ими больших ЭВМ в силу необходимости замены соответствующего программного обеспечения, которое потребовало продолжительного времени.
- *институциональные причины*. Например, контракты между фирмами требуют определенного постоянства в течение времени контракта.
- *механизмы формирования экономических показателей*. Например, инфляция во многом является инерционным процессом.

7.2. Оценка моделей с лагами в независимых переменных

Оценка модели с распределенными лагами во многом зависит от того, конечное или бесконечное число лагов она содержит.

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_k x_{t-k} + \varepsilon_t \text{ (конечное число лагов)}$$

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \varepsilon_t \text{ (бесконечное число лагов)} \quad (7.3)$$

В обеих этих моделях коэффициент β_0 называется *краткосрочным мультипликатором*. Он характеризует изменение среднего значения Y под воздействием единичного изменения переменной X в тот же самый момент времени.

Сумма всех коэффициентов $\sum_j \beta_j$ называется *долгосрочным мультипликатором*. Он характеризует изменение Y под воздействием единичного изменения переменной X в каждом из рассматриваемых временных периодов.

Любую сумму коэффициентов $\sum_j^h \beta_j (h < k)$ называют *промежуточным мультипликатором*.

Модель с конечным числом лагов (7.1) оценивается достаточно просто сведением ее к уравнению множественной регрессии. При этом $X_0^* = x_t$, $X_1^* = x_{t-1}$, …, $X_k^* = x_{t-k}$ и получают уравнение:

$$y_t = \alpha + \beta_0 X_0^* + \beta_1 X_1^* + \ldots + \beta_k X_k^* + \varepsilon_t \quad (7.4)$$

Для оценки моделей с бесконечным числом лагов разработано несколько моделей. Рассмотрим две из них:

1. Метод последовательного увеличения количества лагов

По данному методу уравнение (7.3) рекомендуется оценивать с последовательно увеличивающимся количеством лагов. Признаком завершения процедуры увеличения количества лагов может являться следующее:

- при добавлении нового лага какой-либо коэффициент регрессии β_k при переменной x_{t-k} меняет знак. Тогда в уравнении регрессии оставляют переменные $x_t, x_{t-1}, \ldots, x_{t-k+1}$, коэффициенты при которых знак не поменяли.

- при добавлении нового лага коэффициент регрессии β_k при переменной x_{t-k} становится статистически незначимым. Очевидно, что в уравнении будут использоваться только переменные $x_t, x_{t-1}, \ldots, x_{t-k+1}$, коэффициенты при которых остаются статистически значимыми.

Применение метода последовательного увеличения количества лагов весьма ограничено в силу постоянно уменьшающегося числа степеней свободы, сопровождающегося увеличением стандартных ошибок и ухудшением качества оценок, а также возможности мультиколлинеарности. Кроме того, при неправильном определении числа лагов возможны ошибки спецификации.

2. Метод геометрической прогрессии (метод Койка)

В распределении Койка предполагается, что коэффициенты («веса») β_k при лаговых значениях объясняющей переменной убывают в геометрической прогрессии:

$$\beta_k = \beta_0 \lambda^k, k = 0,1\ldots, \tag{7.5}$$

где $0 < \lambda < 1$ характеризует скорость убывания коэффициентов с увеличением лага (с удалением от момента анализа). Такое предположение достаточно логично, если считать, что влияние прошлых значений объясняющих переменных на текущее значение зависимой переменной будет тем меньше, чем дальше повремени эти показатели имели место.

В данном случае (7.3) преобразуется в:

$$y_t = \alpha + \beta_0 x_t + \beta_0 \lambda x_{t-1} + \beta_0 \lambda^2 x_{t-2} + \ldots + \varepsilon_t \tag{7.6}$$

Параметры уравнения (7.6) можно определять различными способами.

- Одним из них является перебор значений λ из интервала (0;1) с произвольным фиксированным шагом (например, 0,01; 0,001 и др.). Для каждого λ рассчитывается:

$$z_t = x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \lambda^3 x_{t-3} + \ldots + \lambda^p x_{t-p} \tag{7.7}$$

Значение p определяется из условия, что при дальнейшем добавлении лаговых значений x величина изменения z_t менее любого ранее заданного числа.

Далее оценивается уравнение регрессии:

$$y_t = \alpha + \beta_0 z_t + \varepsilon_t \tag{7.8}$$

Из всех возможных значений λ выбирается то, при котором коэффициент детерминации R^2 для уравнения (7.8) будет наибольшим. Найденные при этом параметры α, β_0, λ подставляются в (7.6).

- Более распространенной является схема вычислений на основе преобразования Койка. Для этого определим уравнение (7.6), умноженное на λ и вычисленное для предыдущего периода времени:

$$\lambda y_{t-1} = \lambda\alpha + \beta_0\lambda x_{t-1} + \beta_0\lambda^2 x_{t-2} + \ldots + \lambda\varepsilon_t \qquad (7.9)$$

Из уравнения (7.6) вычтем уравнение (7.9):

$$y_t - \lambda y_{t-1} = (1-\lambda)\alpha + \beta_0 x_t + (\varepsilon_t - \lambda\varepsilon_{t-1}) \Rightarrow$$

$$y_t = (1-\lambda)\alpha + \beta_0 x_t + \lambda y_{t-1} + v_t \qquad (7.10)$$

где $v_t = \varepsilon_t - \lambda\varepsilon_{t-1}$ - скользящая средняя между ε_t и ε_{t-1}.

Преобразование по данному методу уравнения (7.3) в уравнение (7.10) называется *преобразованием Койка*. Таким образом, с помощью данного преобразования уравнение с бесконечным числом лагов сведено к авторегрессионному, для которого требуется определить всего три коэффициента: λ, α, β_0.

7.3. Авторегрессионные модели

Рассмотрим два вида авторегрессионных моделей.

1. Модель адаптивных ожиданий

Ожидания играют существенную роль в экономической активности, что затрудняет моделирование соответствующих экономических процессов. Особенно серьезна эта проблема на макроэкономическом уровне. Например, прогнозирование объема инвестиций требуется учитывать не только процентную ставку, но и экономическую политику государства, на основе которой потенциальные инвесторы принимают свои решения.

Одним из направлений решения рассматриваемой задачи является модель адаптивных ожиданий. В данной модели происходит постоянная корректировка ожиданий на основе получаемой информации о реализации исследуемого показателя. Если реальное значение показателя оказалось больше ожидаемого, то ожидаемое в следующем периоде корректируется в сторону увеличения. В противном случае – наоборот. При этом величина корректировки должна быть пропорциональна разности между реальным и ожидаемым значениями.

В данной модели в уравнение регрессии в качестве объясняющей переменной вместо текущего значения x_t вводится ожидаемое значение x_t^*:

$$y_t = \alpha + \beta x_t^* + \varepsilon_t \tag{7.11}$$

Так как ожидаемые значения не являются фактически существующими, выдвигается предположение, что эти значения связаны следующим соотношением:

$$x_t^* - x_{t-1}^* = \gamma(x_t - x_{t-1}^*) \tag{7.12}$$

Модель (7.12) называется *моделью адаптивных ожиданий* (или модель обучения на ошибках). Коэффициент $0 \le \gamma \le 1$ называется *коэффициентом ожидания*. Иногда в модели (7.12) вместо текущего значения x_t используют предыдущее значение x_{t-1}:

$$x_t^* - x_{t-1}^* = \gamma(x_{t-1} - x_{t-1}^*) \tag{7.13}$$

Перепишем соотношение (7.12) в виде:

$$x_t^* = \gamma x_t + (1-\gamma)x_{t-1}^* \tag{7.14}$$

Из (7.14) видно, что ожидаемое значение x_t^* является взвешенным средним между текущим значением x_t и его ожидаемым значением в предыдущий период x_{t-1}^* с весами γ и $(1-\gamma)$ соответственно. Если $\gamma = 0$, то ожидания являются неизменными: $x_t^* = x_{t-1}^*$. Если $\gamma = 1$, то $x_t^* = x_t$, что означает мгновенно реализуемые ожидания.

