Э К О Н О М Е Т Р И К А (Некоторые подходы)

С О Д Е Р Ж А Н И Е

Введение 2
1. Регрессионный анализ. Общие положения 3
2. Метод наименьших квадратов 9
3. Метод наименьших квадратов, наглядный пример 11
4. Оценка значимости коэффициентов и уравнения регрессии 13
5. Обобщенная линейная модель регрессии классического типа 19
6. Мультиколинеарность и ее уменьшение 26
7. Средства против мультиколлинеарности 31
8. Корреляция рядов данных (Автокорреляция) 35
9. Последствия автокорреляции остатков 39
10. Гетероскедастичность 44
11. Средства против гетероскедастичности 50
12. Подбор функциональной зависимости для уравнения регрессии 54
13. Спецификация: выбор независимых переменных 63
14. Экономические данные 71
15. Системы эконометрических уравнений 76
16. Проблема идентификации 81
Б И Б Л И О Г Р А Ф И Я 88
П Р И Л О Ж Е Н И Я 89

Введение

Основу представленного материала составляют лекции, прочитанные в течение нескольких лет для студентов первого и второго циклов обучения различных специальностей (математика и информатика, международные отношения, менеджмент). При составлении курса основное внимание было уделено одному из разделов эконометрики, который чаще других используется в экономических исследованиях при моделировании реальных процессов и явлений, а именно, регрессионному анализу.

Детально рассмотрены этапы, лежащие в основе регрессионного анализа. Они включают: теоретическое обоснование исследуемого процесса или явления; выявление зависимой переменной и не зависимых переменных, в полной мере объясняющих ее поведение; выбор соответствующей функциональной формы; сбор и подготовка правдоподобной информации; проверка выполнения гипотез, лежащих в основе применения метода наименьших квадратов. Если одна из гипотез нарушена, необходимо определить метод оценки коэффициентов исследуемого уравнения.

После того как подобран соответствующий метод решения, запускается уравнение регрессии, используя специализированный Soft: например, пакет Eviewes или, в отсутствие такового, утилита Excell - Data Anayses, которая предоставляет возможность запуска модуля Regression. После запуска уравнения регрессии выполняется анализ полученных результатов для того, чтобы проверить значимость оцененного уравнения и значимость оцененных коэффициентов при независимых переменных. Сравниваются вычисленные статистики Fiser, Durbin-Watson, t-статистики с табличными значениями, которые соответствуют данной степени свободы и выбранному уровню значимости. В том случае, когда подтверждаются гипотезы о принятии решения, при необходимости, переходят к процедуре прогнозирования. С этой целью вычисляются доверительные интервалы для точечного прогноза, и устанавливается прогнозное значение для зависимой переменной в соответствии со значением исследуемой независимой (независимых) переменной.

Рассматриваются процедуры, использование которых позволяет исключить мультиколлинеарность, автокорреляцию остатков, гетероскедастичность, которые могут быть реализованы самостоятельно, без применения специализированного Soft-а.

В заключене рассматриваются системы эконометрических уравнений, которые наиболее полно отражают объективную экономическую реальность. Исследуется проблема идентификации одновременной системы эконометрических уравнений. Особое внимание уделено рассмотрению одновременных систем эконометрических уравнений, содержащих тождества, поскольку они довольно часто встречаются в экономико-математических моделях различного уровня.

Приведены библиографические ссылки, использованные при создании этих лекций, которые послужили источником данных, примеров и иллюстративного материала.

В приложении к этому курсу лекций приведены практические работы по оценке уравнений регрессии для производственной функции и для функции спроса на импортируемые товары и услуги. Также приведены примеры, содержащие мультиколлинеарность, автокорреляцию остатков и гетероскедастичность, которые подлежат решению.

1. Регрессионный анализ. Общие положения

1.1.Эконометрика: определение и использование

Эконометрика может быть определена как количественный анализ реальных экономических процессов. Профессионалы определяют эконометрику в виде некоторого набора методов, которые позволяют *измерять и анализировать экономические феномены и явления, а также прогнозировать экономические тенденции на будущее*. Эконометрика представляет формальное определение, обладающее глубоким смыслом. Дословно, эконометрика означает "*экономические измерения*" и она занимается количественными измерениями и анализом реальной экономики и феноменов, которые относятся к области реального бизнеса. Она представляет попытку измерить реальную экономику и перебросить мостик через пропасть разделяющую экономическую теорию и реальную деловую активность. Эконометрика позволяет нам исследовать данные, характеризующие реальную деятельность фирм и соотнести их с другими факторами, отражающими действия потребителей и действия правительств.

Эконометрика имеет три основных направления использования:

описание реальной экономики;

тестирование экономических гипотез;

прогноз экономической активности на будущее.

Наиболее простым направлением использования эконометрики является *описание.* Эконометрика позволяет оценить экономическую активность; она позволяет внести числовую информацию в уравнения, которые предварительно содержали только абстрактные символы. Например, спрос потребителей на определенное благо может быть представлен как зависимость между количеством запрашиваемого блага (C), ценой данного блага (P), ценой взаимозаменяемых благ (P_s) и располагаемым доходом (Y_d). Для большинства благ взаимосвязь между потреблением и располагаемым доходом предполагается положительной, поскольку возрастание располагаемого дохода ассоциируется с ростом потребления благ**.** *Эконометрика позволяет оценить эту взаимосвязь, используя наблюдаемые значения для потребления, располагаемого дохода и цен, зафиксированные в прошлом*.

Другими словами, функциональная зависимость типа

$$C = f(P, P_s, Y_d) \qquad (1.1)$$

преобразуется в объясняемую зависимость вида

$$C = -60{,}5-0{,}45*P+0{,}12*P_s+12{,}2*Y_d \qquad (1.2)$$

Такое представление определяет намного более специфическую и мотивируемую картину. Сравнивая уравнения (1) и (2), можно сделать следующий вывод: из выражения (1) можно сделать вывод, что потребление является возрастающей функцией от располагаемого дохода, в то время как уравнение (2) позволяет ожидать рост потребляемых благ в количестве 12,2 единиц на каждую дополнительную единицу располагаемого дохода. Цифра 12,2 называется *оцененным* коэффициентом регрессии. И способность эконометрики оценить этот коэффициент является *неоспоримым ее достоинством*.

Второе и, возможно, наиболее часто используемое направление в эконометрике, это *тестирование гипотез*. Например, можно протестировать будет ли исследуемое благо нормальным, т.е. таким, для которого спрос возрастает одновременно с ростом располагаемого дохода. На первый взгляд, кажется, что эта гипотеза может быть поддержана, поскольку знак коэффициента положителен. Однако прежде чем такой вывод будет сделан, необходимо исследовать «статистическую значимость» этой оценки.

Использование эконометрики для тестирования гипотез является, возможно, *наиболее важной ее функцией*.

Третьим, и, наиболее трудным реализуемым, направлением использования эконометрики является *прогнозирование*: что может случиться в следующем квартале, в следующем году или далее в будущем. Например, экономисты используют эконометрические модели для того, чтобы составить прогноз для таких переменных как: объем продаж, объемы доходов, ВВП, нормы инфляции и т.д. Точность этих прогнозов зависит в наибольшей степени от того, в *какой степени прошлое управляет будущим*. Например, предположим, что руководитель компании, который предлагает благо, смоделированное в уравнении (1), решает: повышать ли ему цены или оставить их на том же уровне. Для этого он спрогнозирует объем продаж при повышении цен и без учета повышения цен, что поможет ему в принятии решения относительно повышения или неизменности цен. В таком ключе *эконометрика может быть использована не только для прогноза, но и для анализа экономических политик*.

1.2. Альтернативные эконометрические подходы

Для получения наиболее реалистичной картины относительно возможных подходов, обратим внимание на этапы, которые необходимо выполнить при любом количественном анализе:

а) спецификация моделей или зависимостей, которые подлежат изучению;
б) сбор данных, необходимых для оценки модели;
в) оценка модели с помощью данных.

Этапы а) и б) одинаковы в работах по количественной оценке, однако, на шаге в) оценки моделей разняться от одной дисциплины к другой. Выбор методики для оценки модели на основе некоторого специфического набора данных, как правило, относится к области эконометрического «искусства». Существуют различные альтернативные подходы при оценке одного и того же уравнения, и каждый из подходов может предоставить результаты, которые отличаются друг от друга.

В дальнейшем обратимся к подходу, который соотносится с регрессионным анализом. Однако, при обращении к эконометрике, надо иметь в виду один очень важный факт, регрессия является только одним из возможных эконометрических инструментов для получения оценок.

1.3. Что такое регрессионный анализ?

Регрессионный анализ используется для выполнения количественных оценок экономических взаимосвязей, которые до того имели место только в сугубо теоретическом смысле. Для того чтобы уточнить направления изменений, необходимо ознакомиться с экономической теорией и основными характеристиками исследуемого блага (например, зависимость объема продаж гибких дисков в зависимости от их цены). Для уточнения изменений, которые должны иметь место на количественном уровне, необходим набор соответствующих данных и метод для оценки функциональной зависимости. Наиболее часто используемым методом для оценки этих функциональных зависимостей является регрессионный анализ.

1.4. Зависимые и независимые переменные, обоснование

Регрессионный анализ является статистическим методом, который пытается «объяснить» изменения одной переменной, *зависимой (объясняемой) переменой* как функции от одной или нескольких переменных, так называемых, *независимых (объясняющих переменных)* путем *оценки одного единственного уравнения* (1) $C = f(P, P_s, Y_d)$. Здесь C является зависимой (объясняемой), а P, P_s, Y_d – независимые переменные (объясняющие). Регрессионный анализ является *подходящим инструментом для экономистов*, поскольку

большинство экономических утверждений могут быть сформулированы в виде функциональной зависимости, *состоящей из одного единственного уравнения*.

В экономике и бизнесе большинство афирмаций являются утверждениями типа причина – эффект: если цена возрастает на одну единицу, то объем спроса убывает в среднем на несколько единиц в зависимости от эластичности спроса по цене. По аналогии, если объем используемого капитала возрастает на единицу, тогда объем производства возрастет на несколько единиц, т.е. имеет место, так называемая предельная производительность капитала. Утверждения такого типа формируют отношения типа, если – тогда или причинные отношения, которые постулируют логически, что изменения в зависимых переменных обусловлены изменениями в некотором специфицированном наборе независимых переменных.

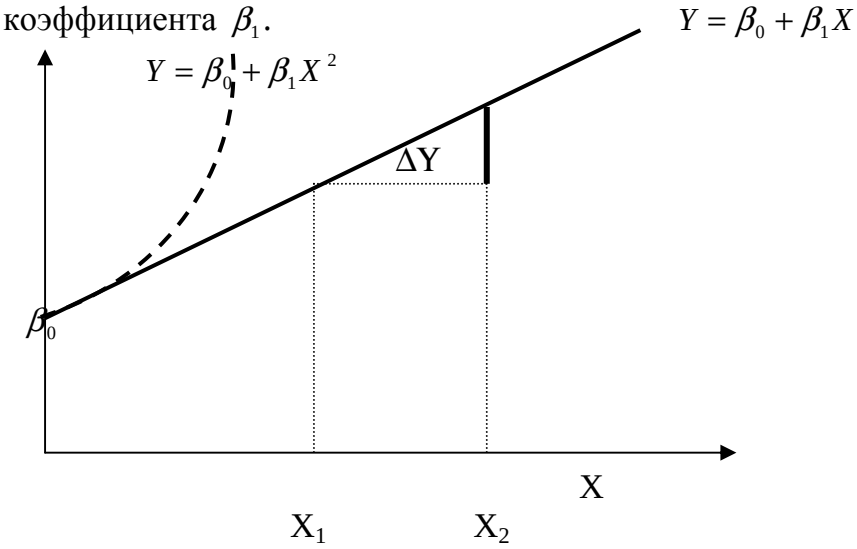
Поскольку многие экономические взаимосвязи являются по своей природе причинными, результат регрессии полагается на его значимость, однако, он не может доказать причинность. Регрессионный анализ может выполнить тестирование значимости количественных соотношений. Обоснование же причинности должно опираться, в большей степени, на экономическую теорию и здравый смысл.

1.5. Линейная модель регрессии из одного уравнения

$$Y = \beta_0 + \beta_1 X \qquad (1.3)$$

Уравнение (3) является самым простым уравнением регрессии, состоящим из одного уравнения. Из уравнения (3) следует утверждение, что зависимая (эндогенная) переменная Y является функцией единственной независимой (экзогенной) переменной X. Модель представлена одним уравнением, поскольку не существует других уравнений для переменной Y, зависящих от X или от других переменных. Модель линейная, поскольку в графической форме она представляет прямую линию, но не кривую.

β_0, β_1 коэффициенты или параметры уравнения, которые определяют координаты прямой линии в любой точке пространства (У,X). β_0 - это константа или свободный член, определяющий точку пересечения прямой с осью координат; этот коэффициент определяет значение независимой переменной Y при X, равном нулю. β_1 - угловой коэффициент, и он определяет значение, на которое изменится переменная Y, когда переменная X измениться на одну единицу. Угловой коэффициент β_1 представляет реакцию переменной Y на изменения переменной X. Для того чтобы объяснить и предсказать изменения зависимой переменной, что является основополагающей задачей при оценке поведенческих уравнений, основное внимание уделяется оценке углового коэффициента β_1.



К примеру, из приведенного рисунка следует, что при увеличении значения переменной X, на $\Delta X = X_2 - X_1$, значение переменной Y, в соответствии с уравнением (3), возрастет на $\Delta Y = Y_2 - Y_1$. Следовательно, в линейных моделях регрессии ответная реакция зависимой переменной Y на изменения в независимой переменной X определяется константой, которая равна угловому коэффициенту $\beta_1 = (Y_2 - Y_1)/(X_2 - X_1) = \frac{\Delta Y}{\Delta X}$.

Следует отличать уравнение, линейное относительно переменных и уравнение, линейное относительно коэффициентов. Линейная регрессия, с необходимостью, должна быть линейной относительно коэффициентов (параметров), но она не обязательно должна быть линейной относительно переменных. Уравнение $Y = \beta_0 + \beta_1 X$ является линейным относительно переменных, в то время как уравнение $Y = \beta_0 + \beta_1 X^2$ не является линейным относительно независимой переменной X, поскольку графически оно представляет квадратичную кривую, но не прямую.

Уравнение является линейным относительно коэффициентов только в том случае, когда они возвышаются в степень (не большую, чем единица), не умножается и не делится на другие коэффициенты, и не являются частью некоторых функций (будь то Log или exp). Уравнение (3) линейно относительно коэффициентов, в то время как уравнение $Y = \beta_0 + X^{\beta_1}$ не является линейным относительно коэффициентов β_0, β_1, поскольку не существует преобразования, которое трансформировало бы его к линейному виду. В общем случае, из всех возможных уравнений с одной независимой переменной, только функция общего вида $f(Y) = \beta_0 + \beta_1 f(X)$ является линейной относительно коэффициентов β_0, β_1.

Все вышеизложенное является важным, поскольку при применении метода линейной регрессии, уравнение, с необходимостью, должно быть линейным. Регрессионный анализ может быть применен к уравнениям, которые не линейны относительно независимой переменной, однако, могут быть представлены в таком виде, что становятся линейными относительно коэффициентов.

1.6. Термин стохастической ошибки. Ошибка спецификации уравнения регрессии

Изменения в зависимой переменной Y могут быть обусловлены не только изменениями в не зависимой переменной X, но и изменениями, которые происходят из других источников. Эти дополнительные изменения появляются, отчасти, вследствие того, что опущены из рассмотрения некоторые важные не зависимые переменные (X_1, X_2, X_3,…), и даже, если эти переменные будут включены в модель, Y продолжает оставаться под воздействием изменений, которые просто не могут быть объяснены посредством модели. Эти изменения могут происходить вследствие: опущенных влияний, ошибок измерения, неправильного выбора вида уравнения или попросту вследствие некоторых событий, которые случайны и полностью непредсказуемы.

В эконометрике такие существенные необъяснимые вариации («ошибки») могут быть учтены путем явного включения в модель регрессии стохастического члена. Стохастическая ошибка является членом, который вводится в уравнение регрессии, для того чтобы отразить все изменения переменной Y, которые не могут быть объяснены с помощью переменной X. Как правило, этот член обозначается символом ε. Добавление члена стохастической ошибки в уравнение регрессии (3) приводит к уравнению регрессии вида:

$$Y = \beta_0 + \beta_1 X + \varepsilon. \qquad (1.4)$$

Уравнение (4) представлено двумя компонентами: детерминированной составляющей и стохастической составляющей. Выражение $Y = \beta_0 + \beta_1 X$ называется детерминированной

составляющей, поскольку значение Y определяется заданными значениями переменной X, которая предполагается полностью определенной, т.е. без случайных воздействий. Эта детерминистская составляющая может быть интерпретирована как ожидаемое значение Y, соответствующее данному значению X либо как среднестатистическое значение Y, ассоциированное с заданным значением X. Детерминистская составляющая может быть обозначена как

$$E(Y/X) = \beta_0 + \beta_1 X \ . \qquad (1.5)$$

К сожалению, в действительности вероятность того, что наблюдаемое значение Y будет равно ожидаемому детерминистскому значению $E(Y/X)$, мала. И, следовательно, в уравнение регрессии необходимо ввести стохастический член

$$Y = E(Y/X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon \ . \qquad (1.6)$$

По крайней мере, четыре источника, отличные от тех, которые обусловлены изменениями переменной X, воздействуют на изменения, происходящие в переменной Y, способствуя увеличению ошибки спецификации, а именно:

– в уравнении опущены из рассмотрения многие незначительные влияния (например, по причине непригодности данных);
– практически невозможно избежать ошибок измерений хотя бы в одной из переменных, входящих в уравнение;
– специфицированное теоретическое уравнение должно обладать функциональной формой отличной от той, которая рассматривается в уравнении регрессии; к примеру, специфицированное уравнение должно быть нелинейным относительно переменных при линейной регрессии или наоборот;
– все попытки формализовать поведение человека, с необходимостью содержат, по крайней мере, небольшое количество непредвиденных или попросту случайных изменений.

1.7. Расширение области применения обозначений

Расширим область применения обозначений с тем, чтобы включить ссылку на некоторое множество установленных наблюдений, и чтобы иметь возможность ввести новые независимые переменные в уравнение регрессии. Тогда единственное линейное уравнение регрессии может быть переписано в следующем виде $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (i=1,2,…,n), в котором

Y_i - наблюдение i зависимой переменной;

X_i - наблюдение i независимой переменной;

ε_i - наблюдение i переменной стохастической ошибки;

β_0, β_1 параметры регрессии;

n - количество наблюдений.

$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$,

$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$,

$Y_3 = \beta_0 + \beta_1 X_3 + \varepsilon_3$,

$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$.

1.8. Оцененное уравнение регрессии

Как только выбрано уравнение регрессии, его необходимо оценить, т.е. необходимо определить параметры специфицированного уравнения. Эта версия «истинного» уравнения регрессии называется оцененным уравнением регрессии, и в основе его

получения лежат наблюдаемые значения Y_s, X_s. В то время как «истинное» уравнение является сугубо теоретическим по своей природе

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{(i=1,2,…,n)}, \tag{1.7}$$

оцененное уравнение регрессии содержит конкретные данные $\widehat{Y}_i = 103{,}4 + 6{,}38\, X_i$. Наблюденные значения для переменных X и Y используются для определения оцененных параметров 103,4 и 6,38, которые, в свою очередь, используются для вычисления оцененных значений $\widehat{Y}_i$ для теоретических значений Y_i.

Рассмотрим отличия между «истинным» уравнением регрессии и оцененным уравнением регрессии. Во-первых, вместо теоретических коэффициентов β_0, β_1 в уравнении регрессии (5), в уравнении (6) появляются оцененные коэффициенты вида 103,4 и 6,38. Поскольку, невозможно выявить значения параметров «истинного» уравнения регрессии, вместо них рассчитываются оценки этих коэффициентов, используя данные, наблюденные в прошлом. Оцененные параметры уравнения регрессии, помеченные как $\widehat{\beta}_0$, $\widehat{\beta}_1$, представляют удачную эмпирическую аппроксимацию, полученную по наблюденным данным Y_s, X_s. В выражении

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \tag{1.8}$$

значения оцененных параметров меняются в зависимости от того, как меняются наблюдения; для каждого нового множества наблюдений вычисляется новая регрессия. Значения $\widehat{Y}_i$ представляют оцененные значения Y_i, и являются результатом применения оцененного уравнения регрессии к i - ому наблюдению.

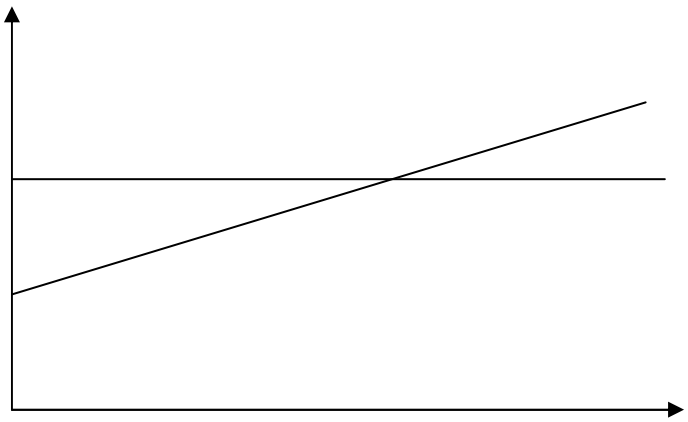
Остаточная переменная определяется как разность между оцененным значением зависимой переменной $\widehat{Y}_i$ и его теоретическим значением (Y_i)

$$Y_i - \widehat{Y}_i = u_i. \tag{1.9}$$

Рассмотрим отличие между остаточной переменной u_i и стохастической ошибкой

$$\varepsilon_i = Y_i - E(Y_i / X_i). \tag{1.10}$$

Остаточная переменная есть разность между наблюдаемым значением Y_i и значением $\widehat{Y}_i$, вычисленным с помощью оцененного уравнения регрессии. В то время, как стохастическая ошибка ε_i является разностью между наблюдаемым значением Y_i и теоретическим, ожидаемым значением зависимой переменной Y. Другими словами, стохастическая ошибка является теоретическим значением, которое невозможно наблюдать, в то время как остаточная переменная является реальным значением, которое вычисляется для каждого наблюдения каждый раз, когда запускается уравнение регрессии. В действительности, большинство регрессионных методов не только выделяет остатки, но и выбирает такие $\widehat{\beta}_0$, $\widehat{\beta}_1$, которые сохраняют остатки на минимально возможном уровне. Чем меньше будут значения остаточных переменных, тем меньше будут отличаться значения $\widehat{Y}_s$ от Y_s. Если выражение (6) подставить в (7), получим другое представление оцененного уравнения регрессии $Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + u_i$.