Подставим (7.14) в (7.11), в результате чего получим:

$$y_t = \alpha + \beta(\gamma x_t + (1-\gamma)x_{t-1}^*) + \varepsilon_t \tag{7.15}$$

Определим по (7.15) значение в предыдущий момент времени, умноженное на $(1-\gamma)$:

$$(1-\gamma)y_{t-1} = \alpha(1-\gamma) + \beta(1-\gamma)(\gamma x_t + (1-\gamma)x_{t-1}^*) + \varepsilon_t(1-\gamma) \tag{7.16}$$

Из (7.15) отнимем (7.16):

$$y_t - (1-\gamma)y_{t-1} = (\alpha - \alpha(1-\gamma)) + \beta\gamma x_t + \beta(1-\gamma)(x_{t-1}^* - \gamma x_{t-1} + (1-\gamma)x_{t-2}^*) + (\varepsilon_t - \varepsilon_t(1-\gamma))$$

Так как из (7.14) $x_{t-1}^* = \gamma x_{t-1} + (1-\gamma)x_{t-2}^*$, то:

$$\begin{gathered} y_t - (1-\gamma)y_{t-1} = \alpha\gamma + \beta\gamma x_t + (\varepsilon_t - \varepsilon_t(1-\gamma)) \Rightarrow \\ y_t = \alpha\gamma + \beta\gamma x_t + (1-\gamma)y_{t-1} + v_t \end{gathered} \tag{7.17}$$

где $v_t = \varepsilon_t - (1-\gamma)\varepsilon_{t-1}$.

На практике при оценивании параметров авторегрессионного уравнения (7.17) вначале оценивается параметр γ , затем коэффициент при x_t $(\beta = \frac{\beta\gamma}{\gamma})$, затем свободный член $(\alpha = \frac{\alpha\gamma}{\gamma})$.

Рассмотрим случай, когда зависимая переменная y_t в текущий момент времени связана с ожидаемым в следующий период времени значением x^*_{t+1} (например, зависимость спроса на деньки от ожидаемой процентной ставки), т.е.:

$$y_t = \alpha + \beta x^*_{t+1} + \varepsilon_t \tag{7.18}$$

Допустим, что ожидаемое в следующий период времени значение переменной определяется как взвешенное среднее ее реального и ожидаемого значения в текущий период времени (аналогично 7.14):

$$x^*_{t+1} = \gamma x_t + (1-\gamma)x^*_t \tag{7.19}$$

Следовательно, $x^*_t = \gamma x_{t-1} + (1-\gamma)x^*_{t-1}$. Отсюда (7.19) примет вид:

$$x^*_{t+1} = \gamma x_t + (1-\gamma)(\gamma x_{t-1} + (1-\gamma)x^*_{t-1}) = \gamma x_t + \gamma(1-\gamma)x_{t-1} + (1-\gamma)^2 x^*_{t-1}$$

Из (7.19) следует, что $x^*_{t-1} = \gamma x_{t-2} + (1-\gamma)x^*_{t-2}$, $x^*_{t-2} = \gamma x_{t-3} + (1-\gamma)x^*_{t-3}$ и т.д. С учетом этого (7.19) примет вид:

$$x^*_{t+1} = \gamma[x_t + (1-\gamma)x_{t-1} + (1-\gamma)^2 x_{t-2} + \ldots] \tag{7.20}$$

Подставив (7.20) в (7.18), получаем:

$$y_t = \alpha + \beta\gamma[x_t + (1-\gamma)x_{t-1} + (1-\gamma)^2 x_{t-2} + \ldots] + \varepsilon_t \tag{7.21}$$

Обозначив $\beta\gamma$ через β_0, и $(1-\gamma)$ через λ, получаем соотношение (7.6).

2. Модель частичной корректировки

В этой модели в уравнение регрессии в качестве зависимой переменной входит не фактическое значение y_t, а желаемое значение y^*_t:

$$y^*_t = \alpha + \beta x_t + \varepsilon_t \tag{7.22}$$

Так как гипотетическое значение y^*_t не является фактически существующим, то относительно него выдвигается предположение *частичной корректировки*:

$$y_t - y_{t-1} = \lambda(y^*_t - y_{t-1}) \tag{7.23}$$

по которому фактическое приращение зависимой переменной пропорционально разнице между ее желаемым значением и значением в предыдущий период. Коэффициент $0 \le \lambda \le 1$ называется *коэффициентом корректировки*. Уравнение (7.23) преобразуется к виду:

$$y_t = \lambda y_t^* + (1-\lambda) y_{t-1} \tag{7.24}$$

Подставив (7.22) в (7.24), получаем *модель частичной корректировки*:

$$y_t = \lambda(\alpha + \beta x_t + \varepsilon_t) + (1-\lambda) y_{t-1} \Rightarrow y_t = \lambda\alpha + \lambda\beta x_t + (1-\lambda) y_{t-1} + \lambda\varepsilon_t \tag{7.25}$$

Очевидно, что чем больше λ, тем быстрее идет корректировка. При $\lambda = 1$ полная корректировка происходит за один период. При $\lambda = 0$ корректировка не происходит вовсе.

7.4. Прогнозирование с помощью временных рядов

Конечной целью статистического анализа временных рядов является прогнозирование будущих значений исследуемого показателя. Различают долгосрочное и краткосрочное прогнозирование. В первом анализируется долговременная динамика исследуемого процесса, и в этом случае главным считается выделение общего направления его изменения (тренда). Для предсказания краткосрочных колебаний проводится более детальный регрессионный анализ с целью выявления большого числа показателей, определяющих поведение исследуемой величины.

Пусть оценивается модель вида $\hat{y}_t = b_0 + b_1 x_t$ в момент времени ($t + p$). Значение $\hat{y}_{t+p}$ - значение по уравнению регрессии, построенному по МНК. Тогда доверительный интервал для действительного значения y_{t+p} имеет вид:

$$\left(\hat{y}_{t+p} \pm t_{\frac{\alpha}{2}, n-2} S_e \sqrt{1 + \frac{1}{n} + \frac{(x_{t+p} - \bar{x})^2}{n S_x^2}} \right) \tag{7.26}$$

где $t_{\frac{\alpha}{2}, n-2}$ - критическое значение, определяемое из приложения 1 для соответствующего уровня значимости α и числа степеней свободы $(n-2)$;

$S_e = \sqrt{\dfrac{\sum e_i^2}{n-2}}$ - стандартная ошибка оценки (стандартная ошибка регрессии);

x_{t+p} - значение объясняющей переменной в момент $(t+p)$;

$S_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}$ - дисперсия переменной x_t.