2. Метод наименьших квадратов

2.1 Оценка линейной модели регрессии с помощью метода наименьших квадратов

Гипотезы:

- Модель является линейной по коэффициентам β_i, (i=0,1) и по стохастической ошибке ε_i, (i=1,…, n).
- Рассматриваются только такие значения X_i, которые не содержат постоянных ошибок, связанных с наблюдениями или измерениями.
- Стохастическая ошибка нормально распределена с нулевым средним $E(\varepsilon_i) = 0$, (i=1,…,n) (в среднем модель хорошо специфицирована).
- Отклонения являются гомоскедастичными: $E(\varepsilon_i^2) = \sigma_i^2$, вариация ошибок σ_i^2 является постоянной. Другими словами, член стохастической ошибки не зависит от эволюции независимой переменной, а это означает, что остаточные дисперсии, вычисленные для различных сегментов значений X_i, одинаковы.
- $E(\varepsilon_i, \varepsilon_j) = 0, \quad i, j = 1,2....,n, i \neq j.$ Значения стохастической ошибки не являются автокорелированными (не зависимы между собой). Последовательные значения стохастической ошибки не зависят друг от друга.
- $E(x_i, \varepsilon_i) = 0$. Значения стохастической ошибки не корелированы со значениями независимой переменной, не зависят от значений независимой переменной.

Итак, в результате статистических наблюдений, получаются ряды данных. Проблема заключается в определении параметров уравнения регрессии по этим данным Метод наименьших квадратов, который впоследствии будем обозначать (МНК), при выполнении объявленных гипотез, обеспечивает возможность определения по рядам данных оценок, которые являются *<u>максимально</u>* *<u>правдоподобными</u>*, *<u>несмещенными</u>*, *<u>согласованными</u>* *и* *<u>эффективными</u>* (с минимальной дисперсией). Эти свойства являются крайне необходимыми для того, чтобы с вычисленными оценками можно было бы согласиться в процессе принятия решений или при эконометрическом моделировании. Обсудим более подробно эти свойства.

<u>Оценки являются несмещенными.</u> Это означает, что $E\left(\widehat{\beta}_k\right) = \beta_k, (k = 1,2,...,n)$, следовательно, оценки коэффициентов, полученные с помощью МНК, центрированы вокруг множества значений истинных коэффициентов.

<u>Оценки являются эффективными</u>. Распределение оцененных коэффициентов вокруг истинных значений параметров является наиболее компактным распределением, которое возможно при несмещенных распределениях. Никакой другой линейный метод оценки коэффициентов не обеспечивает меньшую дисперсию для каждого из оцененных коэффициентов, чем МНК. Т.е. оценки характеризуются как BLUE - Best Linear Unbiased Estimators (Теорема Гаусса-Маркова.).

<u>Оценки являются согласованными.</u> Если бесконечно увеличивать количество наблюдений, то полученные оценки будут стремиться к значениям истинных коэффициентов уравнения регрессии. С ростом числа наблюдений, дисперсия становится меньше, и каждая оценка приближается к истинному значению параметра.

<u>Оценки являются максимально правдоподобными.</u>

$\widehat{\beta}_s$ нормально распределены, $N\left(\beta, VAR[\widehat{\beta}]\right)$

Метод наименьших квадратов, который заключается в определении таких значений параметров β_i, которые обеспечивают минимальное значение суммы квадратов остаточных переменных по всем наблюдениям:

$$\min_{\beta_0,\beta_1}\sum_{i=1}^{n}u_i^2=\min_{\beta_0,\beta_1}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2.$$

Запишем необходимые условия первого порядка для существования экстремума

$$\partial\sum u_i^2/\partial\beta_0=-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)=0.$$

$$\partial\sum u_i^2/\partial\beta_1=-2\sum_{i=1}^{n}\left(-\beta_1x_i^2+x_iy_i-\beta_0x_i\right)=0.$$

$$\sum_{i=1}^{n}y_i=n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}x_i \qquad (2.1)$$

$$\sum_{i=1}^{n}x_iy_i=\hat{\beta}_0\sum_{i=1}^{n}x_i+\hat{\beta}_1\sum_{i=1}^{n}x_i^2 \qquad (2.2)$$

Система уравнений (2.1)-(2.2) называется системой нормальных уравнений. Разделив (2.1) на n и, разрешив это уравнение относительно $\hat{\beta}_0$, получим

$$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}\ , \qquad (2.3)$$

подставив это уравнения в (2.2), определим $\hat{\beta}_1$, $\sum_{i=1}^{n}x_iy_i=\sum_{i=1}^{n}x_i\left(\overline{Y}-\hat{\beta}_1\overline{X}\right)+\hat{\beta}_1\sum_{i=1}^{n}x_i^2$,

$\sum_{i=1}^{n}x_iy_i-\sum^{n}x_i\overline{Y}=\hat{\beta}_1\left(\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\ \overline{X}\right)$, из этого уравнения получим $\hat{\beta}_1=\dfrac{\left(\sum_{i=1}^{n}x_iy_i-\sum^{n}x_i\overline{Y}\right)}{\left(\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\ \overline{X}\right)}$,

после деления числителя и знаменателя на n, получим

$$\hat{\beta}_1=\frac{\overline{XY}-\overline{X}\overline{Y}}{\overline{X^2}-\overline{X}^2}=\frac{Cov(X,Y)}{\sigma_X^2}\ ,\ \hat{\beta}_0=\overline{Y}-\overline{X}\frac{Cov(X,Y)}{\sigma_X^2} \qquad (2.4)$$

Достаточным условием существования экстремума является условие

$$\frac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_i^2}\succ 0, i=0,1;\ \begin{vmatrix}\dfrac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_0^2} & \dfrac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_0\partial\beta_1}\\ \dfrac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_1\partial\beta_0} & \dfrac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_1^2}\end{vmatrix}\succ 0 \qquad (2.5)$$

$$\frac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_1\partial\beta_0}=\sum_{i=1}^{n}x_i \qquad \frac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_0^2}=n \qquad \frac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_1^2}=\sum_{i=1}^{n}x_i^2 \qquad \frac{\partial^2\sum_{i=1}^{n}u_i^2}{\partial\beta_0\partial\beta_1}=\sum_{i=1}^{n}x_i$$

3. Метод наименьших квадратов, наглядный пример

$$\widehat{\beta}_0 = \overline{Y} - \beta_1 \overline{X}; \quad \widehat{\beta}_1 = \frac{Cov(X,Y)}{\sigma_X^2}; \quad Cov(X,Y) = \overline{YX} - \overline{X}\overline{Y}; \quad \sigma_X^2 = \overline{X^2} - \overline{X}^2; \quad \widehat{\beta}_1 = \frac{\overline{YX} - \overline{Y}\overline{X}}{\overline{X^2} - \overline{X}^2}.$$

Параметр β_1 называется коэффициентом регрессии, его значение равно среднему значению зависимой переменной, при условии, что независимая переменная меняется на одну единицу.

Параметр β_0 равен величине, которую принимает зависимая переменная Y, когда значение независимой переменной равно нулю $X = 0$. В том случае, когда независимая переменная не может принимать нулевое значение, предыдущее объяснение лишено смысла. Параметр β_0 может не иметь никакого экономического смысла. Попытка объяснить экономический смысл параметра β_0, особенно в том случае, когда его значение меньше нуля, может привести к абсурдным ситуациям.

В действительности, можно интерпретировать только знак параметра β_0. Если $\beta_0 \prec 0$, тогда относительное изменение зависимой (результирующей) переменной происходит более низкими темпами, чем относительное изменение независимой (факторной) переменной. Другими словами, вариация результирующей (зависимой) переменной, меньше чем вариация независимой (факторной) переменной – коэффициент вариации относительно независимой переменной X, больше чем коэффициент вариации относительно зависимой переменной Y: $V_X \succ V_Y$. Докажем этот феномен, начав сравнивать относительные изменения зависимой переменной Y и независимой переменной X.

$$\frac{dY}{Y} \prec \frac{dX}{X} \Rightarrow \frac{dY}{dX} \prec \frac{Y}{X}; \quad \frac{\beta_1 dX}{dX} \prec \frac{\beta_0 + \beta_1 X}{X}; \quad \beta_1 X \prec \beta_0 + \beta_1 X \Rightarrow 0 \prec \beta_0.$$

Рассмотрим следующий феномен. Группа предприятий, производящая один и тот же продукт, подвержена анализу на предмет производственных затрат в соответствии с функцией регрессии вида: $Y = \beta_0 + \beta_1 X + \varepsilon$. Информацию, необходимую для оценки параметров β_0, β_1 представим в виде таблицы. Отметим важный факт, получивший подтверждение в результате эмпирических исследований, который необходимо учитывать при проведении регрессионного анализа. Количество наблюдений «n» должно быть в 6-7 раз больше, чем количество оцениваемых параметров, соответствующих независимым переменным X.

№ предпр.	Объем пр-ва (тыс.ед.) X	Произв. расходы (млн. лей) Y	$Y \cdot X$	X^2	Y^2	$\widehat{Y}_X$
1	1	30	30	1	900	31,6
2	2	70	140	4	4900	67,9
3	4	150	600	16	22500	141,6
4	3	100	300	9	10000	104,7
5	5	170	850	25	28900	178,4
6	3	100	300	9	10000	104,7
7	4	150	600	16	22500	141,6
Итого	22,00	770	2820	80	99700	770

Запишем систему нормальных уравнений:

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{7} X_i = \sum_{i=1}^{7} Y_i \\ \beta_0 \sum_{i=1}^{7} X_i + \beta_1 \sum_{i=1}^{7} X_i^2 = \sum_{i=1}^{7} X_i Y_i \end{cases}, \qquad \begin{matrix} 7\beta_0 + 22\beta_1 = 770 \\ 22\beta_0 + 80\beta_1 = 2820 \end{matrix},$$

$$\hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X}; \quad \hat{\beta}_1 = \frac{\overline{YX} - \overline{YX}}{\overline{X^2} - \overline{X}^2}; \quad \beta_0 = -5{,}79; \beta_1 = 36{,}84; \hat{Y} = -5{,}79 + 36{,}84X.$$

Значения зависимой переменной представлены в последней колонке. Значение параметра β_0 лишено экономического смысла. В рассматриваемом примере имеем:

$$\overline{X} = 3{,}14; \sigma_X = 1{,}25; V_X = 39{,}8\%; \overline{Y} = 110; \sigma_Y = 46{,}29; V_Y = 42{,}1\%. V_X = \frac{\sigma_X}{\overline{X}}; V_Y = \frac{\sigma_Y}{\overline{Y}}.$$

То, что $\beta_0 \succ 0$, соответствует изменению зависимой переменной более высокими темпами, чем изменение независимой переменной. $\beta_0 \prec 0$ отражает тот факт, что $V_X \succ V_Y$.

Если переменные X и Y выразить через отклонения от средних уровней, то линия регрессии на графике пройдет через начало координат. $Y' = Y - \overline{Y}; X' = X - \overline{X}; \hat{Y}' = \beta_1 X'$. Оценка коэффициента регрессии при этом не изменится.

Оценку коэффициентов регрессии можно получить проще, не обращаясь к методу наименьших квадратов. Альтернативную оценку β_1 можно найти исходя из содержания данного коэффициента: изменение зависимой переменной $\Delta Y = Y_n - Y_1$ сопоставляют с изменением независимой переменной $\Delta X_n = X_n - X_1$.

В нашем примере такого рода альтернативная оценка параметра β_1 составит $\beta_1' = \frac{170 - 30}{5 - 1} = 35$. Эта величина является приближенной, поскольку большая часть информации, имеющейся в данных, не используется при ее расчете. Она основана только на минимаксных значениях переменных. Как правило, уравнению регрессии сопутствует коэффициент линейной корреляции, который характеризует тесноту линейной взаимосвязи между зависимой переменной Y и независимой переменной $X \Rightarrow r_{XY}$. Существует несколько модификаций формулы для коэффициента линейной корреляции. Приведем одну из них $r_{XY} = \beta_1 \frac{\sigma_X}{\sigma_Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$, $\beta_1 = \frac{\mathrm{cov}(X,Y)}{\sigma_X^2}$; известно, что коэффициент линейной корреляции определен на отрезке $-1 \le r_{XY} \le 1$. Если коэффициент $\beta_1 < 0$, $-1 \le r_{XY} \le 0$. Для данных таблицы $r_{XY} = 0{,}991$, это значение очень близко к единице, из чего следует, что между зависимой переменной Y и независимой переменной X существует очень тесная линейная зависимость. Т.е. между затратами на производство и объемом выпущенной продукции существует очень тесная линейная зависимость. Необходимо отметить, что коэффициент линейной корреляции оценивает степень тесноты связи между показателями, взаимосвязь, между которыми рассматривается в линейной форме. Близость значения коэффициента корреляции к нулю, не означает, что между этими показателями не существует другой функциональной зависимости, отличной от линейной, при которой зависимость переменных будет достаточно сильной.

Для оценки качества выбранной линейной функции, вычисляется коэффициент детерминации, который отражает долю дисперсии зависимой переменной Y, объясняемую уравнением регрессии, в общей дисперсии зависимой переменной:

$R_{YX}^2 = \hat{\sigma}_{\hat{Y}}^2 / \hat{\sigma}_Y^2$. Соответственно, величина $1 - R_{YX}^2$ отражает долю дисперсии Y, объясняемую оставшимися факторами, которые не исследуются в модели.

В рассматриваемом примере $R_{YX}^2 = 0{,}982$, и, следовательно, уравнение регрессии объясняет 98,2% от дисперсии зависимой переменной, а на долю оставшихся факторов приходится только 1,8% дисперсии. Величина коэффициента детерминации служит одним из критериев, который *оценивает качество линейной модели*. Чем больше доля отклонений, объясняемых регрессией, тем, соответственно, меньше влияние других факторов, не включенных в уравнение регрессии. Следовательно, линейная модель хорошо аппроксимирует, исходные данные и может быть использована для прогноза значений зависимой (результирующей) переменной.

4. Оценка значимости коэффициентов и уравнения регрессии

После того как линейное уравнение регрессии было специфицировано, производится оценка значимости как уравнения регрессии в целом, так и отдельно каждого параметра уравнения регрессии.

Оценка значимости уравнения регрессии в целом (тестирование правдоподобия модели) производится на основе F-критерия или критерия Фишера, предварительно сформулировав гипотезу о прямой пропорциональной зависимости $H_0 : \{\beta_0 = 0\}$ и гипотезу о независимости переменных $H_0 : \{\beta_1 = 0\}$ в противовес гипотезе специфицированной линейной зависимости - $H_1 : \begin{Bmatrix} \beta_0 \neq 0 \\ \beta_1 \neq 0 \end{Bmatrix}$.

В целях тестирования значимости модели, анализируется разложение суммы квадратов отклонений зависимой переменной Y от среднего значения на две составляющие: первая, которая объясняется уравнением регрессии, и вторая, которая не объясняется уравнением регрессии.

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left((\hat{Y}_i - \bar{Y}) + u_i\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} u_i^2 + 2\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)u_i ,$$

$$\sum_{i=1}^{n} u_i = 0 \Rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i \Rightarrow \bar{Y} = \bar{\hat{Y}} \Rightarrow 2\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)u_i = 2\sum_{i=1}^{n}\left(\bar{\hat{Y}}_i - \bar{Y}\right)u_i = 0 .$$

$$\bar{Y} = \bar{\hat{Y}}$$

Сумма квадратов отклонений значений зависимой переменной от среднего значения $\bar{Y}$ обусловлена как независимой переменной, так и другими факторами. Если независимая переменная не влияет на изменение зависимой переменной, тогда линия, описывающая уравнение регрессии, параллельна оси OX и $\bar{Y} = \hat{Y}$, и вся дисперсия зависимой переменной обусловлена влиянием других факторов. Если другие факторы не воздействуют на зависимую переменную Y, тогда она функционально зависит от независимой переменной X, и сумма квадратов остатков равна нулю. И, как следствие, сумма квадратов отклонений, объясняемых регрессией, совпадает с общей суммой отклонений. Поскольку все точки поля корреляции находятся на линии регрессии, каждый раз имеет место их разложение, обусловленное как независимой переменной X (регрессией), так и другими факторами (необъяснимое регрессией). Очевидно, что линия регрессии хороша для выполнения прогноза тогда, когда сумма квадратов

отклонений, обусловленная регрессией, будет больше, чем сумма квадратов остатков. Уравнение регрессии является значимым и независимая переменная X существенно влияет на зависимую переменную Y, когда коэффициент детерминации R^2_{YX} стремится к единице.

Любая сумма квадратов отклонений напрямую связана со степенью свободы независимого показателя от вариации. *Степени свободы зависят от количества наблюдений «n» и от количества параметров, определяемых в соответствии с ними.* В связи с этим, количество степеней свободы должно отображать, сколько независимых отклонений среди «n» возможных $[(Y_1-\overline{Y}),(Y_2-\overline{Y}),...,(Yn_1-\overline{Y})]$, необходимо для формирования суммы квадратов. Например, для формирования суммы квадратов $\sum_{i=1}(Y_i-\overline{Y})^2$ необходимо $(n-1)$ независимое отклонение, поскольку из множества «n» отклонений после вычисления среднего значения, варьируют независимо только $(n-1)$ отклонение. Предположим, что имеется ряд значений 1,2,3,4,5. Среднее значение равно 3, «n» отклонений от среднего значения равны соответственно: -2; -1; 0; 1; 2. Поскольку в сумме квадратов отклонений от среднего $\sum_{i=1}^{5}(Y_i-\overline{Y})^2$ =0, независимо варьируют только 4 отклонения, то пятое отклонение может быть определено, если известны четыре предыдущих отклонения.

Для вычисления суммы квадратов отклонений, объясняемых регрессией $\sum_{i=1}(\widehat{Y}_i-\overline{Y})^2$, используются, оцененные в соответствии с уравнением регрессии $\widehat{Y}_i=\widehat{\beta}_0+\widehat{\beta}_1X_i$, значения зависимой переменной $\widehat{Y}_i$.

При использовании линейной регрессии $\sum_{i=1}(\widehat{Y}_i-\overline{Y})^2=\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, в этом можно убедиться, если рассмотреть формулу для коэффициента линейной корреляции $r_{XY}=\beta_1\dfrac{\sigma_X}{\sigma_Y}=\dfrac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}\Rightarrow r^2_{XY}=\beta_1^2\dfrac{\sigma_X^2}{\sigma_Y^2}$, откуда следует, что $\sum_{i=1}(\widehat{Y}_i-\overline{Y})^2=\widehat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$.

При заданном количестве наблюдений для X и Y, сумма квадратов отклонений независимой переменной от среднего зависит от единственной константы – коэффициента регрессии β_1, и тогда рассматриваемая сумма имеет единственную степень свободы. К такому же выводу придем, если будем рассматривать уравнение $\widehat{Y}_i=\widehat{\beta}_0+\widehat{\beta}_1X_i$. Параметр β_0 вычисляется по формуле $\beta_0=\overline{Y}-\beta_1\overline{X}$, подставляя это значение в уравнение регрессии, получаем $\widehat{Y}_i=\overline{Y}-\widehat{\beta}_1\overline{X}+\widehat{\beta}_1X_i=\overline{Y}+\widehat{\beta}_1\left(X_i-\overline{X}\right)$.

Из чего следует, что, при известном значении количества наблюдений «n» для переменных X и Y, оцененное значение $\widehat{Y}$ для линейного уравнения регрессии является функцией от одного единственного параметра – коэффициента регрессии. Соответственно и сумма квадратов отклонений независимой переменной имеет только одну степень свободы.

Число степеней свободы, характеризующее общую сумму квадратов отклонений, равно числу степеней свободы суммы квадратов отклонений, которые объясняются уравнением регрессии, и числом степеней свободы относительно суммы квадратов отклонений остатков. Число степеней свободы относительно суммы квадратов остатков для линейной регрессии равно $(n-2)$. Число степеней свободы для общей суммы квадратов отклонений определяется числом наблюдений «n», и, поскольку используется

среднее значение, вычисленное по этим наблюдениям, мы теряем одну степень свободы, т.е. имеем $(n-1)$ степень свободы.

Итак, мы имеем два равенства:

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2$$

$$(n-1) \quad = \quad 1 \quad + \quad (n-2) \quad .$$

Разделив каждую сумму квадратов на соответствующее ей число степеней свободы, получим средний квадрат отклонения или дисперсию, приходящуюся на одну степень свободы.

$$\widehat{\sigma}_{\widehat{Y}}^2=\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2/1\ ; \quad \widehat{\sigma}_u^2=\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2/(n-2); \qquad \widehat{\sigma}_Y^2=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2/(n-1)\ .$$

Определение дисперсий в соответствии со степенями свободы предоставляет возможность их сравнения между собой. Сравнивая дисперсию, объясняемую уравнением регрессии, и остаточную дисперсию, которые приходятся на одну степень свободы, мы получаем F – критерий: $F=\dfrac{\widehat{\sigma}_{\widehat{Y}}^2}{\widehat{\sigma}_u^2}$, этот критерий используется для проверки гипотезы о независимости переменных. $H_0:\widehat{\sigma}_{\widehat{Y}}^2=\widehat{\sigma}_u^2$. Если справедлива гипотеза H_0, тогда дисперсии одинаковы, они не отличаются между собой. Для того, чтобы отклонить гипотезу H_0, необходимо, чтобы $\sigma_{\widehat{Y}}^2 \succ \widehat{\sigma}_u^2$ в несколько раз. Английский ученый ***Snedecor*** разработал таблицы критических значений для F – критерия при различных уровнях значимости гипотезы H_0 и различных степенях свободы. Табличное значение F – критерия является максимальным значением отношения $\sigma_{\widehat{Y}}^2,\widehat{\sigma}_u^2$, которое может иметь место при случайном разбросе и при известной вероятности существования гипотезы H_0. Вычисленное значение этого отношения признано справедливым (отличным от 1), если оно больше табличного значения. В этом случае справедливо заключение о том, что $F_{efect} \succ F_{tab}$ и гипотеза H_0 отвергается.

Если же, $F_{efect} \prec F_{tab}$, то вероятность существования гипотезы H_0 больше чем указанный уровень (например, 0,05) и эта гипотеза не может быть отклонена. В данном случае уравнение регрессии считается статистически незначимым, и гипотеза H_0 не отклоняется. В рассмотренном примере:

$$\sum_{i=1}^{n}(\acute{o}_i-\overline{Y})^2=\sum_{i=1}\acute{o}_i^2-n\overline{Y}^2=15000\,; \qquad \widehat{\sigma}_{\widehat{Y}}^2=\frac{15000}{6}\,;$$

$$\sum_{i=1}^{n}(\widehat{y}_i-\overline{Y})^2=\widehat{\beta}_1^2\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2=14735\,; \qquad \widehat{\sigma}_u^2=\frac{265}{5}=53\,;$$

$$\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2=265\,; \qquad \text{F=1435/53=278;} \qquad F_{\alpha=0,05}=6.61,\ F_{\alpha=0,01}=16.26\,;$$

$$F_{fact}=278 \succ F_{tabl}=6.61\,; \qquad F_{fact}=278 \succ F_{tabl}=16.26\,.$$

Критерий Фишера тесно связан с коэффициентом детерминации R^2, $\sum_{i=1}^{n}(\widehat{y}_i-\overline{Y})^2=R^2\widehat{\sigma}_Y^2$,

$\sum_{i=1}^{n}(\widehat{y}_i-y_i)^2=\left(1-R^2\right)\widehat{\sigma}_Y^2$, и тогда $\qquad F=\dfrac{R^2\cdot(n-2)}{(1-R^2)}$.