После получения прогнозных значений необходимо проверить качество прогноза. Для этого используются следующие показатели:

- Относительная ошибка прогноза, вычисляемая по формуле:

$$\delta_{t+p} = \frac{\hat{y}_{t+p} - y_{t+p}}{y_{t+p}} \cdot 100\% \qquad (7.27)$$

или

$$\delta_{t+p} = \frac{\Delta\hat{y}_{t+p} - \Delta y_{t+p}}{\Delta y_{t+p}} \cdot 100\% \qquad (7.28)$$

где $\Delta\hat{y}_{t+p} = \hat{y}_{t+p} - \hat{y}_t$, $\Delta y_{t+p} = y_{t+p} - y_t$.

Чем больше значение ошибки (выраженное в процентах), тем хуже качество прогноза.

- Стандартная среднеквадратическая ошибка, рассчитываемая по формуле:

$$U = \frac{\frac{1}{k}\sum_{p=1}^{k}(\Delta\hat{y}_{t+p} - \Delta y_{t+p})^2}{\frac{1}{k}\sum \Delta y_{t+p}^2} \qquad (7.29)$$

где k - количество прогнозных периодов.

Значения показателя U лежат в интервале от нуля до единицы. При $U = 0$ прогноз абсолютно точен. Таким образом, чем ближе значение U к нулю, тем точнее прогноз.

Пример. В таблице приведены данные по располагаемому доходу домохозяйств (X) и затратами домохозяйств на розничные покупки (Y) за 22 года:

t	Y_t	X_t	t	Y_t	X_t
1	5,49	9,098	12	5,905	11,305
2	5,54	9,137	13	6,125	11,43
3	5,305	9,095	14	6,185	11,45
4	5,505	9,28	15	6,225	11,697
5	5,42	9,23	16	6,495	11,87

6	5,32	9,348	17	6,72	12,018
7	5,54	9,525	18	6,92	12,525
8	5,69	9,755	19	6,47	12,055
9	5,87	10,28	20	6,395	12,088
10	6,157	10,665	21	6,555	12,215
11	6,342	11,02	22	6,755	12,495

Необходимо оценить уравнение регрессии вида $\hat{y}_t = b + b_1 x_t + \gamma y_{t-1}$ (принять $y_0 = 5{,}4$), проверить значимость коэффициентов b_0, b_1, γ, оценить качество построенной модели при помощи коэффициента детерминации.

Для расчета коэффициентов составим вспомогательную таблицу (при этом рассчитанные средние значения равны $\bar{y}_t = 6{,}04223$, $\bar{y}_{t-1} = 5{,}98067$, $\bar{x} = 10{,}79914$):

t	Y_t	X_t	Y_{t-1}	$(y_t-\bar{y}_t)$	$(x_t-\bar{x})$	$(y_{t-1}-\bar{y}_{t-1})$	$(y_t-\bar{y}_t)^2$	$(x_t-\bar{x})^2$	$(y_{t-1}-\bar{y}_{t-1})^2$	$(x_t-\bar{x})\cdot(y_t-\bar{y}_t)$	$(y_{t-1}-\bar{y}_{t-1})\cdot(y_i-\bar{y}_t)$	$(x_t-\bar{x})\cdot(y_{t-1}-\bar{y}_{t-1})$
1	5,49	9,098	5,4	-0,5522	-1,7011	-0,581	0,3050	2,8939	0,3371	0,9394	0,3206	0,9877
2	5,54	9,137	5,49	-0,5022	-1,6621	-0,491	0,2522	2,7627	0,2407	0,8348	0,2464	0,8155
3	5,305	9,095	5,54	-0,7372	-1,7041	-0,441	0,5435	2,9041	0,1942	1,2563	0,3248	0,7509
4	5,505	9,28	5,305	-0,5372	-1,5191	-0,676	0,2886	2,3078	0,4565	0,8161	0,3630	1,0264
5	5,42	9,23	5,505	-0,6222	-1,5691	-0,476	0,3872	2,4622	0,2262	0,9764	0,2960	0,7463
6	5,32	9,348	5,42	-0,7222	-1,4511	-0,561	0,5216	2,1058	0,3143	1,0481	0,4049	0,8136
7	5,54	9,525	5,32	-0,5022	-1,2741	-0,661	0,2522	1,6234	0,4364	0,6399	0,3318	0,8417
8	5,69	9,755	5,54	-0,3522	-1,0441	-0,441	0,1241	1,0902	0,1942	0,3678	0,1552	0,4601
9	5,87	10,28	5,69	-0,1722	-0,5191	-0,291	0,0297	0,2695	0,0845	0,0894	0,0501	0,1509
10	6,157	10,665	5,87	0,1148	-0,1341	-0,111	0,0132	0,0180	0,0122	-0,0154	-0,0127	0,0148
11	6,342	11,02	6,157	0,2998	0,2209	0,176	0,0899	0,0488	0,0311	0,0662	0,0529	0,0390
12	5,905	11,305	6,342	-0,1372	0,5059	0,361	0,0188	0,2559	0,1306	-0,0694	-0,0496	0,1828
13	6,125	11,43	5,905	0,0828	0,6309	-0,076	0,0069	0,3980	0,0057	0,0522	-0,0063	-0,0477
14	6,185	11,45	6,125	0,1428	0,6509	0,144	0,0204	0,4236	0,0208	0,0929	0,0206	0,0940
15	6,225	11,697	6,185	0,1828	0,8979	0,204	0,0334	0,8062	0,0418	0,1641	0,0374	0,1835
16	6,495	11,87	6,225	0,4528	1,0709	0,244	0,2050	1,1467	0,0597	0,4849	0,1106	0,2617
17	6,72	12,018	6,495	0,6778	1,2189	0,514	0,4594	1,4856	0,2646	0,8261	0,3486	0,6269
18	6,92	12,525	6,72	0,8778	1,7259	0,739	0,7705	2,9786	0,5467	1,5149	0,6490	1,2760
19	6,47	12,055	6,92	0,4278	1,2559	0,939	0,1830	1,5772	0,8824	0,5372	0,4018	1,1797
20	6,395	12,088	6,47	0,3528	1,2889	0,489	0,1244	1,6612	0,2395	0,4547	0,1726	0,6307
21	6,555	12,215	6,395	0,5128	1,4159	0,414	0,2629	2,0047	0,1717	0,7260	0,2125	0,5867
22	6,755	12,495	6,555	0,7128	1,6959	0,574	0,5080	2,8760	0,3299	1,2088	0,4094	0,9740
Σ							5,3998	34,1000	5,2208	13,0114	4,8397	12,5953

По формулам (2.17):

$$b_1 = \frac{(13{,}0114)\cdot 5{,}2208 - 4{,}8397\cdot 12{,}5953}{34{,}100\cdot 5{,}2208 - (12{,}5953)^2} = \frac{6{,}9724}{19{,}3877} = 0{,}36$$

$$\gamma = \frac{4{,}8397\cdot 34{,}10 - 13{,}0114\cdot 12{,}5953}{34{,}100\cdot 5{,}2208 - (12{,}5953)^2} = \frac{1{,}1513}{19{,}3877} = 0{,}06$$

$$b_0 = 6{,}04223 - 0{,}36\cdot 10{,}79914 - 0{,}06\cdot 5{,}98067 = 1{,}80$$

Таким образом, уравнение регрессии с учетом рассчитанных коэффициентов примет вид: $\hat{y}_t = 1{,}8 + 0{,}36x_t + 0{,}06y_{t-1}$.

Для определения статистической значимости коэффициентов и оценки качества уравнения регрессии составим следующую вспомогательную таблицу:

t	Y_t	$\hat{Y}_t$	e_i	e_i^2
1	5,49	5,3960	-0,0940	0,008843
2	5,54	5,4153	-0,1247	0,015542
3	5,305	5,4032	0,0982	0,009642
4	5,505	5,4558	-0,0492	0,002422
5	5,42	5,4497	0,0297	0,00088
6	5,32	5,4871	0,1671	0,02791
7	5,54	5,5448	0,0048	2,29E-05
8	5,69	5,6406	-0,0494	0,002444
9	5,87	5,8383	-0,0317	0,001006
10	6,157	5,9874	-0,1696	0,028757
11	6,342	6,1321	-0,2099	0,044047
12	5,905	6,2456	0,3406	0,11601
13	6,125	6,2646	0,1396	0,019496
14	6,185	6,2849	0,0999	0,009975
15	6,225	6,3773	0,1523	0,023186
16	6,495	6,4419	-0,0531	0,002824
17	6,72	6,5111	-0,2089	0,043635
18	6,92	6,7068	-0,2132	0,045453
19	6,47	6,5496	0,0796	0,006342
20	6,395	6,5348	0,1398	0,019545
21	6,555	6,5760	0,0210	0,000442
22	6,755	6,6862	-0,0688	0,00473
сумма			≈ 0	0,433155

По формулам (2.19) и (2.20) рассчитаем необъясненную дисперсию и стандартные отклонения случайных величин:

$$S^2 = \frac{\sum e_i^2}{n-m-1} = \frac{0,43155}{22-2-1} = 0,0227 \ .$$

$$S_{b0}^2 = \left[\frac{1}{22} + \frac{(10,79914)^2 \cdot 5,2208 + (5,98067)^2 \cdot 34,1 - 2 \cdot 10,79914 \cdot 5,98067 \cdot 12,5953}{34,1 \cdot 5,2208 - (12,5953)^2} \right] \cdot 0,0227$$

$$S_{b0} = \sqrt{S_{b0}^2} = \sqrt{0,238} = 0,4879$$

$$S_{b1}^2 = \frac{5,2208}{34,1 \cdot 5,2208 - (12,5953)^2} \cdot 0,0227 = 0,0061$$

$$S_{b1} = \sqrt{S_{b1}^2} = \sqrt{0,0061} = 0,0781$$

$$S_{\gamma}^2 = \frac{34,10}{34,1 \cdot 5,2208 - (12,5953)^2} \cdot 0,0227 = 0,0399$$

$$S_{\gamma} = \sqrt{S_{\gamma}^2} = \sqrt{0,0399} = 0,1997$$

Определим значение t-статистики для каждого из коэффициентов:

$$t_{b0} = \frac{1,8}{0,4879} = 3,689 \ , \ t_{b1} = \frac{0,36}{0,0781} = 4,609 \ , \ t_{\gamma} = \frac{0,06}{0,1997} = 0,300 \ .$$

Критическое значение определим из приложения 1 для уровня значимости 0,1 и числа степеней свободы $\nu = 22\text{-}2\text{-}1 = 19$: $t_{кр} = t_{\frac{0,1}{2},19} = 1,729$.

Очевидно, что коэффициенты b_0, b_1 являются статистически значимыми, а коэффициент γ является статистически незначимым с уровнем значимости 0,1.

Определим для рассчитанного уравнения коэффициент детерминации (2.23): $R^2 = 1 - \dfrac{0,433155}{5,3998} = 0,92$. Столь высокое значение коэффициента детерминации свидетельствует о высоком качестве модели. Поэтому не будем удалять переменную y_{t-1} из уравнения.

Представим графически зависимость фактической переменной y_t и переменной $\hat{y}_t$ от t:

8. Системы одновременных уравнений

8.1. Необходимость использования систем уравнений

Многие экономические взаимосвязи допускают моделирование одним уравнением. В большинстве случаев моделирование МНК для оценки параметров таких моделей является наиболее подходящей процедурой. Однако ряд экономических процессов моделируется не одним, а несколькими уравнениями, содержащими как повторяющиеся, так и собственные переменные. В силу этого возникает необходимость использования систем уравнений. Кроме того, в одних уравнениях определенная переменная рассматривается как объясняющая, но в то же время она входит в другое уравнение как зависимая переменная. Приведем примеры таких систем.

1. Модель «спрос-предложение»

Данная модель является одной из простейших систем одновременных уравнений. В этом случае, предполагая, что спрос Q^D и предложение Q^S в момент времени t являются линейными функциями от цены P, получаем систему:

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \varepsilon_{t1},\ \alpha_1 < 0 & \text{- функция спроса} \\ q_t^S = \beta_0 + \beta_1 p_t + \varepsilon_{t2},\ \beta_1 > 0 & \text{- функция предложения} \\ q_t^D = q_t^S & \text{- условие равновесия} \end{cases} \quad (8.1)$$

Наличие случайных отклонений в данных моделях связано прежде всего с отсутствием ряда важных объясняющих переменных (дохода, цен сопутствующих товаров, вкусов и т.д.). Очевидно, что изменение одного из этих факторов может отразиться на модели.

Однако модель (8.1) может быть усовершенствована. Например, можно в функцию спроса добавить доход потребителя Y :

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \alpha_2 y_t + \varepsilon_{t1},\ \alpha_1 < 0 & \text{- функция спроса} \\ q_t^S = \beta_0 + \beta_1 p_t + \varepsilon_{t2},\ \beta_1 > 0 & \text{- функция предложения} \\ q_t^D = q_t^S & \text{- условие равновесия} \end{cases} \quad (8.2)$$

2. Кейнсианская модель формирования доходов

Рассмотрим простейшую модель данного типа в предположении, что рассматривается закрытая экономика без государственных расходов:

$$\begin{cases} c_t = \beta_0 + \beta_1 y_t + \varepsilon_t & \text{- функция потребления} \\ y_t = c_t + i_t & \text{- макроэкономическое\_тождество} \end{cases} \quad (8.3)$$

Здесь Y, C, I - совокупный выпуск, объемы потребления и инвестиций соответственно (y_t, c_t, i_t - значения этих переменных в момент времени t).

3. Модели IS-LM

Одной из возможных нестохастических форм модели IS (равновесия на рынке товаров) является следующая модель:

$$\begin{cases} c_t = \beta_0 + \beta_1 y_{(d)t} & \text{- функция потребления} \\ \tau_t = \alpha_0 + \alpha_1 y_t & \text{- функция налогов} \\ i_t = \gamma_0 + \gamma_1 r_t & \text{- функция инвестиций} \\ y_{(d)t} = y_t - \tau_t & \text{- располагаемый доход} \\ g_t = \bar{g} & \text{- государственные расходы} \\ y_t = c_t + i_t + g_t & \text{- макроэкономическое\_тождество} \end{cases} \quad (8.4)$$

Здесь $y_t, c_t, i_t, g_t, \tau_t, y_{(d)t}, r_t$ - значения в момент времени t национального дохода Y, потребления C, желаемого объема чистых инвестиций I, государственных расходов G (в данном случае $G = \bar{g} = const$), объема налогов T, располагаемого дохода $Y_{(d)}$, процентной ставки r.