Оценка значимости уравнения регрессии, как правило, представляется в форме таблицы для дисперсионного анализа.

Источник вариации	Степени свободы	Сумма квадратов отклонений	Дисперсия на одну степень свободы	F - критерии	
				Факт.	Табл α=0.05
Всего	n-1	15 000	2 500		
Объясненная регрессией	1	14 375	14 375	278	6.61
Остаточная	n-2	235	53		

В линейной регрессии оценивается не только значимость уравнения регрессии в целом, но и значимость коэффициентов в отдельности. В связи с этим определяется стандартная ошибка для каждого параметра: $\hat{\sigma}_{\hat{\beta}_0}, \hat{\sigma}_{\hat{\beta}_1}$. Стандартная оценка коэффициента регрессии определяется по формуле

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2 /(n-2)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}} = \sqrt{\frac{S^2}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}}$$

Значение стандартной ошибки совместно с t – распределением Стьюдента при $(n-2)$ степенях свободы применяется для проверки значимости коэффициента регрессии $\hat{\beta}_1$ и для вычисления доверительного интервала.

Для оценки значимости коэффициента регрессии, значение коэффициента регрессии сравнивается со стандартной ошибкой, другими словами, определяется фактическое значение t – критерия Стьюдента: $t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$, которое сравнивается с табличным значением для уровня значимости нулевой гипотезы (α) и $(n-2)$ степенями свободы.

Тот же результат получается, если извлечь корень квадратный из величины F – критерия. $t_{\hat{\beta}_1} = \sqrt{F}$, докажем истинность равенства $R^2_{\hat{\beta}_1} = F$.

$$(t_{\hat{\beta}_1})^2 = (\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}})^2 = \hat{\beta}_1^2 \Big/ \sum\left(y_i - \hat{Y}_i\right)^2 /(n-2) \Big/ \sum\left(x_i - \overline{X}\right)^2 = \frac{\hat{\beta}_1^2 \sum\left(x_i - \overline{X}\right)^2}{\sum\left(y_i - \hat{y}_i\right)^2 /(n-2)} =$$

$$= \frac{\sum\left(\hat{y}_i - \overline{Y}\right)^2}{\sum\left(y_i - \hat{y}_i\right)^2 /(n-2)} = \frac{\hat{\sigma}^2_{\hat{Y}}}{\hat{\sigma}^2_u} = F, \; t_{\hat{\beta}} = \sqrt{F}.$$

Доверительный интервал для уравнения регрессии определяется как $\hat{\beta}_1 \pm t_{tab} * \hat{\sigma}_{\hat{\beta}_1}$.

Поскольку коэффициент регрессии в эконометрических исследованиях имеет понятное экономическое объяснение, то доверительный интервал для него не должен содержать противоречивых результатов, как, например, $-10 \le \hat{\beta}_1 \le 40$. Это означает, что истинное значение коэффициента регрессии может принимать одновременно как положительные, так и отрицательные значения, и даже нулевое значение, что, в принципе, невозможно.

Стандартная ошибка параметра $\hat{\beta}_0$ определяется из следующей формулы:

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)} \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}} = \sqrt{\frac{S^2 \sum x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{X})^2}}.$$

Оценка значимости выполняется также как и для коэффициента $\hat{\beta}_1$, $t_{\hat{\beta}_1} = \dfrac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}}$, при этом значение вычисленного t – критерия сравнивается с табличным значением при $(n-2)$ степенях свободы. Значимость линейного коэффициента корреляции проверяется с помощью значения коэффициента корреляции $m_R = \sqrt{\dfrac{1-R^2}{n-2}}$.

Фактическое значение t – критерия Стьюдента определяется как: $t_R = \dfrac{R}{\sqrt{1-R^2}}\sqrt{n-2}$. Из этой формулы следует, что для регрессии линейной по переменным, $t_R^2 = F$, поскольку $F = \dfrac{R^2}{(1-R^2)\cdot(n-2)}$, плюс к этому, $t_{\hat{\beta}_1}^2 = F$, а $t_R^2 = t_{\hat{\beta}_1}^2$.

Следовательно, проверка гипотезы значимости коэффициентов регрессии и коэффициентов корреляции эквивалентна проверке гипотезы правдоподобия линейной модели регрессии.

Формулу, предложенную для оценки коэффициентов корреляции желательно применять для большого числа наблюдений и тогда, когда значение r сильно отличается от +1 или –1. В противном случае распределение оценок не является нормальным либо типа Стьюдента, поскольку коэффициент корреляции ограничен значениями –1 и +1. Фишер ввел переменную $z = \dfrac{1}{2}\ln(\left(\dfrac{1+r}{1-r}\right)$ для оценки значимости r.

4.1. Прогнозный интервал для линейной модели регрессии

В прогнозных расчетах, основанных на уравнении регрессии, определяется значение $\hat{y}_p$ в виде точечного прогноза y_i для $x_p = x_k$, подставляя в уравнение регрессии $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ соответствующее значение x_p. Однако, очевидно, что точечный прогноз нереален, поэтому этот прогноз дополняется расчетами стандартных ошибок $\hat{Y}_p, \hat{\sigma}_{\hat{Y}}$ и оцененным прогнозным интервалом для значения Y^*, $\hat{Y}_X - \hat{\sigma}_{\hat{Y}_X} \le Y^* \le \hat{Y}_X + \hat{\sigma}_{\hat{Y}_X}$.

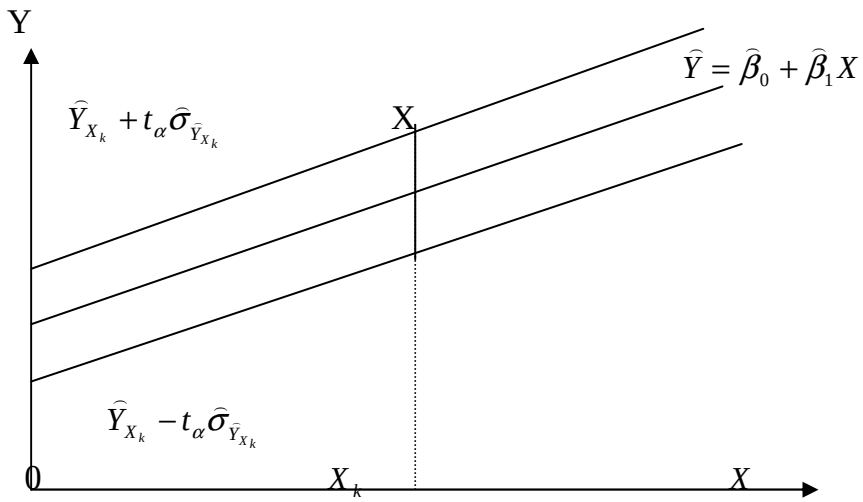
Для нахождения формулы стандартной ошибки $\hat{\sigma}_{\hat{Y}_X}$, обратимся к уравнению регрессии $\hat{Y}_X = \hat{\beta}_0 + \hat{\beta}_1 X$. Заменив $\hat{\beta}_0$ формулой для его вычисления $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 X$, получим $\hat{Y}_X = \overline{Y} - \hat{\beta}_1\overline{X} + \hat{\beta}_1 X = \overline{Y} + \beta_1\left(X - \overline{X}\right)$. Отсюда следует, что стандартная ошибка $\overline{Y} - \hat{\sigma}_{\hat{Y}_X}$ зависит от ошибки $\overline{Y}$ и ошибки коэффициента регрессии $\hat{\beta}_1$, следовательно $\hat{\sigma}_{\hat{Y}_X}^2 = \hat{\sigma}_{\overline{Y}}^2 + \hat{\sigma}_{\hat{\beta}_1}^2\left(X - \overline{X}\right)^2$.

Известно из теории выборок, что $\hat{\sigma}_{\overline{Y}}^2 = \dfrac{\sigma^2}{n}$, используя в качестве σ^2 остаточную дисперсию на одну степень свободы S^2, получаем: $\hat{\sigma}_{\overline{Y}}^2 = \dfrac{S^2}{n}$. Стандартная ошибка коэффициента регрессии определяется формулой $\hat{\sigma}_{\hat{\beta}_1}^2 = \dfrac{S^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$. Будем считать, что прогнозируемое значение $X_p = X_k$, тогда получим следующую формулу для стандартной

ошибки значения $\hat{Y}_{X_k}$,спрогнозированного с помощью уравнения регрессии:

$$\hat{\sigma}^2_{\hat{Y}_{X_k}} = \frac{S^2}{n} + \frac{S^2\left(X_k - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = S^2\left(\frac{1}{n} + \frac{\left(X_k - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)$$

Соответственно, $\hat{\sigma}_{\hat{Y}_{X_k}} = S\sqrt{\frac{1}{n} + \frac{\left(X_k - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$. Полученная формула стандартной ошибки для среднего прогнозируемого значения $\hat{Y}_X$ при известном значении X_k, дает характеристику ошибки расположения линии регрессии. Значение стандартной ошибки $\hat{\sigma}_{\hat{Y}_X}$ достигает минимального значения тогда, когда $X_k = \overline{X}$ и оно возрастает с отдалением точки X_k от среднего значения $\overline{X}$ в любом направлении. Другими словами, насколько больше будет, по абсолютной величине, разность между значениями переменных X_k и $\overline{X}$, настолько больше будет и ошибка $\hat{\sigma}_{\hat{Y}_X}$. С ее помощью будет оценено прогнозное значение $\hat{Y}$, вычисленное по данному значению X_k. Можно ожидать и более удачных прогнозов, если точка X_k будет находиться в центре области наблюдения. И, нет оснований ожидать хороших прогнозов, когда точка X_k отдаляется от точки $\overline{X}$. В том случае, когда точка X_k находиться вне области наблюдения X, использованных для определения линии регрессии, результаты прогноза ухудшаются по мере отдаления точки X_k от области наблюдаемых значений для переменной X.



Для прогнозного значения $\hat{Y}_X$, 95% доверительный интервал при известном значении X_k, определяется с помощью формулы: $\hat{Y}_{X_k} \pm t_\alpha \hat{\sigma}_{\hat{Y}}$. На графике доверительные границы для прогнозного значения $\hat{Y}$ представляют две гиперболы, расположенные по обеим сторонам линии регрессии. Эти две гиперболы, расположенные по обе стороны линии регрессии, определяют 95% доверительный интервал для среднего значения $\hat{Y}$ при заданном значении X.

Наблюденные значения Y_i сосредоточены вокруг среднего значения $\overline{Y}$. Вычисленные же значения $\widehat{Y}$ могут быть рассредоточены вокруг линии $\widehat{Y}$, в пределах значений случайной ошибки u, и при этом, дисперсия остаточной переменной, приходящаяся на одну степень свободы, равна S^2. Поэтому, значение индивидуальной прогнозной ошибки Y, с необходимостью, требует включения не только стандартной ошибки $\widehat{\sigma}_{\widehat{Y}_X}$, но и случайной ошибки S. И тогда, средняя ошибка $\widehat{\sigma}_{\widehat{Y}_{X_k}}$ частного прогнозного значения $\widehat{Y}$, определяется как:

$$\widehat{\sigma}_{\widehat{Y}_{X_k}} = S\sqrt{1+\frac{1}{n}+\frac{\left(X_k-\overline{X}\right)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}}.$$

Выполняя прогноз по линейному уравнению регрессии, необходимо иметь в виду, что прогнозное значение зависит не только от стандартной ошибки частного значения Y, но и от точности прогноза значения независимой переменной X. Ее значение может быть определено, основываясь на анализе других моделей, исходя из конкретной ситуации, анализируя динамику этой переменной.

Формула, рассматриваемая для средней ошибки частного значения зависимой переменной Y ($\widehat{\sigma}_{\widehat{Y}_{X_k}}$), может быть использована для оценки значимости отклонения прогнозного значения, полученного по уравнению регрессии, и гипотетического значения, предложенного исходя из эволюции событий.

$$t_{\widehat{Y}(X_k)} = \frac{\left(\widehat{Y}_{(X_k)} - Y_{hipot}\right)}{\widehat{\sigma}_{\widehat{Y}_{(X_k)}}}; \quad t_{calc\widehat{Y}(X_k)} \succ t_{tab(0,05;n-2)}$$

5. Обобщенная линейная модель регрессии классического типа

5.1. Создание и спецификация обобщенной линейной модели регрессии

Обратимся к случаю, когда, наверняка, более одной независимой переменной, в полной мере, могут объяснить поведение одной зависимой переменной. Случаи, когда поведение одной зависимой переменной может быть объяснено изменениями, происходящими в одной независимой переменной, в повседневной жизни встречаются крайне редко. Спрос на определенное благо, в большей степени может быть объяснен инфляцией, однако это объяснение не является полным, поскольку реклама, агрегированный доход, цены на блага – заменители, международные рынки, качество коммерческих услуг, капризы потребителей, изменение предпочтений (вкусов) – все является важным при моделировании. И, следовательно, чувствуется жизненная необходимость перехода от линейной модели регрессии с одной независимой переменной к линейным моделям регрессии со многими переменными.

Обобщенная модель регрессии с k независимыми переменными может быть представлена следующим уравнением: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$ или

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, \qquad (5.1)$$

где $i = \overline{1,n}$ указывает на количество наблюдений, а x_{i1} указывает на $i-$ое наблюдение переменной X_1, в то время как x_{i2} указывает на $i-$ое наблюдение переменной X_2, k - количество независимых переменных, ε_i - $i-$ая компонента стохастической ошибки.

Самое большое отличие между моделью регрессии с одной независимой переменной и моделью регрессии со многими независимыми переменными состоит в проблеме интерпретации дополнительных угловых коэффициентов. Эти коэффициенты, часто

называются частными коэффициентами регрессии, поскольку коэффициенты множественной регрессии соответствуют частным отклонениям по соответствующей независимой переменной. Коэффициенты должны быть определены для того, чтобы исследователь имел возможность отличить влияние одной независимой переменной на зависимую переменную от влияния другой независимой переменной. А именно, коэффициент множественной регрессии указывает на изменение в зависимой переменной, которое ассоциируется с ростом на одну единицу, происходящим в исследуемой независимой переменной, в то *время как все оставшиеся независимые переменные, входящие в уравнение регрессии, остаются постоянными (не меняются).*

Последняя, подчеркнутая фраза, является ключом к пониманию множественной регрессии. Коэффициент β_1 измеряет воздействие на зависимую переменную Y роста в одну единицу, произошедшего в независимой переменной X_1, при сохранении постоянными значений оставшихся переменных $X_2, X_3, ..., X_k$. Исключение составляют переменные (к примеру, переменная X_{k+1}), опущенные из уравнения регрессии. Коэффициент β_0 является значением зависимой переменной Y, получаемым при равенстве нулю всех независимых переменных и соответствующих им статистических ошибок: $x_{is} = 0, i = 1, n; s = 1, k$ и $\varepsilon_i = 0$, $i = \overline{1, n}$. Как было отмечено ранее, свободный член β_0 будет включен в уравнение регрессии, но на него не будет ссылок при извлечении выводов. Запишем оцененное уравнение регрессии для каждого наблюдения отдельно:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12} + ... + \hat{\beta}_k x_{1k}$$
$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_{21} + \hat{\beta}_2 x_{22} + ... + \hat{\beta}_k x_{2k}$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{k1} + \hat{\beta}_2 x_{k2} + ... + \hat{\beta}_k x_{kk}$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$\hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \hat{\beta}_2 x_{n2} + ... + \hat{\beta}_k x_{nk}$$

Модель множественной линейной регрессии может быть записана в векторной форме $Y = X\beta + \varepsilon$, $\hat{Y} = X\hat{\beta}, Y = \hat{Y} + u$, $Y, \hat{Y}$ - векторы размерности (n), X - матрица размерности - $(n \cdot (k+1))$, β - вектор размерности $(k+1)$, ε, u - векторы размерности (n).

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1\, x_{11} \cdots x_{1k} \\ 1\, x_{21} \cdots x_{2k} \\ 1\, x_{n1} \cdots x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_o \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} i = \overline{1, n}; \quad s = \overline{0, k}, \ \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

$$\beta_1 = \frac{\partial Y}{\partial X_1}; \beta_2 = \frac{\partial Y}{\partial X_2}; \cdots \beta_k = \frac{\partial Y}{\partial X_k}; \quad \hat{Y} = X\hat{\beta} \ ; \ Y - \hat{Y} = u .$$

Примером множественной регрессии может служить функция спроса, которая в большинстве случаев представляется в виде модели вида $C = f(Y, P, M, Z)$, где C - потребление; Y - доход; P - индекс потребительских цен; M - деньги в обращении; Z - ликвидные активы, $\frac{\partial C}{\partial Y} \prec 1$. Множественная регрессия, широко используется в разрешении проблем спроса, в определении дохода по акциям, при изучении динамики производственных затрат, в макроэкономических расчетах. В настоящее время множественная регрессия является одним из наиболее распространенных методов в

эконометрике. *Основная цель множественной регрессии состоит в определении модели множественной регрессии исходя из того, что известно влияние каждой независимой переменной в отдельности и общее воздействие всех независимых переменных на зависимую переменную*.

5.2. Классические гипотезы

Для того чтобы оценки, полученные с помощью метода наименьших квадратов, были лучшими из известных, необходимо выполнение классических гипотез.

Гипотезы:

- Значения переменных $x_{ij}, \left(i=\overline{1,n}; s=\overline{1,k}\right)$ не содержат постоянных ошибок измерений.
- Ожидаемое значение стохастической ошибки ε_i равно нулю, т.е. $E(\varepsilon_i)=0$ (или, другими словами, переменная ε является нормально распределенной переменной с нулевым средним)
- Вариация ошибок постоянна для любого $1 \le i \le n$, $VAR(\varepsilon_i)=\sigma^2_{\varepsilon_i}=\sigma^2=const$.
- Наблюдаемые значения члена стохастической ошибки не коррелированны между собой, (т.е. имеет место независимость стохастических ошибок $E(\varepsilon_i\varepsilon_j)=0; i,j=\overline{1,n}; i \ne j$, нет корреляции ряда значений стохастической ошибки).
- Значения стохастических ошибок не зависят от значений независимых переменных. $E(x_{is}\varepsilon_i)=0, \quad i=\overline{1,n}; s=\overline{1,k}.$
- Отсутствие коллинеарности между независимыми переменными влечет за собой существование регулярной матрицы, которая обеспечивает существование обратной матрицы $\left(X^T X\right)^{-1}$.
- Матрица $\left(X^T X\right)^{-1}/n$ является конечной не сингулярной матрицей.

Член стохастической ошибки, который удовлетворяет, объявленным гипотезам, называется нормальной стохастической ошибкой классического типа.

Из соотношения $n \succ k, n \approx 6(7)\cdot k$ с необходимостью следует, что количество наблюдений должно быть в 6-7 раз больше, чем количество независимых переменных.

5.3. Метод наименьших квадратов

При условии выполнения объявленных классических гипотез, оценка параметров β_s осуществляется, как правило, с помощью метода наименьших квадратов (М.Н.К.). Сущность метода наименьших квадратов состоит в минимизации суммы квадратов отклонений оцененных значений зависимой переменной от теоретических ее значений.

$$\min_{\beta_i}\sum_{i=1}^{n}u_i^2=\min_{\beta_i}\left(u^T u\right)=\min_{\beta_i}\left(Y-X\hat{\beta}\right)^T\left(Y-X\hat{\beta}\right)=S\left(\hat{\beta}\right)\quad=\left(Y^T Y\right)-2\hat{\beta}\left(X^T Y\right)+\hat{\beta}'\left(X^T X\right)\hat{\beta}.$$

Для минимизации этой функции вычисляются производные по β_s и приравниваются нулю:

$$\frac{\partial S\left(\hat{\beta}\right)}{\partial\hat{\beta}}=0;\quad -2\left(X^T Y\right)+2\left(X^T X\right)\cdot\hat{\beta}=0\Rightarrow\left(X^T X\right)\cdot\hat{\beta}=\left(X^T Y\right)\Rightarrow\hat{\beta}=\left(X^T X\right)^{-1}\left(X^T Y\right).$$

5.4. Оценки параметров модели и их свойства

- Оценка $\hat{\beta}$ является линейной несмещенной оценкой $E\left(\hat{\beta}\right)=\beta$.
- Линейная несмещенная оценка $\hat{\beta}$ имеет минимальную дисперсию.

– Оценка $\widehat{\beta}$ является согласованной, т.е. при увеличении количества наблюдений оценка стремится к теоретическому значению $\widehat{\beta} \underset{n\to\infty}{\longrightarrow} \beta$.

– Оценка $\widehat{\beta}$ является нормально распределенной $N\left(\beta, VAR\left(\widehat{\beta}\right)\right)$.

– Оценка $\widehat{\beta}$, полученная с помощью М.Н.К., является оценкой ***BLUE*** (Best Linear Unbiased Estimator), т.е. самой лучшей линейной несмещенной оценкой.

Матрица вариаций и ковариаций коэффициентов регрессии вычисляется как: $\Omega_{\widehat{\beta}} = \sigma_u^2\left(X^T X\right)^{-1}, \sigma_u^2 = \left(u^T u\right), \quad \widehat{\sigma}_u^2 = \dfrac{\left(u^T u\right)}{n-k-1}$. Оценка матрицы вариаций и ковариаций коэффициентов регрессии записывается следующим образом $\widehat{\Omega}_{\widehat{\beta}} = \widehat{\sigma}_u^2\left(X^T X\right)^{-1}$

Эти значения зависят от единицы измерения, поэтому предпочтительно использовать коэффициент детерминации R^2 и коэффициент множественной корреляции r.

$$R^2 = \sum_{i=1}^{n}\left(\widehat{y}_i - \overline{Y}\right)^2 \Big/ \sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 = 1 - \sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2 \Big/ \sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2$$

$r_{YX_1X_2\cdots X_k} = \sqrt{1 - \dfrac{\sigma_u^2}{\sigma_Y^2}}$, где $\sigma_u^2 = \sum\left(y_i - \widehat{y}_i\right)^2$ - остаточная вариация, а $\sigma_Y^2 = \sum\left(y_i - \overline{Y}\right)^2$ общая вариация.

Тест Фишера F используется для проверки значимости уравнения регрессии. С этой целью выдвигаются гипотезы:

$H_0 : \beta_0 = \beta_2 = \cdots = \beta_k = 0$,

$H_1 :$ *не все* $\beta_i \ \left(i = \overline{0,k}\right)$ *равны* 0.

Используемая статистика следующая:

$F_{calc} = \dfrac{\sum_i \left(\widehat{y}_i - \overline{Y}\right)^2 / k}{\sum_i \left(y_i - \widehat{y}_i\right)^2 / \left(n-k-1\right)}$, в случае справедливости гипотезы H_0, данная статистика имеет распределение F (Snedecor) с k и $\left(n-k-1\right)$степенями свободы. Правило для принятия решения при уровне значимости α следующее: принимается гипотеза H_0, если $F_{calc} \prec F_{tab} = F_{\alpha;(k,n-k-1)}$, и гипотеза H_0 отвергается в пользу гипотезы H_1, если $F_{calc} \succ F_{tab} = F_{\alpha;(k,n-k-1)}$, $F_{calc} = \dfrac{R^2/\left(k\right)}{\left(1-R^2\right)/\left(n-k-1\right)}$. С помощью, вычисленного теста F проверяется, насколько хорошо определена модель.

5.5. Построение тестов и интервалы доверия

Тест Стьюдента используется для проверки значимости коэффициентов β_i. Тестируется гипотеза $H_0 : \beta_i = 0, i = \overline{1,k}$ против гипотезы $H_1 : \beta_i \neq 0$, и для принятия гипотезы H_1 при уровне значимости α, необходимо, чтобы выражение $\left|\dfrac{\widehat{\beta}_i}{\widehat{\sigma}_{\widehat{\beta}_i}}\right| \geq 2$, тогда

принимается гипотеза H_1, т.е. делается вывод, что коэффициент при переменной X_i существенно отличается от нуля.

При этом вариации и ковариации оцененных коэффициентов регрессии вычисляются по формуле $\hat{\Omega}_{\hat{\beta}} = \hat{\sigma}_u^2 \left(X^T X\right)^{-1} = \begin{vmatrix} \hat{\sigma}_{\hat{\beta}_1} & \operatorname{cov}(X_1 X_1) & \cdots & \operatorname{cov}(X_1 X_k) \\ \operatorname{cov}(X_2 X_1) & \hat{\sigma}_{\hat{\beta}_2} & \cdots & \operatorname{cov}(X_2 X_k) \\ \dots & \dots & \dots & \dots \\ \operatorname{cov}(X_k X_1) & \operatorname{cov}(X_k X_2) & \cdots & \hat{\sigma}_{\hat{\beta}_k} \end{vmatrix}$.

5.6. Прогноз и его жизненность, прогнозные интервалы доверия

В данном случае проблема заключается в определении значения, которое необходимо присвоить зависимой (эндогенной) переменной, когда известны значения независимых или экзогенных переменных.

Оцененная модель регрессии записывается следующим образом:

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$.

Прогноз для $(i+h)$ значения зависимой переменной определяется как:

$\hat{y}_{i+h} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i+h} + \hat{\beta}_2 x_{2i+h} + \cdots + \hat{\beta}_k x_{ki+h}$, прогнозная ошибка равна $u_{i+h} = y_{i+h} - \hat{y}_{i+h}$.

Исходя из того, что основные гипотезы для обобщенного уравнения регрессии выполнены, полученные прогнозные оценки будут несмещенными.

Вариация прогнозной ошибки равна: $\hat{\sigma}^2_{\hat{y}_{i+h}} = \hat{\sigma}_u^2 \left[X_{i+h}^T \left(X^T X\right)^{-1} X_{i+h} + 1\right]$, где $X_{i+h} = \begin{bmatrix} x_{1i+h} \\ x_{2i+h} \\ \vdots \\ x_{ki+h} \end{bmatrix}$ - это вектор предполагаемых значений независимых переменных.

Прогнозная ошибка нормально распределена, и выражается через теоретическое значение вариации σ_u^2 и через его оцененное значение $\hat{\sigma}_u^2$. Используя предыдущее выражение, получаем: $\dfrac{\hat{y}_{i+h} - y_{i+h}}{\hat{\sigma}_u^2 \left[X_{i+h}^T \left(X^T X\right)^{-1} X_{i+h} + 1\right]} = t_{\hat{y}_{t=h}}$. Эта ошибка следует закону Стьюдента с $(n-k-1)$ степенями свободы.

Прогнозный интервал доверия при уровне значимости α приобретает вид:

$y_{i+h} = \hat{y}_{i+h} \pm t_{\alpha, n-k-1} \sqrt{\hat{\sigma}_u^2 \left[X_{i+h}^T \left(X^T X\right)^{-1} X_{i+h} + 1\right]}$

Условные обозначения

Истинные, не наблюденные переменные		Оценки	
Название	Обозначение	Название	Обозначение
Коэффициент регрессии	β_k	Оцененный коэффициент регрессии	$\hat{\beta}_k$
Ожидаемое значение коэффициента регрессии	$E\left(\hat{\beta}_k\right)$		
Вариация ошибки	σ^2 или $VAR(\varepsilon_i)$	Оцененная вариация ошибки	S^2 или $\hat{\sigma}^2$
Стандартное отклонение ошибки	σ	Оцененное стандартное отклонение	S или $\hat{\sigma}$

Вариация оцененного коэффициента	$\sigma^2(\hat{\beta}_k)$или $VAR(\hat{\beta}_k)$	Стандартная ошибка оцененного коэффициента регрессии	$S^2(\hat{\beta}_k)$или $VAR(\hat{\beta}_k)$
Стандартное отклонение оцененного коэффициентов	$\sigma_{\hat{\beta}_k}$ или $\sigma(\hat{\beta}_k)$	Стандартная ошибка оценки коэффициентов	$\hat{\sigma}_{\hat{\beta}_k}$ или $SE(\hat{\beta}_k)$
Стандартная ошибка	ε_i	Оценка стохастической. ошибки	u_i

5.7. Проверка модели регрессии с помощью тестов

Рассуждая аналогично случаю линейной регрессии с одной независимой переменной, получим, что общая вариация равна вариации, объясняемой с помощью регрессии, плюс вариация остатков:

$$\sum_{i=1}^{n}\left(y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 .$$

Эти величины зависят от единицы измерения, а потому предпочтительнее использование коэффициента детерминации R^2 и коэффициента множественной корреляции r.

Запишем следующую систему нормальных уравнений для множественной регрессии:

$$\sum_{i=1}^{n} y_i = n\cdot\beta_0+\beta_1\sum_{i=1}^{n}x_{i1}+\beta_2\sum_{i=1}^{n}x_{i2}+\cdots+\beta_k\sum_{i=1}^{n}x_{ik}$$

$$\sum_{i=1}^{n} x_{i1}y_i = \beta_0\sum_{i=1}^{n}x_{i1}+\beta_1\sum_{i=1}^{n}x_{i1}x_{i1}+\beta_2\sum_{i=1}^{n}x_{i1}x_{i2}+\cdots+\beta_k\sum_{i=1}^{n}x_{i1}x_{ik}$$

$$\sum_{i=1}^{n} x_{i2}y_i = \beta_0\sum_{i=1}^{n}x_{i2}+\beta_1\sum_{i=1}^{n}x_{i2}x_{i1}+\beta_2\sum_{i=1}^{n}x_{i2}x_{i2}+\cdots+\beta_k\sum_{i=1}^{n}x_{i2}x_{ik}$$

$$\cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots$$

$$\sum_{i=1}^{n} x_{ik}y_i = \beta_0\sum_{i=1}^{n}x_{ik}+\beta_1\sum_{i=1}^{n}x_{ik}x_{i1}+\beta_2\sum_{i=1}^{n}x_{ik}x_{i2}+\cdots+\beta_k\sum_{i=1}^{n}x_{ik}x_{ik}$$

$$\Delta = \begin{vmatrix} n & \sum_{i=1}^{n}x_{i1} & \sum_{i=1}^{n}x_{i2} & \cdots\sum_{i=1}^{n}x_{ik} \\ \sum_{i=1}^{n}x_{i1} & \sum_{i=1}^{n}x_{i1}x_{i1} & \sum_{i=1}^{n}x_{i1}x_{i2} & \cdots\sum_{i=1}^{n}x_{i1}x_{ik} \\ \sum_{i=1}^{n}x_{i2} & \sum_{i=1}^{n}x_{i2}x_{i1} & \sum_{i=1}^{n}x_{i2}x_{i2} & \cdots\sum_{i=1}^{n}x_{i2}x_{ik} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n}x_{ik} & \sum_{i=1}^{n}x_{ik}x_{i1} & \sum_{i=1}^{n}x_{ik}X_{i2}x & \sum_{i=1}^{n}x_{ik}x_{ik} \end{vmatrix}$$

$$\Delta\beta_0 = \begin{vmatrix} \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} y_i x_{i1} & \sum_{i=1}^{n} x_{i1} x_{i1} & \sum_{i=1}^{n} x_{i1} x_{i2} & \cdots \sum_{i=1}^{n} x_{i1} x_{ik} \\ \sum_{i=1}^{n} y_i x_{i2} & \sum_{i=1}^{n} x_{i2} x_{i1} & \sum_{i=1}^{n} x_{i2} x_{i2} & \cdots \sum_{i=1}^{n} x_{i2} x_{ik} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} y_i x_{ik} & \sum_{i=1}^{n} x_{ik} x_{i1} & \sum_{i=1}^{n} x_{ik} x_{i2} & \cdots \sum_{i=1}^{n} x_{ik} x_{ik} \end{vmatrix}$$

Для того, чтобы применить тест Фишера F составим следующую таблицу для анализа вариаций.

Источники вариации	Вариация суммы квадратов отклонений	Степени свободы	Дисперсия на одну степень свободы
Оцененные регрессией	$V_E = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{Y}\right)^2 = \sigma_{\hat{Y}}^2$	k	$\hat{\sigma}_{\hat{Y}}^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{Y}\right)^2 \Big/ k$
Остаточные	$V_R = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sigma_u^2$	$n-k-1$	$\hat{\sigma}_u^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \Big/ (n-k-1)$
Всего	$V_T = \sum_{i=1}^{n} \left(y_i - \bar{Y}\right)^2 = \sigma_Y^2$	$n-1$	$\hat{\sigma}_Y^2 = \sum_{i=1}^{n} \left(y_i - \bar{Y}\right)^2 \Big/ (n-1)$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{Y})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{Y})^2} = 1 - \frac{\sigma_u^2}{\sigma_Y^2};$$

$$1 - R^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{Y})^2}, \quad \bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-k-1)}{\sum_{i=1}^{n} (y_i - \bar{Y})^2 / (n-1)} = 1 - \frac{\hat{\sigma}_u^2}{\hat{\sigma}_Y^2}.$$

R^2 - измеряет степень вариации зависимой переменной Y, которая объясняется регрессией. $\bar{R}^2$ - измеряет степень вариации зависимой переменной Y, которая объясняется регрессией, приходящейся на одну степень свободы.

$$F_{calc} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \Big/ (k)}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \Big/ (n-k-1)} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \cdot (n-k-1)}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \cdot k}.$$

Правило выбора I_0 для уровня значимости α_0 принимается, если $F_{calc} \prec F_{tab} = F_{1-\alpha;(k-1,n-k)}$ и принимается правило выбора I_1, будучи отвергнута гипотеза I_0, если $F_{calc} \succ F_{tab} = F_{1-\alpha;(k-1,n-k)}$.

$$F_{calc} = \frac{R^2 / k}{\left(1 - R^2\right)/\left(n - k - 1\right)}.$$

6. Мультиколинеарность и ее уменьшение

Мультиколлинеарность проявляется в нарушении основной гипотезы, утверждающей, что ни одна из независимых переменных не является линейной функцией от остальных независимых переменных.

Мультиколлинеарность является следствием того, что исследователь не обратил внимания на идентичные данные среди наблюдаемых значений независимых переменных, чего легко можно избежать путем удаления из уравнения регрессии одной из двух коллинеарных независимых переменных

Мультиколлинеарность может присутствовать и тогда, когда две переменные используются для представления третьей переменной в виде суммы последних двух.

Гипотеза также может быть нарушена и тогда, когда независимая переменная имеет нулевую вариацию. В этом случае независимая переменная коллинеарна свободному члену и оценки с помощью М.Н.К. невозможно получить.

Совершенная мультиколлинеарность двух переменных встречается редко, когда независимые переменные тесно коррелируют, но не нарушается классическая гипотеза, а несовершенная мультиколлинеарность наблюдается чаще. В отличие от однофакторных моделей, в многофакторных моделях гипотеза I_6 предполагает независимость экзогенных (независимых) переменных. Нарушение этой гипотезы приводит к появлению феномена мультиколлинеарности, когда эволюция одной зависисимой переменной объясняется изменениями, которые имеют место в более чем одной независимой переменной.

Относительно высокая частота появления мультиколлинеарности среди независимых переменных является следствием высокой степени взаимозависимости в экономике. О наличии мультисоллинеарности свидетельствует:

аналогии в эволюции независимых переменных;

близость к нулю определителя $\left|X^T X\right|$;

величина коэффициента множественной детерминации $\left(R^2\right)$, которая почти совпадает с его величиной в случае, когда одна из независимых переменных опущена;

противоречивость при проверке тестов, а именно, $F-$тест, примененный для теоретических значений – значим, в то время как тест t, примененный к параметрам уравнения регрессии, отмечает, что некоторые из параметров уравнения регрессии не являются значимыми.

6.1. Совершенная и несовершенная мультиколлинеарность

Совершенная мультиколлинеарность является следствием нарушения одной из основных гипотез, а именно I_6, в которой утверждается, что ни одна из независимых переменных не может быть линейной функцией остальных независимых переменных. В данной интерпретации слово «совершенная» означает, что вариация одной из независимых переменных полностью может быть представлена изменениями в других независимых переменных, например

$X_{1i} = \alpha_0 + \alpha_1 X_{2i}$ в уравнении регрессии $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$.

Приведем примеры линейных зависимостей: $X_{1i} = 3X_{2i}$, $X_{1i} = 6 + X_{2i}$, $X_{1i} = 2 + 4X_{2i}$. В данном примере переменные X_1 *и* X_2 являются совершенно коррелированными.

Совершенная мультиколлинеарность может быть обнаружена до того как будет запущена процедура регрессии, путем проверки, не является ли одна из независимых

переменных произведением другой независимой переменной и константы либо суммой другой независимой переменной и константы. В таком случае одна из независимых переменных должна быть опущена, поскольку между ними нет существенного отличия. *Особый случай мультиколлинеарности* отражает тот факт, что одна из переменных, которая по определению взаимосвязана с зависимой переменной, но, тем не менее, эта переменная включена в уравнение регрессии, как независимая переменная. Такая *доминирующая переменная* настолько тесно коррелирована с зависимой переменной, что полностью перекрывает влияние остальных независимых переменных в уравнении регрессии. Этот тип мультиколлинеарности, по существу, отражает коллинеарность между зависимой и независимой переменными. *Доминирующая переменная* привносит тавтологию; она определена таким образом, что зависимая переменная может быть вычислена без каких либо теоретических обоснований относительно нее. Однако надо быть внимательными, поскольку доминирующая переменная может быть ошибочно принята за зависимую переменную большой значимости либо эта переменная виртуально идентична зависимой переменной.

6.2. Несовершенная мультиколлинеарность

Поскольку совершенная мультиколлинеарность встречается крайне редко, то на практике, как правило, при использовании определения мультиколлинеарности подразумевают несовершенную мультиколлинеарность. Несовершенная мультиколлинеарность может быть определена как линейная взаимосвязь между двумя или более независимыми переменными, которая является достаточно сильной, чтобы существенно влиять на оценки параметров уравнения регрессии. Другими словами несовершенная мультиколлинеарность возникает тогда, когда между двумя или более независимыми переменными существует несовершенная линейная зависимость вида: $X_{1i} = \alpha_0 + \alpha_1 X_{2i} + \varepsilon_i$, поскольку уравнение содержит стохастическую ошибку. Из этого следует, что в случае достаточно сильной зависимости между переменными X_1, X_2, наличие стохастической ошибки не позволяет переменной X_1 быть полностью объясняемой с помощью переменной X_2; остаются некоторые необъяснимые вариации.

Несовершенная мультиколлинеарность является следствием как информационных, так и теоретических недоработок. Почти все макроэкономические данные, представленные временными рядами потенциально мультиколлинеарны между собой, в силу тенденции роста этих агрегатов во второй половине 20 века. Трудовые ресурсы, национальный доход, потребление, налоги и все остальные макропоказатели имеют тенденцию роста. Кроме того, рост производительности и рост общего уровня цен являются еще одним подтверждением тенденции одновременного изменения показателей.

Рассмотрим модель, которая отражает влияние налогов на рост накоплений за последние 30 лет в США : $S_t = f(Y_{dt}^+, i_t^+, T_t^-, SS_t^-)$, где

S_t – *номинальные накопления в году* t,

Y_{dt} – *располагаемый доход в году* t,

i_t – *номинальная процентная ставка в году* t,

T_t – *ставка налогообложения в году* t,

SS_t – *отчисления в социальный фонд в году* t.

Выдвигается гипотеза о том, что накопления будут возрастать одновременно с ростом располагаемого дохода, что рост процентной ставки стимулирует рост накоплений, в то

время как рост ставки налогообложения и отчислений в социальный фонд ведут к уменьшению накоплений.

Экономический рост на протяжении рассматриваемого периода будет содействовать тому, что почти все рассматриваемые переменные будут расти во времени; рост населения, располагаемого дохода, отчислений в социальный фонд делает возможной сильную корреляцию среди многих независимых переменных. В таких моделях возможна достаточно сильная мультиколлинеарность и ее последствия должны быть исследованы.

6.3. Последствия мультиколлинеарности

– Оценки останутся несмещенными $E(\hat{\beta}_i) = \beta_i$.
– Возрастет вариация оценок $\hat{\sigma}^2(\hat{\beta}_i)$.
– Вычисленный t-тест уменьшится.
– Оценки становятся чувствительными относительно изменений в спецификациях.
– Аппроксимация уравнения не будет подвержена изменениям.
– Оценки ортогональных переменных не будет подвержены изменениям.
– Чем сильнее мультиколлинеарность, тем последствия тяжелее.

6.4. Выявление мультиколлинеарности

В реальной действительности почти невозможно найти набор данных для независимых переменных, которые не были бы коррелированны между собой. *Поэтому необходимо выявить степень мультиколлинеарности, а не ее наличие.* Вторым важным моментом является выявление причины мультиколлинеарности: это *феномен данных или следствие теории*. Проблема состоит в выявлении независимой переменной, которая является теоретически обоснованной и статистически не мультиколлинеарна.

Высокий коэффициент детерминации $\overline{R}^2$ на одну степень свободы и низкие t- тесты

Количество незначимых t тестов	Вероятность сильной м.к ($\overline{R}^2$- большой)
все	высокая вероятность
некоторые	возможна
ни один	нет проблем

Высокий коэффициент корреляции

Если коэффициент парной корреляции по абсолютной величине велик, тогда возможно возникновение мультиколлинеарности, $r = 0{,}8$ высокая степень коллинеарности.

6.5. Исключение мультиколлинеарности

Если ряды данных сформированы из небольшого количества наблюдений (n<10), тогда рекомендуется включение дополнительных членов до числа (n > 15), так чтобы случайные аналогии, по возможности, были исключены.

В случае существования сильной корреляции между двумя независимыми переменными, одна из них исключается, считая, что исключенная переменная выражается с помощью той, которая остается в модели.

Если данные представлены в виде хронологических рядов, можно приступить к расчету конечных разностей первого порядка $(\Delta = Y_i - Y_{i-1})$ или прологарифмировать

значения $Y_i, X_{1i}, X_{2i}, \cdots, X_{ki}$ в целях уменьшения коллинеарности, обусловленной наличием тренда в данных.

6.6. Процедура выбора независимых переменных в модели множественной корреляции

Ликвидация феномена коллинеарности влечет за собой вычисление коэффициентов корреляции между независимыми переменными r_{X_i/X_j} и r_{Y/X_i} линейными коэффициентами корреляции между зависимой переменной Y и соответствующими ей независимыми переменными X_i. Если $r_{X_i/X_j} \approx 1,\ i \neq j$, то одну из рассматриваемых переменных необходимо будет исключить из числа независимых .

Критерием исключения/включения в уравнение регрессии одной из двух сильно коррелированных переменных служит следующее правило. Если $r_{X_j/X_i} \succ r_{X_j/X_j}$, исключается переменная X_j и сохраняется X_i, в противном случае, исключается переменная X_i и сохраняется переменная X_j. Таким образом, сохранив k линейно независимых экзогенных переменных на первом этапе, имея возможность оценить $(k+1)$ параметр уравнения регрессии, можно переходить ко второму этапу, на котором продолжается отбор экзогенных переменных X_i. Для этих целей имеется несколько способов.

<u>*Первый способ*</u>.

В модель включаются k независимых переменных, порядок включения которых определяется величиной коэффициентов корреляции переменной Y в соответствии с каждой из отобранных независимых переменных $r_{Y/X_1} \succ r_{Y/X_2} \succ r_{Y/X_3} \succ \cdots\ r_{Y/X_k}$, таким образом, получаем k моделей:

$$M(1):\quad Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \hat{u}_i^1 = \hat{Y}_i^1 + \hat{u}_i^1$$

$$\vdots$$

$$M(j):\quad Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \cdots + \hat{\beta}_j X_{i,j} + \hat{u}_i^j = \hat{Y}_i^j + \hat{u}_i^j$$

$$\vdots$$

$$M(k):\quad Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \cdots + \hat{\beta}_j X_{i,j} + \cdots + \hat{\beta}_k X_{i,k} + \hat{u}_i^k = \hat{Y}_i^k + \hat{u}_i^k$$

Известно, что общая вариация переменной Y равна вариации, которая объясняется уравнением регрессии $M(j)$, плюс остаточная вариация $\sigma_{Y^j}^2 = \sigma_{\hat{Y}^j}^2 + \sigma_{u^j}^2$. Из последнего уравнения легко получить коэффициент детерминации $\overline{R}_j^2 = 1 - \dfrac{\hat{\sigma}_{u^j}^2}{\hat{\sigma}_{Y^j}^2}$, который измеряет долю вариации зависимой переменной Y^j, объясняемой уравнением регрессии, в общей вариации. Коэффициент детерминации $\overline{R}_j^2$ представляет собой долю общей вариации, которая не может быть объяснена уравнением регрессии $M(j)$. На основе этих взаимосвязей могут быть сформулированы критерии отбора оптимальной модели уравнения регрессии $M(r)$ из множества моделей $M(j)$, а именно, $\sigma_{\hat{Y}r}^2 = \max\limits_j \sum\limits_{i=1}^{n} \left(\hat{Y}_i^j - \overline{Y}\right)^2$ или $\overline{R}_r^2 = \max\limits_j \overline{R}_j^2$, или $\sigma_{ur}^2 = \min\limits_j \sum\limits_{i=1}^{n} \left(Y_i^j - \hat{Y}_i^j\right)^2$, предварительно проверив уровень значимости этих величин с помощью теста F .