Преобразуем макроэкономическое тождество:

$$y_t = c_t + i_t + g_t = \beta_0 + \beta_1(y_t - \tau_t) + \gamma_0 + \gamma_1 r_t + \bar{g} =$$

$$= \beta_0 + \beta_1(y_t - (\alpha_0 + \alpha_1 y_t)) + \gamma_0 + \gamma_1 r_t + \bar{g}$$

Получаем:

$$y_t = \pi_0 + \pi_1 r_t \tag{8.5}$$

где $\pi_0 = \dfrac{\beta_0 - \alpha_0\beta_1 + \gamma_0 + \bar{g}}{1 - \beta_1(1 - \alpha_1)}, \pi_1 = \dfrac{\gamma_1}{1 - \beta_1(1 - \alpha_1)}$.

Формула (8.5) является выражением кривой IS, задающей такое соотношение между процентной ставкой и уровнем дохода, при котором рынок товаров находится в равновесии.

Линия LM (линия равновесия на рынке денег) задает такое соотношение между процентной ставкой и уровнем дохода, при спрос на деньги равен их предложению. Одна из нестохастических форм данной модели имеет вид:

$$\begin{cases} M_t^D = a + by_t - cr_t & \text{- функция спроса на деньги} \\ M_t^S = \overline{M} & \text{- функция предложени я денег} \\ M_t^D = M_t^S & \text{- условие равновесия} \end{cases} \tag{8.6}$$

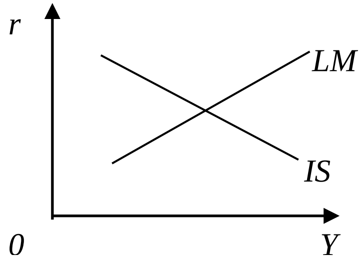
Тогда функцию спроса на деньги можно представить в виде:

$$y_t = \lambda_0 + \lambda_1 \overline{M} + \lambda_2 r_t \tag{8.7}$$

где $\lambda_0 = -a/b, \lambda_1 = 1/b, \lambda_2 = c/b$.

Соотношение (8.7) известно как уравнение LM.

Модель IS-LM представлена на рисунке:



Точка пересечения данных кривых определяет соотношение между процентной ставкой и уровнем дохода, при котором оба рынка находятся в состоянии равновесия.

8.2. Составляющие системы уравнений

При рассмотрении систем одновременных уравнений переменные делятся на два класса: *эндогенные переменные* – переменные, значения которых определяются внутри модели и *экзогенные переменные* – внешние по отношению к модели переменные.

Для системы (8.1) все переменные - q_t^D, q_t^S, P - являются эндогенными, так они определяются внутри системы. Для системы (8.3) переменные C, Y являются эндогенными, а переменная I - экзогенная, так как она определяется вне модели.

Уравнения, составляющие исходную модель, называют *структурными уравнениями модели*. Их подразделяют на *поведенческие уравнения* (описывают взаимодействия между переменными) и *уравнения-тождества* (соотношения, которые должны выполняться во всех случаях).

Уравнения, в которых определена схема определения эндогенных переменных через экзогенные или предопределенные переменные, называются *уравнениями в приведенной форме (приведенными уравнениями)*.

Предопределенные переменные – лаговые эндогенные переменные, значения которых определены до рассмотрения соотношения.

8.3. Косвенный метод наименьших квадратов (КМНК)

Качественно оценить параметры системы одновременных уравнений на основе МНК невозможно в силу получения смещенных и несостоятельных оценок. Поэтому для получения «хороших» оценок требуется использования других методов. Одним из них является косвенный метод наименьших квадратов (КМНК).

Для иллюстрации КМНК рассмотрим модель (8.3). Данную модель можно свести к виду:

$$\begin{cases} y_t = \dfrac{\beta_0}{1-\beta_1} + \dfrac{1}{1-\beta_1} i_t + \dfrac{\varepsilon_t}{1-\beta_1} \\ c_t = \dfrac{\beta_0}{1-\beta_1} + \dfrac{\beta_1}{1-\beta_1} i_t + \dfrac{\varepsilon_t}{1-\beta_1} \end{cases} \tag{8.8}$$

Обозначив: $\dfrac{\beta_0}{1-\beta_1}=\lambda_{10}, \dfrac{1}{1-\beta_1}\lambda_{11}, \dfrac{\varepsilon_t}{1-\beta_1}=v_t,$

$\dfrac{\beta_0}{1-\beta_1}=\lambda_{20}$, $\dfrac{\beta_1}{1-\beta_1}=\lambda_{21}$, получаем:

$$\begin{cases} y_t = \lambda_{10} + \lambda_{11} i_t + v_t \\ c_t = \lambda_{20} + \lambda_{21} i_t + v_t \end{cases} \qquad (8.9)$$

Определив оценки $\hat{\lambda}_{10}, \hat{\lambda}_{11}$, $\hat{\lambda}_{20}, \hat{\lambda}_{21}$ по МНК для каждого из уравнений системы (8.9), можно определить оценки коэффициентов β_0, β_1:

$$b_0 = \hat{\beta}_0 = \frac{\hat{\lambda}_{20}}{\hat{\lambda}_{11}}, \; b_1 = \hat{\beta}_1 = \frac{\hat{\lambda}_{21}}{\hat{\lambda}_{11}} \qquad (8.10)$$

Определение оценок по указанной схеме называется *косвенным методом наименьших квадратов*. Смысл такого названия очевиден: оценки b_0, b_1 вычисляются не напрямую по МНК, а через коэффициенты приведенных уравнений. Полученные оценки b_0, b_1 являются состоятельными.

Пример. Рассмотрим следующую модель (первая функция – предложение, вторая – спрос):

$$\begin{cases} q_t = \beta_0 + \beta_1 p_t + \varepsilon_{1t} \\ q_t = \alpha_0 + \alpha_1 p_t + \alpha_2 y_t + \varepsilon_{2t} \end{cases}$$

где q_t, p_t - эндогенные переменные – количество товара и цена в году t; y_t - экзогенная переменная – доход потребителей; $\varepsilon_{1t}, \varepsilon_{2t}$ - случайные отклонения.

На основании следующих статистических данных необходимо оценить коэффициенты функции предложения, используя для этого МНК и КМНК.

	p_t	q_t	y_t	p_t^2	y_t^2	$p_t q_t$	$p_t y_t$	$q_t y_t$
	1	8	2	1	4	8	2	16
	2	10	4	4	16	20	8	40
	3	7	3	9	9	21	9	21
	4	5	5	16	25	20	20	25
	5	1	2	25	4	5	10	2
Сумма	15	31	16	55	58	74	49	104
Среднее	3	6,2	3,2	11	11,6	14,8	9,8	20,8

Построим приведенные уравнения данной системы. Для этого вычтем из функции предложения функцию спроса:

$$(\beta_0 - \alpha_0) + (\beta_1 - \alpha_1)p_t - \alpha_2 y_t + (\varepsilon_{1t} - \varepsilon_{2t}) = 0$$

Выразим отсюда p_t: $p_t = \dfrac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \dfrac{\alpha_2}{\beta_1 - \alpha_1} y_t + \dfrac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \alpha_1}$.