<u>*Второй способ*</u>.

Этот способ основан на предположении о том, что $(k+1)$ независимые переменные, определяющие влияние на зависимую переменную Y, являются линейно независимыми. При выполнении этих условий, существует обратная матрица $\left(X^T X\right)^{-1}$, с помощью которой вычисляются оценки параметров уравнения регрессии $\left(\widehat{\beta}_j\right)$ и соответствующие им дисперсии, получив

$M_k: \quad Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i,1} + \cdots + \widehat{\beta}_r X_{i,r} + \cdots + \widehat{\beta}_k X_{i,k} + \widehat{u}_i^k$. Затем, с помощью t теста, тестируется значимость оценок $\left(\widehat{\beta}_j\right)$ для уровня значимости α и для $(n-(k+1))$ степеней свободы.

Если $\dfrac{\left|\widehat{\beta}_j\right|}{\sigma_{\widehat{\beta}_j}} \geq t_{\alpha,(n-(k+1))}$, тогда $\left(\widehat{\beta}_j\right)$ существенно отличаются от нуля. Предполагая, что $\left(\widehat{\beta}_j\right)$ отличны от нуля для $j = 0,1,\cdots,r$, и что $\left(\widehat{\beta}_j\right)$ не отличаются существенно от нуля для $j = r+1,\cdots,k$, а это означает, что независимые переменные $\left(X_j\right), j \succ r$ не оказывают существенного воздействия на зависимую переменную Y и могут быть выведены из уравнения регрессии, и тогда, модель будет составлена на основе независимых переменных $\left(X_j\right), j \leq r$.

<u>*Третий способ.*</u> Заключается в проверке тестов на мультиколлинеарность.

Тест Klein

Этот тест основан на сравнении коэффициента детерминации R^2 для модели регрессии с k независимыми переменными: $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_k X_k + u$ и коэффициентов парной корреляции $r^2_{X_i/X_j}$ между независимыми переменными при $i \neq j$. Если $R^2_Y \prec r^2_{X_i/X_j}$, то существует вероятность мультиколлинеарности.

Тест Farrar и Glauber

Первый этап. Вычисляется определитель матрицы коэффициентов корреляции

$$D = \begin{vmatrix} 1 & r_{X_1/X_2} & r_{X_1/X_3} & \cdots & r_{X_1/X_k} \\ r_{X_2/X_1} & 1 & r_{X_2/X_3} & \cdots & r_{X_2/X_k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{X_k/X_1} & r_{X_k/X_2} & \cdots r_{X_r/X_{k-1}} & & 1 \end{vmatrix}$$

Если значение определителя мало отличается от нуля, вероятность мультиколлинеарности велика. Например, если для модели из двух переменных, числовые ряды значений независимых переменных абсолютно коррелированы, то определитель $D = \begin{vmatrix} 1 & r_{X_1/X_2} \\ r_{X_2/X_1} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 0$, в противном случае, когда ряды ортогональны, определитель $D = \begin{vmatrix} 1 & r_{X_1/X_2} \\ r_{X_2/X_1} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1$.

<u>Второй этап.</u> Выполняется тест χ^2 и проверяется справедливость следующих гипотез:

$I_0: \quad D = 1$ (ряды ортогональны)

$I_1: \quad D \prec 1$ (ряды взаимозависимы)

Эмпирическое значение χ^2, вычисленное для выборки из n наблюдений и K независимых переменных ($K = k+1$, если постоянный член включен в уравнение регрессии), определяется следующим образом: $\chi^2_{cfkc} = -[n-1-1/6(2*k+5)]*Ln(D)$. Если $\chi^2_{calc} \leq \chi^2_{tab}$ при n*(k-1)/2 степенях свободы и уровне значимости α, тогда отвергается

гипотеза I_0 и вероятно наличие мультиколлинеарности. Если $\chi^2_{calc} \succ \chi^2_{tab}$, тогда принимается гипотеза ортогональности.

Для исключения мультиколлинеарности также применяется: метод последовательной регрессии

7. Средства против мультиколлинеарности

Что можно предпринять, чтобы ослабить последствия, которые оказывает сильная мультиколлинеарность на исследуемое уравнение регрессии? На этот вопрос нельзя ответить однозначно, поскольку мультиколлинеарность является таким феноменом, который меняется от одного набора данных к другому даже, если уравнение регрессии не меняется.

7.1. Не предпринимать ничего

На первом этапе, когда выявлена сильная мультиколлинеарность, необходимо решить, надо ли что-то предпринимать. Как удалось заметить, в случае, когда любое средство против мультиколлинеарности вызывает определенные сомнения, правильным действием будет – не предпринимать ничего.

Другим важным мотивом в пользу не принятия никаких мер является тот, что наличие мультиколлинеарности в уравнении не всегда уменьшает t-статистику настолько, чтобы она стала незначимой или не настолько меняет коэффициент регрессии $\hat{\beta}_s$, чтобы его значение существенно отличалось от ожидаемого значения. Другими словами простое присутствие мультиколлинеарности не всегда что-то обозначает. К примеру, если простой коэффициент корреляции между двумя независимыми переменными равен 0.97, в то время как каждая переменная имеет значимую индивидуальную статистику на 95% уровне доверия, не имеет смысла применять какое либо средство для ликвидации мультиколлинеарности.

При наличии сильной мультиколлинеарности, самое простое средство состоит в исключении одной или более независимых переменных из уравнения регрессии. Однако, к сожалению, удаление мультиколлинеарных переменных, которые, согласно теории, принадлежат уравнению регрессии, является опасной операцией, поскольку может подвергнуть уравнение смещениям в спецификации. Удалением такой переменной, преднамеренно создаются смещения в оценках. Поэтому любые попытки избежать мультиколлинеарности, неоправданны по сравнению с риском исключения независимых переменных. И, как следствие, опытные специалисты в эконометрике, в большинстве случаев, сохранят мультиколлинеарность в уравнении регрессии, несмотря на возможное уменьшение t-статистик.

Последний аргумент в пользу утверждения ничего не предпринимать в целях предотвращения мультиколлинеарности является теоретическим, который применим к любому уравнению регрессии. Всякий раз, когда запускается регрессия, существует риск обнаружить подходящую, хотя и случайную, спецификацию для выбранного набора данных, которая, тем не менее, не является истинной. Увеличение числа попыток запуска регрессии увеличивает шансы получения случайных результатов. Следовательно, последовательная спецификация приемлема тогда, когда существует сильная мультиколлинеарность, поскольку в этом случае оценка коэффициентов чувствительна к изменениям в спецификации. Итак, зачастую лучше оставить уравнение регрессии неизменным, противодействуя различного вида мультиколлинеарности, за исключением чрезмерной мультиколлинеарности.

Это пожелание тяжело принимается начинающими исследователями, когда они стоят пред выбором окончательного варианта регрессии и низким уровнем значимости t-

статистик. Тем не менее, по сравнению с возможной альтернативой получения смещений, обусловленных исключением, важных с теоретической точки зрения, переменных или случайных результатов, низкие t-статистики кажутся незначительной проблемой.

7.2. Исключение одной или более мультиколлинеарных переменных

Наиболее надежный способ, чтобы избежать присутствия существенной мультиколлинеарности в уравнении, состоит в удалении всех мультиколлинеарных переменных. Мультиколлинеарность обусловлена корреляцией между независимыми переменными; уравнение при удалении коррелированных переменных не подвержено более коллинеарности и все проблемы, связанные с мультиколлинеарностью снимаются. Коэффициенты при оставшихся независимых переменных измеряют почти все общее воздействие на зависимую переменную исключенных независимых коллинеарных переменных.

Чтобы продемонстрировать, как работает эта методика, рассмотрим следующий пример:

$$\widehat{C}_i = -367.83 + 0.5113 * Y_{di} + 0.0427 * LA_i$$

$$\begin{array}{llll} & (1.0307) & (0.0942) & \\ t = & 0.496 & 0.453 & \overline{R}^2 = 0.835 \end{array}$$

С – потребление; Y_d - располагаемый доход; LA - ликвидные активы

$$\widehat{C}_i = -471.43 + 0.9714 * Y_{di}$$

$$\begin{array}{lll} & (0.157) & \\ t = & 6.187 & \overline{R}^2 = 0.861 \end{array}$$

$$\widehat{C}_i = -199.44 + 0.08876 * LA_i$$

$$\begin{array}{lll} & (0.01443) & \\ t = & 6.153 & \overline{R}^2 = 0.860 \end{array}$$

Отметим, что исключение одной из коррелированных переменных не только удаляет мультиколлинеарность между двумя независимыми переменными, но и увеличивает значение t-статистик при коэффициентах оставшихся независимых переменных. Исключая независимую переменную Y_d, мы увеличиваем t_{LA} от значения 0.453 до значения 6.153. В то же время, исключение независимой переменной меняет значение оставшегося коэффициента, и такие резкие изменения не являются исключением.

Предположим, что есть желание исключить какую-то переменную, как решить какую из переменных исключать? В случае сильной мультиколлинеарности не имеет значения, какая из переменных будет исключена. Не имеет смысла выбирать исключаемую переменную по принципу, что одна больше подходит или, что другая более значима (или знак при переменной тот, который предполагался) в истинном уравнении регрессии. Теоретическое обоснование модели должно быть основанием для принятия решений подобного рода. В представленном примере существует теоретическая мотивация того, что располагаемый доход определяет потребление в большей степени, чем ликвидные активы. Во многих случаях, решение об исключении одной из коррелированных переменных является хорошим решением. Например, некоторые начинающие исследователи включают в уравнение регрессии слишком много независимых переменных, желая избежать смещения в оставшихся переменных. И, как следствие,

чаще всего они имеют одну или несколько независимых переменных в уравнении, которые по существу преследуют одни и те же цели. В этом случае, коррелированные переменные вполне могут быть обоснованными как с теоретической, так и статистической точки зрения. Однако переменные могут быть определены как бесполезные, поскольку только одна из этих переменных необходима для определения влияния на зависимую переменную. Например, для функции агрегированного спроса не имеет смысла вводить в качестве независимых переменных располагаемый доход и валовой внутренний продукт, поскольку они измеряют один и тот же показатель – доход. Более тонким является решение об одновременном не включении в уравнение агрегированного спроса показателей располагаемого дохода и населения, поскольку вновь они оценивают, в самом деле, одни и те же понятия – объем агрегированного рынка. С ростом населения возрастет и доход. Исключение такого рода бесполезных мультиколлинеарных переменных не приведет ни к чему, кроме как к увеличению ошибки спецификации.

7. 3. Преобразование мультиколлинеарных переменных

Очень часто в уравнениях, для которых последствия мультиколлинеарности являются настолько серьезными, что необходимо принимать меры по их ликвидации, все независимые переменные являются исключительно важными с теоретической точки зрения. В таких случаях ни бездействие, ни исключение переменных не являются полезными. Тем не менее, иногда возможно преобразование независимых переменных из рассматриваемого уравнения, которое ослабит мультиколлинеарность. Приведем два наиболее часто используемые преобразования:

– Создание некоторой линейной комбинации из числа коррелированных переменных.
– Преобразование уравнения в конечные разности.

Техника создания линейной комбинации из двух или более мультиколлинеарных переменных состоит:

а) в создании новой переменной, которая является функцией от мультиколлинеарных переменных;

б) в использовании полученной переменной для замены старых переменных в уравнении регрессии.

Например, если переменные X_1 и X_2 сильно коррелированны, то новая переменная $X_3 = X_1 + X_2$ или, в общем случае, любая комбинация вида ($X_3 = k_1 X_1 + k_2 X_2$) может заменить обе коррелированные переменные во вновь оцененной модели. Такой способ используется в тех случаях, когда уравнение предполагается использовать для данных, отличных от наблюдаемых значений, поскольку в этом случае мультиколлинеарность может отсутствовать либо может следовать тем же шаблонам, как и в случае данных, зарегистрированных внутри интервала наблюдений. Большим недостатком этого подхода является то, что обе переменные имеют одинаковый коэффициент в оцененном уравнении регрессии, а именно, $X_{3i} = X_{1i} + X_{2i} \Rightarrow Y_i = \widehat{\beta}_0 + \widehat{\beta}_3 X_{3i} = \widehat{\beta}_0 + \widehat{\beta}_3 (X_{1i} + X_{2i}) + u_i$. Необходимо аккуратно подходить к выбору линейной комбинации для переменных, у которых ожидаются различные коэффициенты регрессии (как и разные знаки) или большая разница в значениях переменных (отличие на порядок и больше), которая не может быть нивелирована с помощью коэффициентов пропорциональности (k_3) в уравнении общего вида ($X_3 = k_1 X_1 + k_2 X_2$). Например, если двумя коррелированными переменными являются ВВП и инфляция, то простая сумма может полностью поглотить вариацию инфляции (в зависимости от единиц измерения переменных).

$X_{3i} = GNP_i + INF_i = 3.250 * GNP_i + 0.08 * INF_i$. Если GNP удвоится, то X_3 также увеличится в два раза. Если же удвоиться инфляция, то X_3 почти не измениться. Следовательно, при подборе линейной комбинации, необходим тщательный расчет средних ожидаемых значений для коэффициентов регрессии, которые входят в эту линейную комбинацию. Другими словами, переменные могут либо погашать друг друга, либо ликвидировать друг друга.

Чтобы познакомиться с примером такого типа сформируем линейную комбинацию между располагаемым доходом и ликвидными активами как функцию от потребления, и запустим регрессию с линейной комбинацией независимых переменных. Для того чтобы сбалансировать переменные, располагаемый доход умножим на 10, в результате чего получим:

$$\widehat{C}_i = -355.43 + 0.0467 * X_{3i}$$

$$(0.0073)$$

$$t = \quad 6.362 \qquad R_2 = 0.868$$

Сравнивая с предыдущими результатами, видим, что удаление мультиколлинеарности вновь существенно повышает t-статистику независимой переменной, в то время как оказывает незначительное воздействие на значимость уравнения. Интересно отметить, что оцененный коэффициент может быть вычислен по предыдущим оценкам уравнения с помощью линейной комбинации.

Другой вид преобразований, который может быть рассмотрен как возможное средство для удаления сильной мультиколлинеарности, состоит в преобразовании формы уравнения регрессии. Посмотрим, как преобразование уравнения с помощью конечных разностей первого порядка снизит уровень мультиколлинеарности набора данных (можно было бы рассмотреть и log преобразования, а также преобразования с помощью других функциональных форм, но все они основаны на тех же принципах). Конечные разности первого порядка есть не что иное, как изменение в переменной предыдущего периода и текущего периода, которая обозначается как $\Delta X_t = X_t - X_{t-1}$.

Если уравнение в целом или несколько переменных из уравнения представить в конечных разностях первого порядка, то степень мультиколлинеарности существенно уменьшится по двум причинам. Первая, любое изменение в определении переменных (за исключением линейных преобразований) уменьшает уровень мультиколлинеарности. Вторая, мультиколлинеарность чаще всего имеет место (понятно, однако, что не исключительно) во временных рядах данных, в которых разности первого порядка совершенно не похожи на постоянно возрастающие смещения, по сравнению с агрегатами, по которым они считаются. Например, годовой прирост ВВП составляет 5-6%% , в то время как годовые изменения в нем (конечные разности первого порядка) могут сильно колебаться. И как следствие, преобразование уравнения в целом или какой-то его части, к конечным разностям в состоянии снизить мультиколлинеарность в моделях с временными рядами данных.

Если сильная мультиколлинеарность иногда может быть уменьшена с помощью преобразования в конечные разности или с помощью других преобразований, изменение функциональной формы уравнения с целью избавления от мультиколлинеарности в большинстве случаев не приводит к теоретическим осложнениям. Например, моделирование запасов по своей природе отличается от моделирования изменений в запасах капитала, которые представляют инвестиции, несмотря на то, что одно уравнение следует из другого. Если основной целью запуска регрессии является моделирование конечных разностей, тогда модель регрессии может быть специфицирована в таком виде.

При этом стоит заметить, что при вычислении конечных разностей первого порядка, число степеней свободы уменьшается на единицу.

7.4. Увеличение числа наблюдений

Другой способ борьбы с мультиколлинеарностью состоит в попытке увеличения количества наблюдений таким образом, чтобы снизить уровень мультиколлинеарности.

Основная идея этого способа, состоит в том, что большее количество наблюдений (часто требующее нового сбора данных) позволит выполнить более точные оценки, в то время как больший набор данных, естественно, уменьшит тем или иным способом вариации оцененных коэффициентов, уменьшая последствия мультиколлинеарности, даже если уровень мультиколлинеарности остается тем же.

Тем не менее, для большинства приложений из экономики и бизнеса такой подход не возможен. Набор данных составлен из сопоставимых величин, в то время как новые данные трудно найти либо они очень дорого стоят.

8. Корреляция рядов данных (Автокорреляция)

Корреляция рядов, также называемая автокорреляцией, существует в научных исследованиях, для которых является важным порядок следования наблюдений. Поэтому чаще всего автокорреляция возникает на множестве данных, описываемых временными рядами. *По существу, из корреляции рядов данных следует, что член стохастической ошибки для одного временного периода симметрично зависит от члена стохастической ошибки для другого временного периода.* И, поскольку временные ряды используются во многих эконометрических приложениях, очень важно понять смысл корреляции рядов данных и ее последствия для оценок, полученных по М.Н.К.

Попытаемся ответить на следующие вопросы:

Какова сущность проблемы?

Каковы последствия наличия этой проблемы?

Насколько опасна эта проблема?

Какие существуют способы, чтобы решить эту проблему?

Корреляция наблюдений члена стохастической ошибки во времени называется корреляцией ряда данных. Далее будет обсуждаться природа этой проблемы и отличие двух форм корреляции ряда данных: «совершенной» и «несовершенной».

8.1 Совершенная корреляция ряда данных

Совершенная корреляция ряда данных возникает тогда, когда нарушена классическая гипотеза относительно некоррелированности наблюдений члена стохастической ошибки в <u>корректно</u> <u>специфицированном уравнении регрессии</u>. Напомним, что классическая гипотеза утверждает: $E(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$. Если произведение любых двух наблюдаемых значений члена стохастической ошибки не равно нулю, то будем говорить, что имеет место автокорреляция остатков.

Наиболее часто встречается автокорреляция остатков первого порядка: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, где ε_t член стохастической ошибки в исследуемом уравнении, ρ - параметр, описывающий функциональную зависимость между наблюдаемыми значениями члена стохастической ошибки, u_t - член классической стохастической ошибки, который не автокоррелирован. Эта функциональная форма представляет так называемую схему Маркова первого порядка с коэффициентом автокорреляции первого порядка ρ. Такая автокорреляция характеризуется тем, что одно из наблюдаемых значений члена

стохастической ошибки непосредственно воздействует на другое наблюденное значение стохастической ошибки. Величина коэффициента ρ указывает на то, как велика степень автокорреляции в уравнении регрессии. Если коэффициент ρ равен нулю, то автокорреляция отсутствует (поскольку ε_t равно u_t- члену классической стохастической ошибки). Если ρ по абсолютной величине не больше единицы, предыдущее значение члена стохастической ошибки становится определяющим при нахождении текущего значения члена стохастической ошибки и существует большая вероятность автокорреляции. Если же ρ больше единицы по абсолютной величине, то появляется тенденция абсолютного роста члена стохастической ошибки, и потому такие значения ρ нет смысла рассматривать. И, следовательно, будут рассматриваться значения ρ из интервала $-1 \prec \rho \prec 1$. Знак ρ указывает на характер корреляции в уравнении. Положительный знак ρ указывает на то, что член стохастической ошибки имеет тенденцию сохранения положительного знака в будущем при переходе от одного временного интервала к другому. Такая тенденция означает, что, если ε_t будет иметь большие значения в некоторый период времени, то следующие наблюдаемые значения будут стремиться к сохранению части из этих больших оригинальных значений и будут иметь тот же знак, что и оригинальное значение. Например, в модели с временными рядами исключительно большой шок в экономике в определенный момент времени будет иметь продолжение и в последующие моменты времени. Если это имеет место, то значения члена стохастической ошибки будут иметь тенденцию к сохранению положительного знака для нескольких последующих наблюдаемых значений, затем знак для нескольких значения наблюдений станет отрицательным, а через некоторое время он опять примет положительное значение. Этот феномен называется положительной автокорреляцией остатков.

Например, положительная автокорреляция может встречаться в уравнениях спроса на сезонные товары или блага (как новогодние гирлянды), которые не имеют фиктивной переменной. Как бы то ни было, в большинстве приложений, описываемых временными рядами, отрицательная автокорреляция встречается гораздо реже, чем положительная автокорреляция.

Автокорреляция может принимать и другие формы, отличные от автокорреляции первого порядка. Например, в квартальных моделях наблюдаемые значения стохастической ошибки могут функционально зависеть от тех же наблюдаемых квартальных данных, но предыдущего года: $\varepsilon_t = \rho\varepsilon_{t-4} + u_t$.

Также наблюдаемые значения члена стохастической ошибки уравнения могут зависеть от нескольких предыдущих наблюдаемых значений члена стохастической ошибки: $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + u_t$, такая корреляция носит название автокорреляции второго порядка.

8.2. Несовершенная автокорреляция

Под несовершенной автокорреляцией будем понимать такую автокорреляцию, которая вызвана ошибками спецификации, к которым можно отнести: опущенные переменные или неверно подобранную функциональную форму. В то время как совершенная автокорреляция вызвана распределением стохастической ошибки в истинном (теоретическом) уравнении регрессии (которое не может быть изменено), несовершенная автокорреляция обусловлена ошибками спецификации, которые в большинстве случаев могут быть скорректированы.

Как получается, что ошибки спецификации провоцируют автокорреляцию? Вспомним, что член стохастической ошибки может быть представлен как следствие не

включения в уравнение регрессии важных, с теоретической точки зрения, независимых переменных, нелинейности, ошибок измерения и попросту стохастических отклонений независимой переменной. А это означает, что если мы не включаем важную переменную или используем неадекватную функциональную форму, тогда часть влияния опущенной переменной, которая не может быть отражена оставшимися независимыми переменными, поглощается членом стохастической ошибки. Следовательно, при неверной спецификации уравнения регрессии, член стохастической ошибки будет включать в себя часть влияния всех не включенных в уравнение регрессии независимых переменных и долю влияния, оцененного разностью между функциональной формой, выбранной исследователем, и теоретической функциональной формой. В этом случае автокорреляция остатков обусловлена выбором спецификации исследователем, а не членом стохастической ошибки и полностью ассоциирующимся с верной спецификацией.

Средства борьбы с автокорреляцией зависят от типа автокорреляции: совершенной либо несовершенной. Не является неожиданностью тот факт, что при несовершенной автокорреляции, лучшим средством является отказ от попытки исключить из уравнения регрессии независимые переменные. И, как правило, большинство исследователей пытаются убедиться в том, что они получили возможно наилучшую спецификацию до проведения изнурительной и, требующей больших временных затрат, процедуры по проверке автокорреляции остатков.

Для того, чтобы убедится в том, как исключение независимых переменных может вызвать автокорреляцию остатков, предположим, что истинное уравнение регрессии имеет вид: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$, в котором ε_t является классическим членом стохастической ошибки. Если, независимая переменная X_2 будет случайно исключена из уравнения, (или для независимой переменной X_2 не существует данных), тогда $Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t^*$, где $\varepsilon_t^* = \beta_2 X_{2t} + \varepsilon_t$. Следовательно, член стохастической ошибки, используемый при не включении в уравнение регрессии важных независимых переменных, не является классическим членом стохастической ошибки ε_t, в то же время, являясь функцией независимой переменной X_2. И, как следствие, член стохастической ошибки $\varepsilon^*{}_t$ может быть автокоррелированным в то время, как истинный член стохастической ошибки не является автокоррелированным. В частности, новый член стохастической ошибки $\varepsilon^*{}_t$ будет подвержен автокорреляции тогда, когда:

а) переменная X_2 сама автокоррелированна (что, само по себе, с большой вероятностью возможно при использовании временных рядов);

б) величина ε_t мала по сравнению с величиной $\beta_2 X_{2t}$.

Эти тенденции поддерживаются, если существуют одна или более опущенных переменных. Отметим следующий факт, значение ошибки $\varepsilon^*{}_t$ отлично от нуля, поскольку существует несовершенная автокорреляция, оценка свободного члена β_0, полученная по М.Н.К., будет скорректирована в соответствии с возникшей проблемой. Во-вторых, поскольку несовершенная автокорреляция остатков связана с ошибками спецификации типа опущенных переменных, то вероятны смещения оценок коэффициентов уравнения регрессии. Одновременно с удалением ошибок спецификации, исчезнут смещения оценок и несовершенная автокорреляция остатков.

Рассмотрим спрос на рыбу, который является примером того как опущенные переменные могут вызвать несовершенную автокорреляцию остатков в неправильно специфицированном уравнении регрессии. Пусть $F_t = \beta_0 + \beta_1 RP_t + \beta_2 \ln Y_{dt} + \beta_3 D_t + \varepsilon_t$, где F_t - потребление рыбы на одного жителя в текущем году t, RP_t - цена рыбы по отношению к говядине в текущем году t, Y_{dt} - реальный располагаемый доход на одного

жителя в текущем году t, D_t - фиктивная переменная, которая равна нулю до принятия Папой вердикта относительно рыбы и равна единице после, ε_t классический член стохастической ошибки (не автокоррелированный). Предположим, что данное уравнение корректно специфицировано, что случится с уравнением, если исключить из него переменную реального располагаемого дохода? $F_t = \beta_0 + \beta_1 RP_t + \beta_3 D_t + \varepsilon_t^*$. Очевидно, что произойдут смещения оценок при независимых переменных RP_t и D_t, которые зависят от степени корреляции переменных RP_t и D_t с переменной Y_{dt}. Поскольку $\varepsilon^*{}_t = \varepsilon_t + \beta_2 \ln Y_{dt}$, новый член стохастической ошибки включает значительную часть влияния, оказываемую располагаемым доходом на потребление рыбы, то возникает вторичный эффект, обусловленный этим. Уместно ожидать, что располагаемый доход (и натуральный логарифм располагаемого дохода) может следовать умеренному шаблону автокорреляции остатков: $\ln Y_{dt} = f(\ln Y_{dt-1}) + u_t$ А, если располагаемый доход автокоррелирован (и если его влияние не меньше чем ε_t), тогда $\varepsilon^*{}_t$, с большой степенью вероятности, будет автокоррелирован, что может быть записано, как: $\varepsilon_t^* = \rho\varepsilon_{t-1}^* + u_t$, где ρ представляет коэффициент автокорреляции остатков, а u_t - член классической стохастической ошибки. Этот пример показал, что вполне вероятно, чтобы опущенная переменная была источником несовершенной автокорреляции остатков.

Другой, не менее распространенный вид автокорреляции остатков связан с неправильным выбором функциональной формы. Некорректная функциональная форма может стать источником несовершенной автокорреляции остатков. Предположим, что истинное уравнение регрессии представлено в полной логарифмической форме: $\ln Y_t = \beta_0 + \beta_1 \ln X_{1t} + \varepsilon_t$, однако вместо этого уравнения запускается уравнение регрессии следующего вида: $Y_t = \alpha_0 + \alpha_1 X_{1t} + \varepsilon_t^*$, новый член стохастической ошибки $\varepsilon^*{}_t$ сейчас является функцией от члена классической стохастической ошибки ε_t и от разности между линейной формой и полной логарифмической формой. Как следует из рисунка, эти разности умеренно автокоррелированы. За положительными разностями следуют положительные, а за отрицательными разностями следуют отрицательные разности. Следовательно, использование линейной функциональной формы, тогда когда предпочтительнее использование нелинейной формы, как правило, вызывает несовершенную автокорреляцию остатков. Использование неправильной функциональной формы группирует положительные и отрицательные остатки вместе приводя к несовершенной автокорреляции.

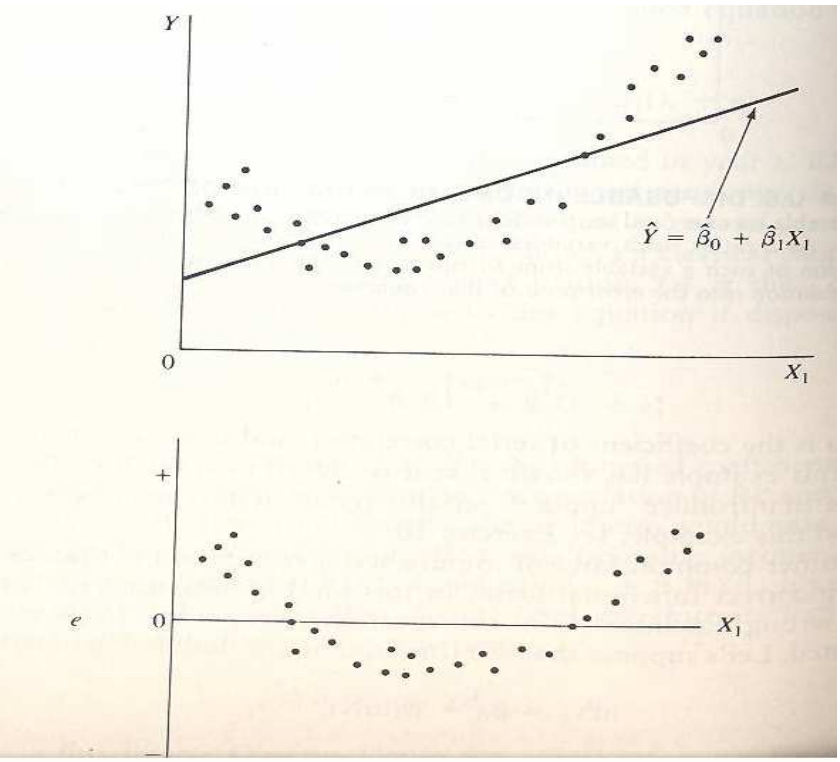


Рис. Неправильно подобранная функциональная форма - источник автокорреляции.

9. Последствия автокорреляции остатков

По своему характеру, последствия автокорреляции остатков полностью отличаются от рассмотренных ранее задач: не включенные переменные, не существенные переменные и мультиколлинеарность. Каждая из этих проблем меняет оцененные коэффициенты и стандартные ошибки определенным образом, и анализ этих изменений, в большинстве случаев, предоставляет достаточно информации, позволяющей решить соответствующую проблему. Как будет отмечено, автокорреляция остатков с большой вероятностью имеет внутренние симптомы, она воздействует на уравнение регрессии, таким образом, который нелегко наблюдать, анализируя результаты как таковые.

Существуют следующие последствия автокорреляции остатков:

а) совершенная автокорреляция остатков не вызывает смещений в оцененных коэффициентах;

б) автокорреляция остатков способствует росту вариаций распределений β.

в) автокорреляция остатков ведет к недооценке вариаций (и стандартных ошибок) коэффициентов, оцененных по М.Н.К.

Синтез последствий автокорреляции остатков:

9.1. Совершенная автокорреляция остатков не вызывает смещений оценок коэффициентов

Предположим, что член стохастической ошибки в уравнении вида: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$ подвержен совершенной автокорреляции остатков первого порядка: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, где u_t классический член стохастической ошибки (не автокоррелированый). Если уравнение корректно специфицировано и оценено с помощью М.Н.К., то полученные оценки коэффициентов будут несмещенными: $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$. Совершенная автокорреляция остатков не вызывает смещений оценок коэффициентов в процессе их оценки. Этот вывод справедлив как для положительной автокорреляции, так и для отрицательной автокорреляции остатков

первого порядка. Если же автокорреляция несовершенна, то смещения, как бы то ни было, могут быть внесены при использовании неверной спецификации.

Отсутствие смещений не означает с необходимостью, что оценки, полученные с помощью М.Н.К. для уравнения, в котором присутствует автокорреляция остатков, будут близки к истинным значениям коэффициентов; единственная наблюденная оценка в действительности может следовать из большого числа возможных значений. Кроме того, стандартная ошибка этих оценок будет, как правило, увеличена вследствие автокорреляции остатков. Это повышает вероятность того, что $\hat{\beta}$ будет существенно отличаться от истинных значений β. Несмещенные оценки в этом случае имеют распределение $\hat{\beta}_s$, которое центрировано вокруг истинных значений β.

9.2. Автокорреляция остатков увеличивает вариации распределений $\hat{\beta}$

Несмотря на то, что нарушение классической гипотезы не вызывает смещений, оно, тем не менее, может воздействовать на основной вывод теоремы Гаусса-Маркова, которая касается минимальных вариаций. В частности, нельзя доказать, что оценки $\hat{\beta}_s$, полученные М.Н.К., обладают минимальной вариацией в случае, когда классическая гипотеза нарушена. И, как следствие, член стохастической ошибки автокоррелирован тогда, когда М.Н.К. не обеспечивает минимальную вариацию для оцененных коэффициентов.

Корреляция значений члена стохастической ошибки заставляет зависимую переменную колебаться в той же фазе, в которой процедура М.Н.К. предписывает независимым переменным. Следовательно, с большой вероятностью, можно утверждать, что М.Н.К. при наличии автокорреляции недооценивает истинные значения β. Для балансировки, $\hat{\beta}_s$ остаются несмещенными, поскольку с равной вероятностью возможна как переоценка, так и недооценка истинных значений. В любом случае, эти ошибки увеличивают вариацию распределения оценок, в то же время увеличивая величину, на которую любые, полученные оценки, будут отличаться от истинных значений β. Можно доказать, что если член стохастической ошибки распределен как: $\varepsilon_t = \rho\varepsilon_{t-4} + u_t$, тогда вариация $\hat{\beta}_s$ является функцией от коэффициента ρ. И, чем больше значение ρ, тем больше вариация $\hat{\beta}_s$.

9.3. При наличии автокорреляции остатков, М.Н.К. недооценивает вариации (стандартные ошибки) коэффициентов

Если автокорреляция остатков увеличивает вариации (стандартные ошибки) коэффициентов $\hat{\beta}_s$, тогда можно предположить, что $\hat{\sigma}(\hat{\beta}_s)$, полученные по М.Н.К., также возрастут, однако не всегда, в то же время эти $\hat{\sigma}(\hat{\beta}_s)$ демонстрируют тенденцию к уменьшению. Следовательно, автокорреляция остатков увеличивает стандартные отклонения оцененных коэффициентов, но в завуалированном виде, через оценки М.Н.К.

Использование М.Н.К. для оценки коэффициентов уравнения регрессии с автокоррелированными остатками может привести к недооценке стандартных ошибок. Автокорреляция остатков соответствует той области наблюдений, которая обеспечивает лучшую аппроксимацию, чем та, которую может обеспечить уравнение, в отсутствии эффекта автокорреляции остатков. Лучшие приближения следуют не только из недооценки стандартных ошибок коэффициентов $\hat{\beta}_s$, но и из стандартных ошибок остаточных переменных, так что ни на t - статистики, ни на F - статистику нельзя положиться при наличии автокорреляции остатков.

В частности, феномен недооценки методом наименьших квадратов $\hat{\sigma}(\hat{\beta}_s)$будет содействовать переоценке t - статистик оцененных коэффициентов, поскольку: $t = \frac{(\hat{\beta} - \beta_{H_0})}{\hat{\sigma}(\hat{\beta})}$. Если слишком маленькое значение $\hat{\sigma}(\hat{\beta})$ вызывает слишком высокое значение t - статистики для определенного коэффициента, тогда с большой вероятностью будет отвергнута гипотеза $H_0 : (\beta = 0)$, в то время как она на самом деле является справедливой. По существу, в этом случае М.Н.К. приводит исследователя к конфузу по поводу значимости конкретного результата. Автокорреляция остатков не только увеличивает стандартные отклонения, но и зачастую ведет к ошибочным выводам, которые затрудняют, получение этого роста с помощью М.Н.К.

9.4. Тест Durbin-Watson

Наиболее часто используемым тестом для определения автокоррелированности остатков является d - тест или тест Durbin-Watson.

9.4.1.Статистика Durbin-Watson

Статистика Durbin-Watson d используется для определения автокоррелированности остатков первого порядка путем исследования остатков оценок для конкретного уравнения. Важно использовать d - статистику Durbin-Watson только тогда, когда предположения, подтверждающие наличие этого феномена имеются налицо:

а) Модель регрессии включает свободный член.

б) Существует автокорреляция первого порядка: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, где ρ коэффициент автокорреляции и u_t член классической ошибки (не автокоррелированный); $\rho = \sum_{t=2}^{n} u_t u_{t-1} / \sum_{t=2}^{n} u_{t-1}^2$.

в) Модель регрессии не должна содержать запаздывающих зависимых переменных в качестве независимых переменных. (В этом случае d - статистика смещена к значению 2, но может быть использован тест $n - Durbin - Watson$ или другие). Формула для d - статистики Durbin-Watson при n наблюдениях имеет вид: $d = \sum_{t=2}^{n} (u_t - u_{t-1})^2 / \sum_{t=1}^{n} u_t^2$, в которой u_t - остатки, полученные по М.Н.К. Отметим, что числитель имеет на одно наблюдение меньше, чем знаменатель, поскольку одно наблюдение необходимо для вычисления u_{t-1}. d - статистика равна нулю, если существует положительная предельная автокорреляция остатков; равна двум, если не существует автокорреляции остатков; равна четырем, если существует отрицательная автокорреляция остатков. Чтобы убедиться в этом, введем соответствующие остатки в уравнение:

а) $d =$ 0. Предельная положительная автокорреляция. В этом случае, $u_t = u_{t-1}$, $(u_t - u_{t-1}) = 0$ и, следовательно, d = 0, ($\rho = 1$).

б) $d \approx$ 4. Предельная отрицательная автокорреляция. В этом случае, $u_t = -u_{t-1}$ и $(u_t - u_{t-1}) = 2u_t$. Подставляя в уравнение, получим $d = \sum(2u_{t^2}) \Big/ \sum u_t^2 \Rightarrow d \approx 4$, $\rho = -1$.

в) Не существует автокорреляции: $d \approx 2$, $\rho = 0$.

9.4.2. Использование d - тест Durbin-Watson

Тест Durbin-Watson часто не используется исходя из двух предпосылок. Первая, исследователи почти никогда не тестируют гипотезу о наличии отрицательной

автокорреляции остатков, поскольку в проблемах экономики и бизнеса это очень сложно объяснять с теоретической точки зрения. Ее существование означает, что несовершенная автокорреляция остатков, вероятно, обусловлена ошибками спецификации. Вторая, d- тест или тест Durbin-Watson иногда является не убедительным. Правило предварительного принятия решения, каждый раз имеет только области “подтверждения” или «опровержения», в то время как сам d- тест или тест Durbin-Watson предоставляет третью возможность, называемую областью неубедительности. Принимая это во внимание, делаем вывод что d- тестом или тестом Durbin-Watson можно пользоваться также как и t - тестом или F - тестом.

Для тестирования автокорреляции остатков, необходимо выполнение следующих этапов:

– Из уравнения, подверженного тестированию по М.Н.К., получаем остаточную переменную и вычисляем d - статистику.
– Определяем объем выборки и количество независимых переменных, а затем консультируем статистические таблицы, для того чтобы найти соответственно максимальное критическое значение d_U и минимальное критическое значение d_L.
– Будучи объявлена гипотеза H_0 :, которая отвергает положительную автокорреляцию остатков $H_0 : \rho \le 0$ (не существует положительной автокорреляции остатков) и гипотеза $H_A : \rho \succ 0$ (существует положительная автокорреляция остатков). Самым подходящим правилом для принятия решения является:

если $d \prec d_L$ отвергается гипотеза H_0 :

если $d \succ d_U$ не отвергается гипотеза H_0 :

если $d_L \le d \le d_U$ область неопределенности.

В некоторых случаях, двусторонний тест будет более приемлемым для определения автокорреляции остатков, тогда будут выполнены первые два этапа, и не будет выполнен третий этап.

9.4.3. Пусть выполнены двусторонние альтернативные гипотезы:

$H_0 : \rho = 0$ (не существует автокорреляции остатков)

$H_A : \rho \ne 0$ (существует автокорреляция остатков)

Самым подходящим правилом для принятия решения является:

если $d \prec d_L$ отвергается гипотеза H_0 :

если $d \succ 4 - d_L$ отвергается гипотеза H_0 :

если $4 - d_U \succ d \succ d_U$ не отвергается гипотеза H_0 :,

во всех остальных случаях область неопределенности.

9.5. Оценка моделей с автокоррелированными остатками.

Наличие автокорреляции остатков первого порядка предполагает, что значение члена стохастической ошибки ε_t в момент времени t зависит от значения члена стохастической ошибки ε_{t-1} в момент времени $t-1$, следовательно, существует модель регрессии $\varepsilon_t = \beta_0 + \beta_1 \varepsilon_{t-1} + u_t$, в котором β_0, β_1 - параметры уравнения регрессии. В соответствии с формулами М.Н.К., имеем $\beta_0 = \bar{\varepsilon}_t - \beta_1 \bar{\varepsilon}_{t-1}$; $\beta_1 = \dfrac{\overline{\varepsilon_t \varepsilon_{t-1}} - \bar{\varepsilon}_t \bar{\varepsilon}_{t-1}}{\overline{\varepsilon_{t-1}^2} - \bar{\varepsilon}_{t-1}^2}$, здесь $\bar{\varepsilon}_t = \bar{\varepsilon}_{t-1} = 0$, поскольку ε_t остаточные переменные, полученные по М.Н.К. $\sum \varepsilon_t = 0 \Rightarrow \bar{\varepsilon}_t = \bar{\varepsilon}_{t-1} = 0$. Тогда имеем

$\beta_0 = 0$, а $\beta_1 = \dfrac{\overline{\varepsilon_t \varepsilon_{t-1}}}{\overline{\varepsilon_{t-1}^2}} = \dfrac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=2}^{n} \varepsilon^2{}_{t-1}} \approx \rho_1^{\varepsilon}$ - коэффициент автокорреляции остатков первого порядка. Следовательно, имеем $\varepsilon_t = \rho_1^{\varepsilon} \varepsilon_{t-1} + u_t$, u_t - член стохастической ошибки. Отметим, что $\left|\rho_1^{\varepsilon}\right| \prec 1$, учитывая последнее уравнение, получим $Y_t = \beta_0 + \beta_1 X_{1t} + \rho_1^{\varepsilon} \varepsilon_{t-1} + u_t$.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t \qquad (9.1)$$

Применим основной подход к оценке уравнения регрессии, когда имеет место автокорреляция остатков. Обратимся к модели регрессии при $t = t-1$:

$$Y_{t-1} = \beta_0 + \beta_1 X_{1t-1} + \varepsilon_{t-1}, \qquad (9.2)$$

умножим обе части уравнения на ρ_1^{ε},

$$\rho_1^{\varepsilon} Y_{t-1} = \rho_1^{\varepsilon} \beta_0 + \rho_1^{\varepsilon} \beta_1 X_{1t-1} + \rho_1^{\varepsilon} \varepsilon_{t-1}, \qquad (9.3)$$

вычтем полученное уравнение из предыдущего

$Y_t - \rho_1^{\varepsilon} Y_{t-1} = \beta_0 - \rho_1^{\varepsilon} \beta_0 + \beta_1 X_{1t} - \rho_1^{\varepsilon} \beta_1 X_{1t-1} + \varepsilon_t - \rho_1^{\varepsilon} \varepsilon_{t-1}$ (4) или $Y_t' = \beta_0' + \beta_1 X_{1t}' + u_t'$ (9.5), в котором

$$\begin{aligned} Y_t' &= Y_t - \rho_1^{\varepsilon} Y_{t-1} \\ X_t' &= X_t - \rho_1^{\varepsilon} X_{t-1} \qquad (*) \\ u_t' &= u_t - \rho_1^{\varepsilon} u_{t-1} \\ \beta_0' &= \beta_0 \left(1 - \rho_1^{\varepsilon}\right). \end{aligned}$$

Поскольку u_t' является членом стохастической ошибки не коррелированным, для оценки параметров уравнения регрессии (9.5) можно применять простой М.Н.К.

В заключение отметим, что в случае автокорреляции остатков для оценки параметров уравнения регрессии используется О.М.Н.К. (обобщенный метод наименьших квадратов). Для его использования необходимо выполнение следующих условий:

а) Преобразовать переменные Y_t и X_t к виду (*).

б) Применить М.Н.К. к уравнению (9.5) для оценки параметров β_0', β_1' .

в) Вычислить параметр $\beta_0' / \left(1 - \rho_1^{\varepsilon}\right) = \beta_0$.

г) Записать исходное уравнение.