Подставим полученное значение p_t в функцию предложения:

$$q_t = \beta_0 + \beta_1 \left(\frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{\alpha_2}{\beta_1 - \alpha_1} y_t + \frac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \alpha_1}\right) + \varepsilon_{1t}.$$

Таким образом, приведенные уравнения имеют вид:

$$\begin{cases} p_t = \pi_{10} + \pi_{11} y_t + v_{1t} \\ q_t = \pi_{20} + \pi_{21} y_t + v_{2t} \end{cases}$$

где $\pi_{10} = \dfrac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1}$, $\pi_{11} = \dfrac{\alpha_2}{\beta_1 - \alpha_1}$, $v_{1t} = \dfrac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \alpha_1}$

$$\pi_{20} = \beta_0 + \beta_1 \pi_{10}, \pi_{21} = \beta_1 \pi_{11}, v_{2t} = \beta_1 v_{1t} + \varepsilon_{1t}$$

По имеющимся статистическим данным, приведенным в таблице, оценим коэффициенты приведенных уравнений (по обычному МНК, формула (1.11)):

$$\hat{\pi}_{11} = \frac{\overline{yp} - \bar{y} \cdot \bar{p}}{\overline{y^2} - (\bar{y})^2} = \frac{9{,}8 - 3{,}2 \cdot 3}{11{,}6 - (3{,}2)^2} = 0{,}1471, \hat{\pi}_{10} = \bar{p} - \hat{\pi}_{11}\bar{y} = 2{,}5293,$$

$$\hat{\pi}_{21} = \frac{\overline{yq} - \bar{y} \cdot \bar{q}}{\overline{y^2} - (\bar{y})^2} = \frac{20{,}8 - 3{,}2 \cdot 6{,}2}{11{,}6 - (3{,}2)^2} = 0{,}7059, \quad \hat{\pi}_{20} = \bar{q} - \hat{\pi}_{21}\bar{y} = 3{,}9411.$$

Теперь можно оценить оценки b_0 и b_1 параметров β_0, β_1:

$$\beta_1 = \frac{\pi_{21}}{\pi_{11}}, \beta_0 = \pi_{20} - \beta_1 \pi_{10} \Rightarrow$$

$$b_1 = \frac{\hat{\pi}_{21}}{\hat{\pi}_{11}} = \frac{0{,}7059}{0{,}1471} = 4{,}7988, b_0 = \hat{\pi}_{20} - b_1 \hat{\pi}_{10} = 3{,}9411 - 4{,}7988 \cdot 2{,}5293 = -8{,}1965$$

Следовательно, функция предложения, рассчитанная по косвенному методу наименьших квадратов, имеет вид:

$$\hat{q}_t = -8{,}1965 + 4{,}7988\, p_t.$$

Далее необходимо проверить статистическую значимость коэффициентов и качество регрессии. Эти вычисления производятся аналогично примеру в главе 1.

В то же время, если рассчитать оценки b_0 и b_1 параметров β_0, β_1 непосредственно по МНК, то получим функцию предложения:

$$\hat{q}_t = 11{,}9 - 1{,}9 p_t$$

Полученные результаты позволяют сделать вывод о том, что применение МНК в несоответствующих ситуациях может существенно исказить картину действительности.

8.4. Проблема идентификации

Изменение формы уравнений хотя и позволяет устранить проблему коррелированности объясняющей переменной и случайного отклонения, но может привести к другой, не менее серьезной проблеме – *проблеме идентификации*. Под проблемой идентификации понимается возможность численной оценки параметров структурных уравнений по оценкам коэффициентов приведенных уравнений.

Исходную систему уравнений называют *идентифицируемой (точно определенной)*, если по коэффициентам приведенных уравнений можно однозначно определить значения коэффициентов структурных уравнений. Обычно это удается сделать, когда количество уравнений для определения коэффициентов структурных уравнений в точности равно количеству этих коэффициентов.

Исходную систему уравнений называют *неидентифицируемой (недоопределенной)*, если по коэффициентам приведенных уравнений можно получить несколько вариантов значений коэффициентов структурных уравнений. Обычно это происходит, когда количество уравнений для определения коэффициентов структурных уравнений меньше числа определяемых коэффициентов.

Исходную систему уравнений называют *сверхидентифицируемой (переопределенной)*, если по коэффициентам приведенных уравнений невозможно определить значения коэффициентов структурных уравнений. В этом случае система, связывающая коэффициенты структурных уравнений с коэффициентами приведенных уравнений, является несовместной. Обычно в таких случаях число уравнений для оценки коэффициентов структурных уравнений больше числа определяемых коэффициентов.

Для быстрого формального определения идентифицируемости структурных уравнений применяются следующие необходимые и достаточные условия. Пусть система одновременных уравнений включает в себя N уравнений относительно N эндогенных переменных. пусть в системе имеется M экзогенных или предопределенных переменных. Пусть количество эндогенных и экзогенных переменных в проверяемом на идентифицируемость уравнении равно n и m соответственно. Переменные, не входящие в данное уравнение, но входящие в другие уравнение системы, назовем *исключенными переменными* (из данного уравнения). Их количество равно $(N-n)$ для эндогенных и $(M-m)$ для экзогенных переменных.

Первое необходимое условие

Уравнение идентифицируемо, если оно исключает, по крайней мере, $(N-1)$ переменную (эндогенную или экзогенную), присутствующую в модели: $(N-n)+(M-m)\ge N-1$

Второе необходимое условие

Уравнение идентифицируемо, если количество исключенных из уравнения экзогенных переменных не меньше количества эндогенных переменных в этом уравнении, уменьшенного на единицу: $M-m\ge n-1$.

Знаки равенства в обоих необходимых условиях соответствуют точной идентификации уравнения. Знак «>» свидетельствует о переопределенности. Знак «<» свидетельствует о недоопределнности.

Пример.

А) для модели «спрос-предложение» проверим условия идентифицируемости:

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \varepsilon_{t1} \\ q_t^S = \beta_0 + \beta_1 p_t + \varepsilon_{t2} \end{cases}$$

Для системы $N=2, M=0$. Для каждого из уравнений $n=2, m=0$. Следовательно, для обоих уравнений не выполняется первое условие: $(2-2)+(0-0)\ge 2-1$. Это означает, что оба они не идентифицируемы.

Б) В ту же модель введем экзогенную переменную I - доход потребителя:

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \alpha_2 i_t + \varepsilon_{t1} \\ q_t^S = \beta_0 + \beta_1 p_t + \varepsilon_{t2} \end{cases}$$

Для системы $N=2, M=1$. Для первого уравнения $n=2, m=1$. Для второго $n=2, m=0$. Тогда для первого уравнения первое условие не выполняется: $(2-2)+(1-1)\ge 2-1$. Для второго уравнения выполняются первое условие: $(2-2)+(1-0)=2-1$; и второе условие $(1-0)=2-1$. Это означает, что первое уравнение не идентифицируемо, а второе может быть определено однозначно, т.е. является идентифицируемым.

В) В модели:

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \alpha_2 i_t + \varepsilon_{t1} \\ q_t^S = \beta_0 + \beta_1 p_t + \beta_2 p_{t-1} + \varepsilon_{t2} \end{cases}$$

Для системы $N = 2, M = 2$. Для каждого из уравнений $n = 2, m = 1$. Следовательно, для обоих уравнений выполняется первое условие: $(2-2)+(2-1) = 2-1$; и второе условие $(2-1) = 2-1$. Это означает, что оба они идентифицируемы.

Г) В предыдущую модель в функцию спроса введем s_t - объем сбережений к моменту времени t:

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \alpha_2 i_t + \alpha_3 s_t + \varepsilon_{t1} \\ q_t^S = \beta_0 + \beta_1 p_t + \beta_2 p_{t-1} + \varepsilon_{t2} \end{cases}$$

Для системы $N = 2, M = 3$. Для первого уравнения $n = 2, m = 2$. Соответственно для него первое условие: $(2-2)+(3-2) = 2-1$; второе условие: $(3-2) = 2-1$. Для второго $n = 2, m = 1$. Соответственно для второго уравнения первое условие: $(2-2)+(3-1) > 2-1$; второе условие $(3-1) > 2-1$. Это означает, что первое уравнение точно идентифицируемо, а второе является переопределенным.

8.5. Оценка систем уравнений

1. МНК для рекурсивных моделей

Одним из случаев успешного применения МНК для оценки структурных коэффициентов модели является его использование для *рекурсивных (треугольных) моделей*. В этих моделях эндогенные переменные последовательно (рекурсивно) связаны друг с другом. Первая переменная Y_1 зависит лишь от экзогенных переменных $X_i, i = 1,2,\ldots,m$, и случайного отклонения ε_1. Вторая эндогенная переменная Y_2 определяется лишь значениями экзогенных переменных $X_i, i = 1,2,\ldots,m$, случайным отклонением ε_2, а также эндогенной переменной Y_1. Третья эндогенная переменная Y_3 определяется значениями экзогенных переменных $X_i, i = 1,2,\ldots,m$, случайным отклонением ε_3, а также эндогенных переменных Y_1 и Y_2 и т.д.

В этих моделях структурные уравнения оцениваются поэтапно ($Y_1 \to Y_2 \to Y_3 \to \ldots \to Y_N$). Применение МНК для таких моделей позволяет получить несмещенные и состоятельные оценки.

Модели данного типа встречаются достаточно редко.

2. Двухшаговый метод наименьших квадратов

Рассмотрим данный метод на примере для модели IS-LM для закрытой экономики при фиксированной налоговой ставке (t):

$$\begin{cases} Y = \alpha_0 + \alpha_1 r + \alpha_2 G + \alpha_3 t + \varepsilon_1 \quad (\alpha_1 < 0) & (8.11\text{-}1) \\ Y = \beta_0 + \beta_1 r + \beta_2 M + \varepsilon_2 \quad (\beta_1 > 0) & (8.11\text{-}2) \end{cases}$$

Второе уравнение является переопределенным. Для его оценки рекомендуется использовать *двухшаговый метод наименьших квадратов* (ДМНК).

Шаг 1.

В уравнении $Y = \beta_0 + \beta_1 r + \beta_2 M + \varepsilon_2$ переопределенной переменной является процентная ставка r. Ее можно оценить, лишь опираясь на экзогенные переменные (например, вычесть из уравнения (8.11-1) уравнение (8.11-2)):

$$r = \lambda_0 + \lambda_1 M + \lambda_2 G + \lambda_3 t + v \qquad (8.12)$$

Коэффициенты $\lambda_0, \lambda_1, \lambda_2, \lambda_3, v$ предлагается найти самостоятельно по аналогии с ранее рассмотренными примерами.

Применяя для (8.12) МНК, получаем оценку $\hat{r}$ переменной r:

$$\hat{r} = \hat{\lambda}_0 + \hat{\lambda}_1 M + \hat{\lambda}_2 G + \hat{\lambda}_3 t \qquad (8.13)$$

где $\hat{r}$ - условная средняя при фиксированных значениях M, G, t.

Шаг 2.

Подставляя оценку (8.13) в уравнение (8.11-2), имеем:

$$Y = \beta_0 + \beta_1 \hat{r} + \beta_2 M + \varepsilon_2 \qquad (8.14)$$

Данная замена позволяет преодолеть такую существенную проблему переопределенных моделей, как коррелированность объясняющей переменной со случайным членом (что приводит к получению смещенных и несостоятельных оценок). Действительно, оценка $\hat{r}$ выражается только через экзогенные переменные и, следовательно, не коррелирует со случайным членом. Фактически ее можно рассматривать как новую экзогенную переменную.

Заменив в модели (8.11) уравнение (8.11-2) на (8.14), получаем систему, которую можно решать при помощи МНК.

При наличии в модели более одной переопределенной переменной на первом этапе необходимо оценить все такие переменные.

Приложения

Приложение 1

Распределение Стьюдента (t-распределение)

		уровень значимости						
		0.40	0.25	0.10	0.05	0.025	0.01	0.005
число степеней свободы	1	0,325	1,000	3,078	6,314	12,706	31,821	63,657
	2	0,289	0,816	1,886	2,920	4,303	6,965	9,925
	3	0,277	0,765	1,638	2,353	3,182	4,541	5,841
	4	0,271	0,741	1,533	2,132	2,776	3,747	4,604
	5	0,267	0,727	1,476	2,015	2,571	3,365	4,032
	6	0,265	0,718	1,440	1,943	2,447	3,143	3,707
	7	0,263	0,711	1,415	1,895	2,365	2,998	3,499
	8	0,262	0,706	1,397	1,860	2,306	2,896	3,355
	9	0,261	0,703	1,383	1,833	2,262	2,821	3,250
	10	0,260	0,700	1,372	1,812	2,228	2,764	3,169
	11	0,260	0,697	1,363	1,796	2,201	2,718	3,106
	12	0,259	0,695	1,356	1,782	2,179	2,681	3,055
	13	0,259	0,694	1,350	1,771	2,160	2,650	3,012
	14	0,258	0,692	1,345	1,761	2,145	2,624	2,977
	15	0,258	0,691	1,341	1,753	2,131	2,602	2,947
	16	0,258	0,690	1,337	1,746	2,120	2,583	2,921
	17	0,257	0,689	1,333	1,740	2,110	2,567	2,898
	18	0,257	0,688	1,330	1,734	2,101	2,552	2,878
	19	0,257	0,688	1,328	1,729	2,093	2,539	2,861
	20	0,257	0,687	1,325	1,725	2,086	2,528	2,845
	21	0,257	0,686	1,323	1,721	2,080	2,518	2,831
	22	0,256	0,686	1,321	1,717	2,074	2,508	2,819
	23	0,256	0,685	1,319	1,714	2,069	2,500	2,807
	24	0,256	0,685	1,318	1,711	2,064	2,492	2,797
	25	0,256	0,684	1,316	1,708	2,060	2,485	2,787
	26	0,256	0,684	1,315	1,706	2,056	2,479	2,779
	27	0,256	0,684	1,314	1,703	2,052	2,473	2,771
	28	0,256	0,683	1,313	1,701	2,048	2,467	2,763
	29	0,256	0,683	1,311	1,699	2,045	2,462	2,756
	30	0,256	0,683	1,310	1,697	2,042	2,457	2,750
	40	0,255	0,681	1,303	1,684	2,021	2,423	2,704
	50	0,255	0,680	1,296	1,676	2,009	2,403	2,678
	60	0,255	0,679	1,296	1,671	2,000	2,390	2,660
	80	0,254	0,679	1,292	1,664	1,990	2,374	2,639
	100	0,254	0,678	1,290	1,660	1,984	2,365	2,626
	120	0,254	0,677	1,289	1,658	1,980	2,358	2,467
	200	0,254	0,676	1,286	1,653	1,972	2,345	2,601

Приложение 2

Распределение Фишера (F-распределение)

$\alpha = 0{,}10$	число степеней свободы v1																	
число степеней свободы v2	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,50	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,08	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,04	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	2,01	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,98	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,96	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,94	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,92	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,88	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,85	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,84	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,81	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,79	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,73	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,68	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,62	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26