О.М.Н.К. является аналогом М.Н.К. для конечных разностей. С той лишь разницей, что из значений переменных Y_t и X_t вычитаются не полные значения в предыдущий момент времени X_{t-1}, Y_{t-1}, а только некоторая часть из них - $\rho_1^{\varepsilon} Y_{t-1}, \rho_1^{\varepsilon} X_{t-1}$. Когда $\rho_1^{\varepsilon} = 1$, это есть просто метод конечных разностей первого порядка, поскольку $Y_t' = Y_t - Y_{t-1}; X_t' = X_t - X_{t-1}$. Следовательно, если значение d - теста или теста Durbin-Watson равно 0, применение метода конечных разностей первого порядка вполне оправдано. Если же $\rho_1^{\varepsilon} = -1$, член стохастической ошибки отрицательно коррелирован, и тогда, описанный метод, модифицируется следующим образом.

$$\begin{aligned} Y_t' &= Y_t - (-1) Y_{t-1} = Y_t + Y_{t-1} \\ X_t' &= X_t - (-1) X_{t-1} = X_t + X_{t-1} \\ \beta_0' &= \beta_0 \left(1 - (-1)\right) = 2\beta_0 \\ Y_t + Y_{t-1} &= 2\beta_0 + \beta_1 \left(X_t + X_{t-1}\right) + u_t \text{ и, как следствие,} \\ \left(Y_t + Y_{t-1}\right)/2 &= \beta_0 + \beta_1 \left(X_t + X_{t-1}\right)/2 + u_t/2 \end{aligned}$$

По существу, в последней модели определяются средние между двумя периодами для каждого ряда, а затем, для полученных данных, с помощью М.Н.К. оцениваются параметры β_0, β_1.

Основная проблема при применении этого метода заключается в определении оценки ρ_1^{ε}. Существует много методов для определения этой оценки. Однако базовый метод состоит в оценке коэффициента непосредственно из показателей, полученных для основного уравнения, т.е. $\rho_1^{\varepsilon} = 1 - d/2$, здесь d - тест Durbin-Watson.

9.6. Односторонние гипотезы

$H_0: \quad \rho \le 0$ (автокорреляция остатков отсутствует);

$H_A: \quad \rho \succ 0$ (автокорреляция остатков присутствует);

$d \prec d_L$ гипотеза H_0 отвергается;

$d_L \le d \le d_U$ область неопределенности;

$d \succ d_U$ гипотеза H_0 принимается.

9.7. Альтернативные двусторонние гипотезы

$H_0: \quad \rho = 0$ (автокорреляция остатков отсутствует);

$H_A: \quad \rho \ne 0$ (автокорреляция остатков присутствует);

$d \prec d_L$ (гипотеза H_0 отвергается);

$d \succ 4 - d_L$ (гипотеза H_0 отвергается);

$4 - d_U \succ d \succ d_U$ (гипотеза H_0 принимается) во всех остальных случаях существует область неопределенности.

10. Гетероскедастичность

Гетеросчкедастичность является следствием нарушения классической гипотезы, утверждающей, что член стохастической ошибки имеет постоянную вариацию. Предположение постоянства вариации для различного числа наблюдений члена стохастической ошибки не всегда реалистично. Например, если рассматривать модель, в которой измеряется высота баскетболиста с ошибкой в 1 inch (2.54 см) и использование той же ошибки в 1 inch при измерении высоты мышки. Вполне вероятно, что ошибка, ассоциированная с высотой баскетболиста, берет свое начало от распределения с большей вариацией, чем та, которая ассоциируется с высотой мышки. Как будет показано, необходимо различать гетероскедастичность и гомоскедастичность, поскольку *применение М.Н.К. к моделям с гетероскедастичностью не является методом с минимальной вариацией (оценки при этом остаются несмещенными).*

Гетероскедастичность часто возникает тогда, когда в наблюдаемых данных существует очень большой разброс между самыми маленькими и самыми большими значениями. Большие скачки между значениями в выборке увеличивает вероятность того, что распределение члена стохастической ошибки будет иметь большую вариацию для больших значений, в то время как распределение члена стохастической ошибки будет иметь маленькую вариацию для небольших значений наблюдений.

Легко можно получить большую разницу между самыми большими и самыми маленькими значениями переменных на множестве перекрестных данных. Напомним, что **в моделях с перекрестными данными, переменные наблюдаются в один и тот же момент времени, но для различных объектов (к примеру, личности, государства, области и др.)**. Трудн*о избежать явления гетероскедастичности в экономических*

тематиках, исследуемых перекрестным способом, поскольку модели с перекрестными данными в одном и том же примере включают наблюдения различной величины.

Тем не менее, это не означает, что гетероскедастичность не встречается в моделях с временными рядами и не исключает того, что не включенные переменные могут быть причиной гетероскедастичности для любого типа данных. Как бы то ни было, в *общем случае, появление гетероскедастичности наиболее вероятно в моделях с перекрестными данными, чем в моделях с временными рядами*.

В связи с этим, попробуем ответить на те же вопросы, касающиеся гетероскедастичности, которые были освещены для явлений мультиколлинеарности и автокорреляции остатков.

а) Какова сущность проблемы?
б) Каковы последствия этой проблемы?
в) Как диагностируется эта проблема?
г) Каковы доступные средства для устранения проблемы?

10.1. Совершенная и несовершенная гетероскедастичность

Совершенная гетероскедастичность вызывается членом стохастической ошибки *правильно специфицированного уравнения регрессии,* в то время как *несовершенная гетероскедастичность* является следствием опущенных переменных.

10.1.1.Совершенная гетероскедастичность

Совершенная гетероскедастичность является функцией зависящей от члена стохастической ошибки *корректно специфицированного уравнения регрессии.* Когда мы говорим гетероскедастичность, то подразумеваем совершенную гетероскедастичность. Такого типа гетероскедастичность возникает тогда, когда *в корректно специфицированном уравнении нарушена классическая гипотеза относительно постоянства вариаций члена стохастической ошибки:* $VAR(\varepsilon_i) = \sigma^2, (i = 1,2,\cdots,n)$

Если это предположение имеет место, то все наблюдаемые значения члена стохастической ошибки могут быть нарисованы в виде распределения с нулевым средним и вариацией σ^2, которая не меняется от одного наблюдения к другому. Это свойство носит название гомоскедастичности. Если же присутствует гетероскедастичность, то вариация члена стохастической ошибки не является постоянной, она зависит от рассматриваемого наблюдения: $VAR(\varepsilon_i) = \sigma_i^2, (i = 1,2,\cdots,n)$. Отметим, что разность между двумя формулами заключается в том, что в последнем случае мы имеем индекс "i" при σ^2, который свидетельствует о том, что вариация члена стохастической ошибки в условиях гетероскедастичности может меняться с изменением наблюдения, вместо того, чтобы оставаться постоянной для любого наблюдения.

Другой способ представления гетероскедастичности заключается в изображении множества наблюдений члена стохастической ошибки с более плоским распределением по сравнению с другими. Самая простая ситуация это та, при которой наблюдения члена стохастической ошибки могут быть сгруппированы в два различных распределения «широкое» и «узкое». Эту версию проблемы назовем дискретной гетероскедастичностью, при которой оба распределения будут центрированы вокруг нулевого значения, однако одно из них будет иметь большую вариацию, чем другое.

Гетероскедастичность, однако, принимает более сложные формы, их число практически не лимитировано, и анализ хотя бы небольшого процента представляет огромную проблему. Обсудим общие принципы гетероскедастичности, концентрируясь на наиболее часто используемых моделях *совершенной гетероскедастичности*. Это,

однако, не означает, что специалисты в эконометрике рассматривают только один тип моделей гетероскедастичности.

В модели совершенной гетероскедастичности вариация члена стохастической ошибки соотносится с экзогенной переменной Z_i. В классическом уравнении регрессии $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, вариация члена стохастической ошибки может быть равна $VAR(\varepsilon_i) = \sigma^2 Z_i^2$, *где Z_i может быть или может не быть одной из независимых переменных X_s из уравнения регрессии*. Переменная Z_i носит название фактора пропорциональности, поскольку вариация члена стохастической ошибки меняется пропорционально квадрату Z_i. Чем больше Z_i, тем больше вариация распределения наблюдения "i" члена стохастической ошибки. В зависимости от значения наблюдения, которое принимает переменная Z_i, наблюденный член стохастической ошибки будет принимать столько же различных значений, по одному для каждого из различных значений Z_i. Каким может быть фактор пропорциональности Z_i? Каким образом экзогенная переменная типа Z_i меняет в целом распределение члена стохастической ошибки?

Рассмотрим функцию, которая соотносит потребление домашних хозяйств с их доходом. Абсолютная величина вариация расходов домашних хозяйств с малым доходом, безусловно, отличается от абсолютной величины вариации расходов домашних хозяйств с большим доходом, поскольку изменение на 10% в расходах семей с большим доходом привлекает больше денег, чем изменение на 10% в расходах семей с небольшим доходом. Наряду с этим, доля бюджета семей с небольшими доходами, которые необходимо потратить на товары первой необходимости, намного больше, чем доля бюджета, используемая на те же нужды в семьях с большим доходом. В этом случае Y_i будут представлять расходы на потребление, а фактор пропорциональности Z_i будет представлять доходы домашних хозяйств. Если доходы домашних хозяйств растут, то вариация члена стохастической ошибки в уравнении будет построена таким образом, чтобы отразить расходы. Этот пример помогает объяснить тот факт, что гетероскедастичность может возникнуть в моделях с перекрестными данными, поскольку существует большой разброс в данных зависимых переменных, включенных для исследования в уравнение регрессии. К примеру, большие экзогенные отклонения, которые могут возникнуть для семей с низкими доходами, могут казаться незначительными для семей с большими доходами.

Гетероскедастичность также может возникнуть в моделях с временными рядами, по крайней мере, в двух случаях, которые отличаются от тех случаев, что описаны для моделей с перекрестными данными с большим числом вариаций в значениях зависимой переменной.

1.Гетероскедастичность может появиться в моделях с данными в форме временных рядов при наличии большого количества изменений в зависимой переменной (*темпы изменений зависимой переменной очень велики*). Если имеет место исключительный рост промышленного производства, вполне вероятно, что вариация члена стохастической ошибки возрастет теми же темпами. Как бы то ни было, такие изменения не могут иметь место во временных рядах с низкими темпами роста.

2. Гетероскедастичность может появиться в любой модели с данными в виде временных рядов либо перекрестными данными, в которых *качество собранных данных меняется катастрофически в пределах выборки*. Как только техника подготовки данных становится лучше, вариация члена стохастической ошибки будет убывать, поскольку

ошибки измерений являются частью стохастической ошибки. С уменьшением ошибок измерения, уменьшается вариация члена стохастической ошибки.

10.2. Несовершенная гетероскедастичность

Гетероскедастичность, обусловленная ошибками спецификации, как, например, опущенные переменные, представляет несовершенную гетероскедастичность. Аналогично тому, как неподходящая функциональная форма может стать источником несовершенной автокорреляции остатков, ошибки спецификации способны вызвать несовершенную гетероскедастичность, т.е. для различных ситуаций применим один и тот же подход.

Опущенная переменная может вызвать гетероскедастичность члена стохастической ошибки, поскольку часть опущенного влияния не представлено ни одной из включенных переменных, и поглощено стохастической ошибкой. Если это влияние содержит гетероскедастичную компоненту, то член стохастической ошибки в преобразованном уравнении может быть гетероскедастичным даже, если член стохастической ошибки в истинном уравнении не является таковым. Это отличие является важным, поскольку, в условиях несовершенной гетероскедастичности, правильные действия состоят в попытке определить исключенную переменную и включить ее в уравнение регрессии. *Потому очень важно* ***быть уверенным, что переменные специфицированы верно****, прежде чем выявлять либо исправлять совершенную гетероскедастичность.*

Будем считать, что проводится перекрестное исследование объемов импорта, определенного числа стран различной величины в заданном году. Для простоты, предположим, что лучшая модель импорта в перекрестной постановке для стран представлена функцией, которая положительно зависит от соответствующих ВВП и от относительных цен (которые включают влияние обменного курса) между этими странами и остальным миром. В этом случае истинная модель будет иметь следующий вид: $M_i = f(GDP_i^+, PR_i^+) = \beta_0 + \beta_1 GDP_i + \beta_2 PR_i + \varepsilon_i$.

Здесь:

M_i - импорт в ($) страны i_s;

GDP_i - валовой внутренний продует в ($) страны i_s;

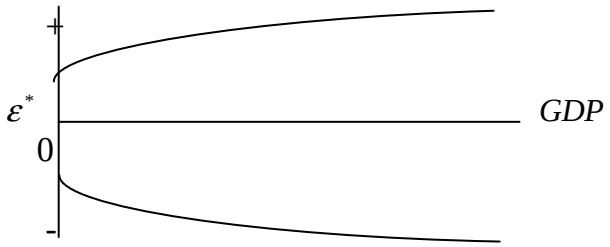
PR_i - отношение внутренней цены, реализуемых товаров (конвертированных в ($) с помощью обменного курса) к мировой цене этих же товаров, исчисленных в ($), для страны i_s;

ε_i - классический член стохастической ошибки.

Теперь предположим, что из уравнения исключается GDP и запускается процедура регрессии. Тогда уравнение примет вид: $M_i = \beta_{0i} + \beta_2 PR_i + \varepsilon_i^*$, в котором член стохастической ошибки преобразованного уравнения ε^* является функцией от GDP, исключенной переменной, и не гетероскедастичного члена стохастической ошибки ε: $\varepsilon_i^* = \varepsilon_i + \beta_1 GDP_i$.

Пока цены не воздействуют как переменные «proxy» относительно GDP, член стохастической ошибки не вбирает в себя эффект воздействия опущенной переменной. Если это новое воздействие имеет вариацию, возрастающую с ростом GDP, что вполне вероятно, новый член стохастической ошибки $\varepsilon*$ будет гетероскедастичным. Влияние такого эффекта также зависит от величины $\beta_1 GDP_i$, компоненты, сравнимой по абсолютной величине с классической компонентой ε_i. Чем больше доля опущенной переменной в ε_i^*, тем наличие гетероскедастичности более вероятно. Наблюденные

значения члена стохастической ошибки ε_i^* для такого случая представлены на следующем рисунке



Как видно из рисунка, большим значениям GDP соответствуют и большие значения вариаций члена стохастической ошибки.

10.3.. Последствия гетероскедастичности

Если факт наличия гетероскедастичности члена стохастической ошибки установлен, тогда как это сказывается на оценках коэффициентов? Последствия гетероскедастичности, по большому счету, почти идентичны тем, что вызываются автокорреляцией остатков, за исключением двух проблем, которые полностью отличаются.

а) *Совершенная гетероскедастичность не вызывает смещений в оценке коэффициентов*. Даже, если член стохастической ошибки в уравнении регрессии подвержен совершенной гетероскедастичности, то это не может вызвать смещений в оценке коэффициентов, вычисленных по М.Н.К. И, как следствие, уравнение регрессии, подверженное совершенной гетероскедастичности, обладает следующими свойствами: $E(\hat{\beta}_s) = \beta_s, \forall\, s$. Однако отсутствие смещений не может гарантировать качество оценок коэффициентов, в частности, из-за того, что гетероскедастичность увеличивает вариации оценок, но их распределение все-таки центрировано вокруг истинных значений коэффициентов. Уравнение, подверженное несовершенной гетероскедастичности, обусловленной исключенной переменной, конечно же, будет иметь некоторое смещение вследствие изменения спецификации уравнения регрессии.

б) *Гетероскедастичность увеличивает вариации распределений* β_s.

Совершенная гетероскедастичность не вызывает смещений оценок, полученных по М.Н.К., но может повлиять на свойство минимальной вариации. Если член стохастической ошибки в уравнении регрессии гетероскедастичен в соответствии с членом пропорциональности Z: $VAR(\varepsilon_i) = \sigma^2 Z_i^2$, тогда вариация $\hat{\beta}_s$ является функцией от Z: $VAR^{**}(\hat{\beta}_s) = f(Z^2) \cdot \left[VAR(\hat{\beta}_s)\right]$, здесь $VAR^{**}(\hat{\beta}_s)$ - вариация с гетероскедастичностью; $f(Z^2)$ - положительная функция от фактора пропорциональности Z, который является причиной гетероскедастичности в уравнении; $\left[VAR(\hat{\beta}_s)\right]$ - вариация без гетероскедастичности. А потому, не является неожиданностью, что утверждение о минимальной вариации не может быть доказано, если не выполнена классическая гипотеза относительно наличия гетероскедастичности.

в) *Наличие гетероскедастичности приводит к тому, что М.Н.К. недооценивает вариации (и стандартные ошибки) коэффициентов*.

Гетероскедастичность вызывает прирост вариаций $\hat{\beta}_s$ в завуалированном виде, который не улавливается М.Н.К., а потому почти всегда М.Н.К. недооценивает эти вариации. *Следовательно, ни t - статистики, ни F - статистика не являются надежными при*

наличии гетероскедастичности. В действительности, как правило, применение М.Н.К. приводит к большим значениям t - статистик, если наблюдаемые значения члена стохастической ошибки гетероскедастичны, иногда заставляя исследователя опровергнуть нулевую гипотезу, которая в действительности не должна быть отвергнута.

Гетероскедастичность вызывает специфический сегмент последствий, поскольку Z и распределение члена стохастической ошибки возрастают таким образом, что увеличивают вероятность появления больших по абсолютной величине значений члена стохастической ошибки. Если случайно этот сегмент значений положителен, когда значения одной из независимых переменных значительно превышают среднее значение, то оценки $\hat{\beta}_s$ этой переменной, полученные по М.Н.К., имеют тенденцию к увеличению по сравнению с теми, что получены раннее. Если же, сегмент этих больших наблюдаемых значений члена стохастической ошибки отрицателен, когда значения одной из независимых переменных X_s значительно ниже среднего значения, тогда оценки $\hat{\beta}_s$ этой переменной, полученные по М.Н.К., имеют тенденцию к уменьшению по сравнению с ожидаемыми значениями. Поскольку предполагается, что наблюдаемые значения члена стохастической ошибки не коррелированны с наблюдаемыми значениями независимых переменных, переоценки коэффициентов, полученных по М.Н.К., равно вероятны, как и их недооценки, в то время как, при наличии гетероскедастичности, сами оценки остаются не смещенными. Как бы то ни было, *гетероскедастичность способствует тому, что* $\hat{\beta}_s$ *отдаляются от истинных значений и, следовательно, вариация распределения* $\hat{\beta}_s$ *возрастает*.

10.4. Тестирование гетероскедастичности

Для выявления гетероскедастичности используются различные тесты, поскольку это явление принимает различные формы и ее точное проявление в уравнении регрессии каждый раз неизвестно. Использование фактора пропорциональности Z является только одним из множества подходов, используемых для спецификации гетероскедастичности. А потому, не существует единого подхода к тестированию гетероскедастичности. Из всего множества подходов рассмотрим только четыре. Сначала рассмотрим тест Park.

10.4.1. Тест Park

Пусть $VAR(\varepsilon_i) = \sigma^2 Z_i^2$, где ε_i член стохастической ошибки исследуемого уравнения регрессии; σ^2 - величина вариации члена стохастической ошибки гомоскедастичного уравнения регрессии; Z - фактор пропорциональности. Тест Park является формальной процедурой, которая пытается оценить остатки на наличие в них гетероскедастичности в таком же ключе, в котором d - статистика $Durbin-Watson$ тестирует остатки, с целью выявления автокорреляции. Тест Park содержит три основных этапа. На первом этапе уравнение регрессии оценивается с помощью М.Н.К.. На втором этапе используется процедура логарифмирования квадратов остатков, которые будут представлять зависимую переменную в новом уравнении регрессии с единственной независимой переменной-фактором пропорциональности Z. И на третьем этапе результаты, полученные после запуска вспомогательной регрессии, тестируются на наличие гетероскедастичности. Как бы то ни было, нет необходимости запускать тест Park для каждого оцененного уравнения регрессии. Прежде чем использовать тест Park, было бы неплохо поставить следующие вопросы и дать на них ответы.

а) Существуют ли явные ошибки спецификации? Если есть подозрение, что в оцененном уравнении были опущены переменные или регрессия перезапускалась из-за последствий

спецификации, проверка теста Park отодвигается до тех пор, пока спецификация достаточно хороша.
б) Может ли природа предмета исследования быть подвержена гетероскедастичности? Модели с перекрестными данными являются наиболее подозреваемыми на наличие гетероскедастичности (например, с большими вариациями в значениях зависимых переменных) нежели другие.
в) И, наконец, предоставляет ли графическое исследование какие-либо доказательства в пользу гетероскедастичности? Иногда можно сэкономить много времени, если построить график зависимости остатков от фактора пропорциональности Z. Анализируя график можно сделать вывод о наличии гетероскедастичности без того, чтобы проверять тест Park.

Если существуют какие-то подозрения на гетероскедастичность лучше всего проверить тест Park. А поскольку он не запускается автоматически компьютерными программами по регрессии, то надо знать, как запускать этот тест самостоятельно.
Первый этап: Вычисляются остатки оцененного уравнения регрессии. $u_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$.
Второй этап: Вычисленные остатки используются при формировании зависимой переменной для второго уравнения регрессии. А именно, предлагается запустить следующее уравнение регрессии в полных логарифмах: $\ln(u_i^2) = \alpha_0 + \alpha_1 \ln(Z_i) + v_i$, в котором u_i остатки из первого уравнения регрессии; Z_i - наилучший выбор для фактора пропорциональности Z; v_i - классический член стохастической ошибки (гомоскедастичный).
Третий этап: С помощью теста Стьюдента - t проверяется значимость коэффициента при переменной $\ln(Z)$, которая объясняет поведение зависимой переменной $\ln(u_i^2)$ в рассматриваемом уравнении. Если коэффициент при переменной $\ln(Z_i)$ является значимым, т. е. существенно отличается от нуля, то это свидетельствует о том, что существует сегмент гетероскедастичности в остатках относительно Z. В противном случае, гетероскедастичность, соотнесенная с этим частным Z, не подтверждается. Как бы то ни было, можно доказать, что специфицированный член стохастической ошибки в рассматриваемом уравнении регрессии является гомоскедастичным.

Тест Park непросто использовать каждый раз при возникновении необходимости. Самая большая проблема заключается в идентификации фактора пропорциональности Z. Хотя, в большинстве случаев, Z является независимой переменной в исходном уравнении регрессии, этот факт не гарантирован для любого оригинального уравнения регрессии. Частный вид фактора пропорциональности Z может быть найден только в результате исследования уравнения регрессии на гетероскедастичность. Хороший фактор пропорциональности, по всей вероятности, изменяется в том же ритме, что и вариация члена стохастической ошибки.

Например, в модели с перекрестными данными для различных стран хорошим фактором пропорциональности Z будет тот, который измеряет объем выборки в зависимости от исследуемой зависимой переменной.

11. Средства против гетероскедастичности

Будут представлены несколько способов борьбы с гетероскедастичностью, которые в то же время продемонстрируют, что существуют ситуации, для которых проблема не может быть подогнана под определенную процедуру. Искусство эконометрики и состоит в умении различать одну ситуацию от другой.

Первый шаг на пути освобождения уравнения от гетероскедастичности состоит в попытке выяснения, какого типа гетероскедастичность присутствует в уравнении регрессии: совершенная или несовершенная. Если окажется, что гетероскедастичность является несовершенной, то выявляются важные переменные, которые были исключены из уравнения регрессии, и они вновь включаются в него. Если же гетероскедастичность - совершенная, то рассматриваются два обобщенных средства.