Приложение 2 (продолжение)

Распределение Фишера (F-распределение)

$\alpha = 0{,}05$		число степеней свободы v1																		
		1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120
число степеней свободы v2	1	161	200	216	225	230	234	237	239	271	242	243	244	246	248	249	250	251	252	253
	2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
	3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55
	4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66
	5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,71	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40
	6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70
	7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27
	8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97
	9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75
	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58
	11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45
	12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34
	13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25
	14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	3,52	3,37	3,29	3,21	3,13	3,05	2,96
	16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06
	17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01
	18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97
	19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90
	22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84
	24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,21	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79
	26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75
	28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68
	40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58
	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47
	120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,87	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35

Приложение 3

Распределение Дарбина-Уотсона

$\alpha = 0{,}01$ (n – объем выборки, m – число объясняющих переменных в уравнении регрессии)

n	m=1		m=2		m=3		m=4	
	d1	d2	d1	d2	d1	d2	d1	d2
6	0,390	1,142						
7	0,433	1,036	0,294	1,676				
8	0,497	1,003	0,343	1,489	0,229	2,102		
9	0,554	0,998	0,408	1,389	0,279	1,873	0,183	2,433
10	0,604	1,001	0,466	1,333	0,340	1,733	0,230	2,193
11	0,633	1,010	0,319	1,297	0,396	1,640	0,286	2,030
12	0,697	1,023	0,369	1,274	0,449	1,373	0,339	1,913
13	0,738	1,038	0,616	1,261	0,499	1,326	0,391	1,826
14	0,776	1,034	0,660	1,234	0,347	1,490	0,441	1,737
15	0,811	1,070	0,700	1,232	0,391	1,464	0,488	1,704
16	0,844	1,086	0,737	1,232	0,633	1,446	0,332	1,663
17	0,874	1,102	0,772	1,233	0,672	1,432	0,374	1,630
18	0,902	1,118	0,803	1,239	0,708	1,422	0,613	1,604
19	0,928	1,132	0,833	1,263	0,742	1,413	0,630	1,384
20	0,932	1,147	0,863	1,271	0,773	1,411	0,683	1,367
21	0,973	1,161	0,890	1,277	0,803	1,408	0,718	1,334
22	0,997	1,174	0,914	1,284	0,831	1,407	0,748	1,343
23	1,018	1,187	0,938	1,291	0,838	1,407	0,777	1,334
24	1,037	1,199	0,960	1,298	0,882	1,407	0,803	1,328
25	1,033	1,211	0,981	1,303	0,906	1,409	0,831	1,323
26	1,072	1,222	1,001	1,312	0,928	1,411	0,833	1,318
27	1,089	1,233	1,019	1,319	0,949	1,413	0,878	1,313
28	1,104	1,244	1,037	1,323	0,969	1,413	0,900	1,313
29	1,119	1,234	1,034	1,332	0,988	1,418	0,921	1,312
30	1,133	1,263	1,070	1,339	1,006	1,421	0,941	1,311
35	1,193	1,307	1,140	1,370	1,083	1,439	1,028	1,312
40	1,246	1,344	1,198	1,398	1,148	1,437	1,098	1,318
50	1,324	1,403	1,283	1,446	1,243	1,491	1,203	1,338
100	1,322	1,362	1,303	1,383	1,482	1,604	1,462	1,623

Приложение 4

ПРОГРАММА ЛЕКЦИОННОГО КУРСА

Тема 1. Парная линейная регрессия

Понятие корреляционных и функциональных зависимостей. Парная и множественная регрессия. Причины отклонений в регрессионных моделях. Корреляционное поле. Линейная регрессия. Эмпирическое уравнение регрессии. Метод наименьших квадратов. Определение эмпирических коэффициентов регрессии при помощи МНК. Свойства оценок МНК. Проверка качества уравнения регрессии. Анализ точности определения оценок коэффициентов регрессии. Проверка гипотез относительно коэффициентов линейного уравнения регрессии. Определение интервальных оценок коэффициентов линейного уравнения регрессии. Определение доверительных интервалов для зависимой переменной. Коэффициент детерминации.

Тема 2. Множественная линейная регрессия

Расчет коэффициентов множественной линейной регрессии. Расчет коэффициентов множественной линейной регрессии для уравнения с двумя переменными. Анализ качества эмпирического уравнения множественной линейной регрессии. Определение выборочных дисперсий эмпирических коэффициентов регрессии. Проверка статистической значимости коэффициентов уравнения регрессии. Проверка общего качества уравнения регрессии. Статистика Дарбина-Уотсона.

Тема 3. Нелинейная регрессия

Понятие нелинейной регрессии. Линейные относительно параметров модели. Логарифмическая модель и определение ее коэффициентов. Эластичность зависимой переменной. Полулогарифмические модели: лог-линейная модель и линейно-логарифмическая модель. Обратная модель. Показательная модель. Примеры экономических ситуаций, описываемых с помощью нелинейных регрессионных моделей. Выбор формы модели.

Тема 4. Гетероскедастичность

Гомоскедастичность и гетероскедастичность. Последствия гетероскедастичности. Способы обнаружения гетероскедастичности. Графический анализ остатков для обнаружения гетероскедастичности. Тест ранговой корреляции Спирмена. Тест Голдфелда-Квандта. Методы смягчения проблемы гетероскедастичности. Метод взвешенных наименьших квадратов.

Тема 5. Автокорреляция

Понятие автокорреляции. Причины возникновения автокорреляции. Виды автокорреляции. Способы обнаружения автокорреляции. Графический метод для обнаружения автокорреляции. Критерий Дарбина-Уотсона. Методы устранения автокорреляции. Авторегрессионная схема первого порядка. Поправка Прайса-Винстена. Метод Хилдрета-Лу. Метод первых разностей.

Тема 6. Фиктивные переменные в регрессионных моделях

Понятие фиктивных переменных. Необходимость их использования. Модели дисперсионного анализа. Модели ковариационного анализа. Модели ковариационного анализа при наличии у фиктивной переменной двух альтернатив. Ловушка фиктивной переменной. Модели ковариационного анализа при наличии у качественных переменных более двух альтернатив. Регрессия с одной количественной и двумя качественными переменными. Сравнение двух регрессий. Тест Чоу. Использование фиктивных переменных в сезонном анализе.

Тема 7. Динамические модели

Временной ряд. Лаговые переменные. Виды динамических моделей. Причины наличия лагов. Оценка моделей с лагами в независимых переменных. Модели с конечным и бесконечным числом лагов. Понятие краткосрочного, долгосрочного и промежуточного мультипликаторов. Метод последовательного увеличения количества лагов для оценки моделей с бесконечным числом лагов. Метод Койка для оценки моделей с бесконечным числом лагов. Авторегрессионные модели. Модель адаптивных ожиданий. Модель частичной корректировки. Прогнозирование с помощью временных рядов. Проверка качества прогноза.

Тема 8. Системы одновременных уравнений

Понятие систем одновременных уравнений и необходимость их использования. Модель «спрос-предложение». Кейнсианская модель формирования доходов. Модели IS-LM. Эндогенные и экзогенные переменные. Структурные уравнения модели. Приведенные уравнения. Косвенный метод наименьших квадратов. Проблема идентификации: неидентифицируемость и сверхидентифицируемость. Условия идентифицируемости. Рекурсивные модели. Метод наименьших квадратов для рекурсивных моделей. Двухшаговый метод наименьших квадратов.