11.1. Использование взвешенного метода наименьших квадратов

Если же гетероскедастичность является совершенной, то рассматривается взвешенный метод наименьших квадратов. Каждый член исследуемого уравнения регрессии делится на фактор пропорциональности Z (либо на функцию, зависящую от фактора пропорциональности Z). Вновь полученное уравнение регрессии, которое представлено преобразованными зависимой и независимой переменными, оценивается.

11.2. Переопределение переменных

Эффект гетероскедастичности остатков во многих случаях может быть исключен путем переопределения переменных. Этот метод прямой в отличие от взвешенного метода наименьших квадратов, который является косвенным. Переопределение переменных будет основываться на соответствующей теории и на смещении центра уравнения регрессии в зависимости от теоретического базового представления, которое должно быть обоснованным. Следовательно, первое, что необходимо предпринять, если тест Park указывает на возможность наличия гетероскедастичности в исследуемом уравнении регрессии, это тщательно исследовать уравнение на наличие ошибок спецификации. Тот факт, что тест Park указывает на возможность наличия гетероскедастичности, ни в коей мере не означает включение новой переменной в уравнение регрессии без того, чтобы внимательно изучить спецификации уравнения. Если в результате анализа спецификаций уравнения регрессии, сделан вывод о необходимости включения в него дополнительной независимой переменной, тогда эта переменная будет включена в уравнение регрессии.

11.3. Взвешенный метод наименьших квадратов

Рассмотрим уравнение с совершенной гетероскедастичностью, обусловленной фактором пропорциональности Z: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, (11.1)

$VAR(\varepsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2$ (11.2),

здесь Z_i - фактор пропорциональности, σ^2 - постоянная вариация члена классической стохастической ошибки (гомоскедастичной) ε_i. Поскольку существует совершенная гетероскедастичность, то исходное уравнение регрессии может быть определено как

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + Z_i u_i$. (11.3)

Член стохастической ошибки $Z_i u_i$ в полученном уравнении гетероскедастичен, поскольку вариация $\sigma^2 Z_i^2$ не является постоянной. Самый простой подход к преобразованию специфицированного уравнения в гомоскедастичное, заключается в делении каждого члена уравнения регрессии на фактор пропорциональности Z_i. В результате чего получается новый член стохастической ошибки u_i с постоянной вариацией σ^2 и новое уравнение регрессии, которое удовлетворяет всем классическим гипотезам, а его запуск подтверждает факт исключения гетероскедастичности. Этот метод носит название взвешенного метода наименьших квадратов (В.М.Н.К.), и он является одной из модификаций классического метода наименьших квадратов.

В.М.Н.К. влечет за собой деление каждого члена исследуемого уравнения регрессии на любую переменную, которая преобразовала бы член стохастической ошибки в гомоскедастичный, а затем запуск преобразованного уравнения регрессии. Если известно, что исследуемое уравнение регрессии подвержено совершенной гетероскедастичности, тогда метод состоит из трех этапов.

а) каждый член уравнения (11.3) делится на фактор пропорциональности Z_i,

$$Y_i / Z_i = \beta_0 / Z_i + \beta_1 X_{1i} / Z_i + \beta_2 X_i / Z_i + u_i \qquad (11.4)$$

член стохастической ошибки в этом уравнении уже гомоскедастичен,

б) пересчитываются значения переменных в соответствии с преобразованным уравнением,

в) с помощью М.Н.К. оценивается преобразованное уравнение.

Оценки, полученные на третьем этапе с помощью взвешенного метода наименьших квадратов, обманчивы, поскольку для получения нового уравнения регрессии важно, является ли фактор пропорциональности независимой переменной в исходном уравнении регрессии. Если фактор пропорциональности не является независимой переменной в исходном уравнении регрессии, тогда уравнение регрессии на третьем этапе примет вид:

$$Y_i / Z_i = \beta_0 / Z_i + \beta_1 X_{1i} / Z_i + \beta_2 X_i / Z_i + u_i . \qquad (11.5)$$

В любом случае, отметим, что в полученном уравнении член стохастической ошибки гомоскедастичен.

Как было замечено ранее, опущенный свободный член может спровоцировать эффект опущенной постоянной переменной, нелинейность и ошибки измерений по отношению к другим оцененным коэффициентам. Для того чтобы избежать ситуации, при которой постоянный член провоцирует изменения в оцененных коэффициентах, используется альтернативный подход, который предполагает, что в исходное уравнение регрессии добавляется свободный член до того, как уравнение будет оценено. Следовательно, когда в исходном уравнении регрессии Z не идентифицируется ни с одной из независимых переменных X_s, тогда на третьем этапе с помощью В.М.Н.К запускается следующая спецификация

$$Y_i / Z_i = \alpha_0 + \beta_0 / Z_i + \beta_1 X_{1i} / Z_i + \beta_2 X_i / Z_i + u_i . \qquad (11.6)$$

Если Z независимая переменная в исходном уравнении регрессии, тогда нет необходимости добавлять свободный член в уравнение, поскольку в нем уже есть свободный член. Вернемся опять к уравнению (11.4), если $Z = X_1$ (или $Z = X_2$), тогда один из угловых коэффициентов становится свободным членом в уравнении регрессии, т.к. $X_1 / Z = 1$. $Y_i / Z_i = \beta_0 / Z_i + \beta_1 + \beta_2 X_i / Z_i + u_i$ (11.7)

Однако, несмотря на то, что используется взвешенная модификация М.Н.К., оцененные коэффициенты уравнения (11.7) следует интерпретировать очень осторожно. Отметим, что коэффициент $\hat{\beta}_1$ является свободным членом в уравнении (11.7) хотя он представляет угловой коэффициент при переменной X_1 в уравнении (11.1). В данном случае в уравнении (11.7) будет исследоваться коэффициент $1/Z_i$, а при необходимости и оценка свободного члена в уравнении (11.5). Компьютер афиширует коэффициент $\hat{\beta}_0$ в качестве «углового коэффициента» и $\hat{\beta}_1$ в качестве «свободного члена 2, в то время как в действительности, в уравнении (11.1) оценены противоположные коэффициенты.

Возникают три серьезные проблемы при использовании В.М.Н.К.

1. Задача идентификации фактора пропорциональности является очень сложной.

2. Функциональная форма, которая соотносит фактор пропорциональности с вариацией свободного члена в исходном уравнении, может не быть линейной (предположим

квадратичной), а когда исследуются другие функциональные формы, необходимы другие преобразования переменных.
3. Иногда В.М.Н.К. применяется для уравнения с несовершенной гетероскедастичностью. В этом случае можно доказать, что полученные оценки подвержены небольшим уменьшениям в смещениях, если опущена одна из независимых переменных, и оценки больше тех, которые получены для корректно специфицированного уравнения.

11.4. Прямой подход: переопределение переменных

Другой подход, применяемый для исключения гетероскедастичности, состоит в пересмотре основной теории и переопределении независимых переменных в исходном уравнении для того, чтобы ликвидировать гетероскедастичность. Переопределение переменных часто является полезным, поскольку позволяет сконцентрироваться на поведенческом аспекте взаимосвязи в исходном уравнении. Переопределение переменных является трудным и обескураживающим процессом, поскольку отвергает всю выполненную работу. Тем не менее, как только теоретическая работа выполнена, альтернативные подходы предоставляют возможные пути для того, чтобы избежать проблемы, которые казались непреодолимыми в начале. К сожалению, трудно указать процедуры более обобщенные, чем «полный пересмотр исследуемого проекта». Представим нечисловой пример по этому поводу.

Рассмотрим модель с перекрестными данными для общих затрат правительств разных городов. При таком анализе логично использовать при исследовании следующие переменные: агрегированный доход; численность населения и среднюю зарплату в каждом городе. При этом, чем больше будут доходы резидентов и бизнесменов города, тем больше будут и расходы правительства этого города. В данном случае недостаточно знать, что большие города имеют большие расходы (в абсолютных величинах) по сравнению с малыми городами. Аппроксимация этой зависимости с помощью уравнения регрессии также отражает утрированный вес больших городов и соответственно большую долю квадратов остатков. Это так, поскольку М.Н.К. минимизирует сумму квадратов остатков, а т.к. остатки для больших городов, с большой вероятностью, могут быть большими попросту из-за величины городов, оценки будут особенно чувствительны к остаткам больших городов. Этот феномен часто называют «ложной корреляцией», обусловленной размерностью.

Плюс к этому, остатки могут указывать на гетероскедастичность. Средство против такого рода гетероскедастичности заключается не в автоматическом использовании взвешенного М.Н.К. и не в исключении наблюдений относящихся к большим городам. Имеет смысл рассмотреть новую формулировку модели, которая обеспечит нормированное снижение фактора (величина города) и подчеркнет соответствующее поведение. В этом случае затраты на одного жителя будут представлять логически обоснованную зависимую переменную и доход на одного жителя будет логически обоснованной независимой переменной. Далее представим такое преобразование. Новая форма преобразованного уравнения ставит большие города на ту же ступень, что и маленькие города, и, таким образом предоставляет им тот же вес при определении оценок. Если независимая переменная не является функцией от величины города, она не будет подвержена преобразованию. К примеру, если в уравнение входит переменная, отражающая среднюю зарплату, то она не будет подлежать преобразованию. Отметим, что это преобразование, в некотором смысле, подобно взвешенному М.Н.К. Отличие состоит в том, что не существует ни одного члена в уравнении обратного численности населения (как в М.Н.К.) и не все независимые переменные делятся на численность населения. Из исходного уравнения,

$$EXP_i = \beta_0 + \beta_1 POP_i + \beta_2 INC_i + \beta_3 WAGE_i + \varepsilon_i . \quad (11.8)$$

Версия В.М.Н.К. будет следующей:

$$EXP_i / POP_i = \beta_1 + \beta_0 / POP_i + \beta_2 INC_i / POP_i + \beta_3 WAGE_i / POP_i + u_i , \quad (11.9)$$

в то же время, непосредственно преобразованное уравнение принимает вид:

$$EXP_i / POP_i = \alpha_0 + \alpha_1 INC_i / POP_i + \alpha_2 WAGE_i + u_i . \quad (11.10)$$

Как нетрудно заметить, при использовании В.М.Н.К. уравнение (11.9) полностью делится на независимую переменную - население, в то время как, в преобразованном в соответствии с теорией уравнении, делятся на население только переменные дохода и расходов. Непосредственно преобразованное уравнение (11.10) действительно решает проблему возможной гетероскедастичности в модели. Такое решение будет рассматриваться как случайное, для того чтобы вновь обратиться к уравнению регрессии с целью концентрации на принципиальном исследовании его поведения.

Заметим, что преобразованное уравнение (11.10) может быть подвержено гетероскедастичности, поскольку вариация остаточного члена может быть больше для наблюдений с большими значениями на душу населения для затрат и доходов чем для тех наблюдений, которые имеют меньшие значения на душу населения для тех же показателей. Поэтому даже в преобразованном уравнении необходимо обосновать тестирование гетероскедастичности. И все-таки, гетероскедастичность в преобразованном уравнении не является правдоподобной, поскольку речь будет идти о небольшой вариации для значений нормально ассоциированных с гетероскедастичностью.

Аккуратное преобразование переменных с целью исправления гетероскедастичности, не ликвидируя «ложной корреляции», обусловленной большими значениями, все-таки иногда может представлять удачный подход в решении этих проблем. Отметим, что не каждая переменная в уравнении трактуется одинаково (в отличие от М.Н.К.). Каждая переменная в модели с перекрестными данными может быть рассмотрена с позиции возможного преобразования, результатом которого является значимое и полное представление уравнения регрессии.

12. Подбор функциональной зависимости для уравнения регрессии

12.1 Спецификация уравнения регрессии: выбор функциональной формы

Выбор функциональной формы для уравнения регрессии является составной частью в процессе спецификации уравнения регрессии. Выбор функциональной формы, в большинстве случаев, будет обусловлен обоснованной экономической теорией или обоснованной теорией соответствующей деловой активности, и, только в редких случаях, качеством прогноза, который реализует функциональная форма.

В парной регрессии выбор вида математической функции $Y = f(X)$ может быть осуществлен тремя способами:

- графическим;
- аналитическим, т.е. исходя из теории изучаемой взаимосвязи;
- экспериментальным.

При изучении зависимости между двумя признаками, графический метод подбора вида уравнения регрессии достаточно нагляден. Он основан на поле корреляции.

Значительный интерес представляет аналитический метод выбора типа уравнения регрессии, который основан на изучении материальной природы связи исследуемых признаков.

При обработке информации на компьютере, выбор вида уравнения регрессии обычно осуществляется экспериментальным путем, т.е. путем сравнения остаточной дисперсии, рассчитанной по разным моделям.

Если уравнение регрессии проходит через все точки корреляционного поля, что возможно только при функциональной связи, когда все точки лежат на линии регрессии $\widehat{Y} = f(X)$, т.е. фактические значения результирующего признака, совпадают с теоретическим, они полностью обусловлены влиянием фактора X, а $\widehat{\sigma}_u^2 = 0$.

12.2 Альтернативные функциональные формы

Выбор функциональной формы для уравнения регрессии является необходимым элементом при спецификации уравнения регрессии. При использовании М.Н.К. необходимо, чтобы уравнение регрессии было линейным относительно коэффициентов, однако, известно много функциональных форм, которые, будучи линейными относительно коэффициентов, не являются линейными относительно переменных и могут быть использованы в качестве уравнения регрессии. Приведем детальное описание наиболее часто используемых функциональных форм для того, чтобы иметь возможность правильного подобрать одну из них при спецификации уравнения регрессии.

Как правило, основанием для выбора функциональной формы будет обоснованная экономическая теория или подтвержденная теория деловой активности, и только, в редких случаях, будет выбрана та функциональная форма, которая предоставляет лучший прогноз. Исследование логической взаимосвязи между зависимой и независимыми переменными на основе разных функциональных форм, определит выбор той, которая в большей степени, соответствует экономической теории.

Далее будут рассмотрены наиболее часто употребляемые функциональные формы: графики, уравнения, примеры для того, чтобы иметь возможность сравнить их.

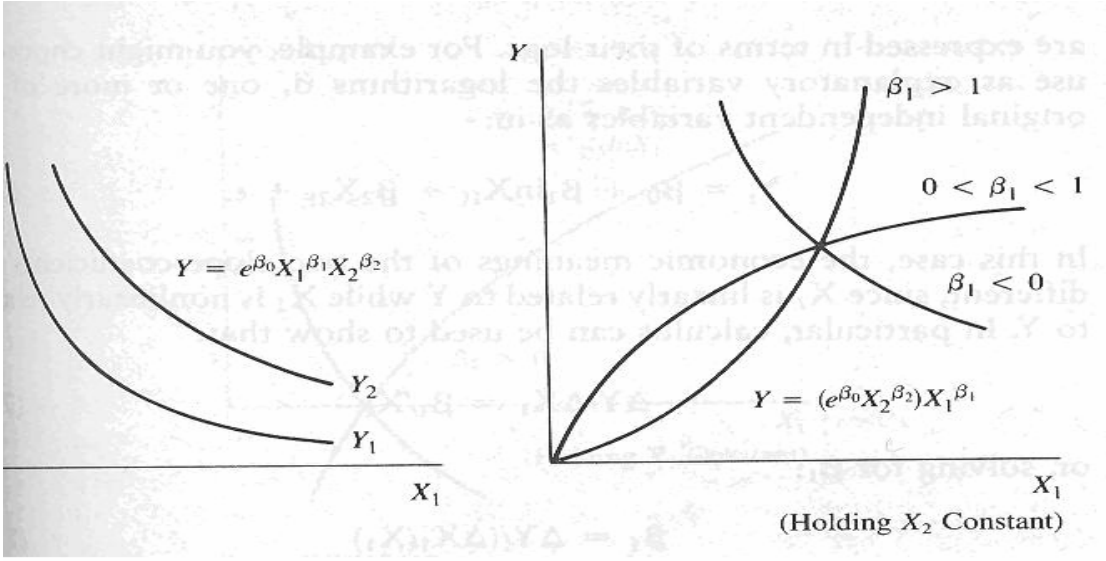
12.2.1 Линейная форма

Линейная модель регрессии предполагает постоянство угловых коэффициентов ($\frac{\partial Y}{\partial X_i} = \frac{\Delta Y}{\Delta X_i} = \beta_i$, $i = 1,2,...,K$), которые характеризуют взаимосвязь между зависимой переменной и независимыми переменными. Угловой коэффициент является постоянным, а эластичности зависимой переменной Y по отношению к независимым переменным X_i (изменение в процентах зависимой переменной, обусловленное изменением на один процент одной из независимых переменных, в то время как остальные переменные не меняются, остаются постоянными) не является постоянными: $E_{Y,X_i} = \frac{\partial Y}{\partial X_i}\frac{X_i}{Y} = \frac{\partial Y}{\partial X_i} / \frac{Y}{X_i} = \beta_i \frac{X_i}{Y}$. Если предполагаемая зависимость между определяемой переменной Y и независимыми переменными X_i такова, что угловой коэффициент предположительно постоянный, тогда может быть использована линейная форма.

К сожалению, очень часто из теоретических соображений можно предсказать лишь знак, сопутствующий той или иной переменной, но не вид функциональной зависимости. Когда теоретическое обоснование не достаточно для определения вида функциональной зависимости, может быть использована линейная зависимость до тех пор, пока не будут получены достаточные теоретические аргументы в пользу другой зависимости. Исключение составляют те случаи, когда теория, здравый смысл, опыт свидетельствуют о том, что использование какой либо другой функциональной формы не оправдано. Поскольку линейная функциональная зависимость используется по умолчанию, то на нее ссылаются как на неявную функциональную форму.

12.2.2 Полная логарифмическая функциональная форма

Наиболее часто используемой функциональной формой (нелинейной по переменным, но линейной по коэффициентам) является полная логарифмическая форма. Полная логарифмическая зависимость наиболее часто используется для спецификации уравнения регрессии, поскольку, в противоположность линейной модели, в этой модели эластичности, а не угловые коэффициенты, являются постоянными. А это означает, что $E_{Y/X_i} = \beta_i = const$. Согласно гипотезе о постоянстве эластичностей, соответствующая функциональная примет вид: $Y = e^{\beta_0} X_1^{\beta_1} X_2^{\beta_2} e^{\varepsilon}$.



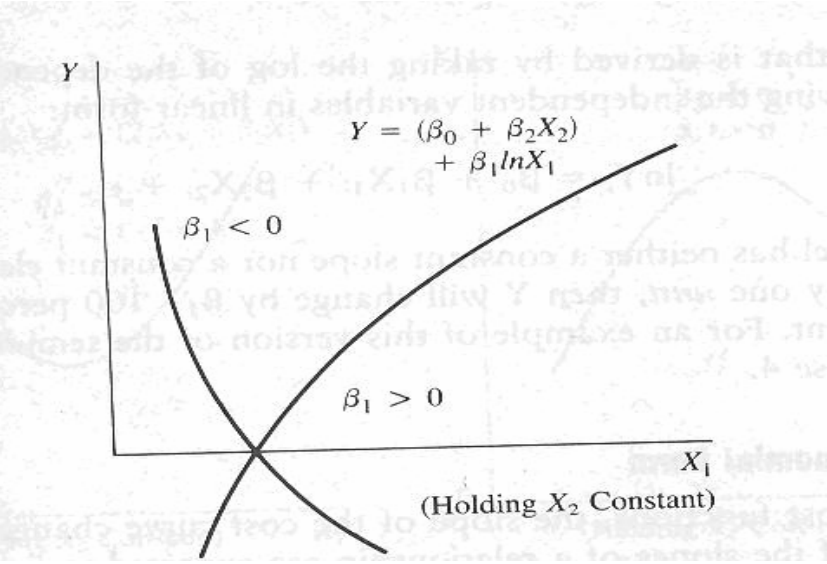
Прологарифмировав обе части уравнения, получим уравнение, линейное относительно коэффициентов, которое известно под названием полной логарифмической формы $\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon$, здесь $\ln Y$ натуральный логарифм Y. В полном логарифмическом уравнении, частные коэффициенты регрессии, к примеру β_k, могут быть интерпретированы как эластичности, поскольку:

$\beta_k = \dfrac{\partial \ln Y}{\partial \ln X_k} = \dfrac{\Delta Y / Y}{\Delta X_k / X_k} = E_{Y,X_k}$. Параметры β_k в полном логарифмическом уравнении имеют следующий смысл. Если значение независимой переменной X_k меняется на один процент, в то время как значения остальных независимых переменных остаются постоянными, значение зависимой переменной Y изменится на β_k%. На рисунке, приведенном выше, дано экономическое представление производственной функции (кривые безразличия). Изокванты производственной функции демонстрируют различные комбинации факторов производства (возможно, капитала и труда), которые могут быть использованы для фабрикации определенного объема продукции. Такого типа производственные функции называются производственными функциями типа Cobb-Douglas. Рисунок слева демонстрирует зависимость между переменной Y и переменной X_1, когда независимая переменная X_2 не меняется либо не была включена в модель. Необходимо заметить, что наклон кривой зависит от знака и величины коэффициента β_1.

Прежде чем использовать полную логарифмическую модель, необходимо проверить наблюдаемые значения для зависимой переменной и независимых переменных на наличие нулевых значений. Полная логарифмическая модель может быть использована только, если все значения переменных отличны от нуля.

12.2.3 Полулогарифмическая форма

Полулогарифмическая форма является разновидностью полной логарифмической формы. В ней только некоторые переменные (зависимая и/или независимые) выражены в логарифмических терминах. Рассмотрим полулогарифмическую форму следующего вида: $Y_i = \beta_o + \beta_1 \ln X_{1i} + \beta_2 X_{2i} + \varepsilon_i$. В этом уравнении два угловых коэффициента, и экономический смысл каждого из них разный. А именно, независимая переменная X_2 находится в линейной зависимости относительно Y, в то время как независимая переменная X_1 не линейно зависит от переменной Y. Можно установить, что $\frac{\Delta Y}{\Delta X_1} = \beta_1 / X_1$ или $\beta_1 = \Delta Y / (\Delta X_1 / X_1)$. Другими словами, если переменная X_1 меняется на 1%, то переменная Y меняется на $\beta_1 / 100$ (напомним, что значение X_1 должно быть положительным, чтобы можно было проводить операцию логарифмирования). $E_{Y/X_1} = \frac{\Delta Y}{\Delta X_1}\frac{X_1}{Y} = \frac{\beta_1}{Y}$ есть эластичность переменной Y относительно переменной X_1, и она убывает с ростом Y. На следующем рисунке изображена зависимость между Y и X_1 при условии постоянства X_2. Отметим, что для $\beta_1 \succ 0$ влияние на Y изменений в X_1 убывает с ростом X_1. Следовательно, полулогарифмическая форма будет использована тогда, когда предполагаемая зависимость между переменной Y и переменной X_1 является убывающей.



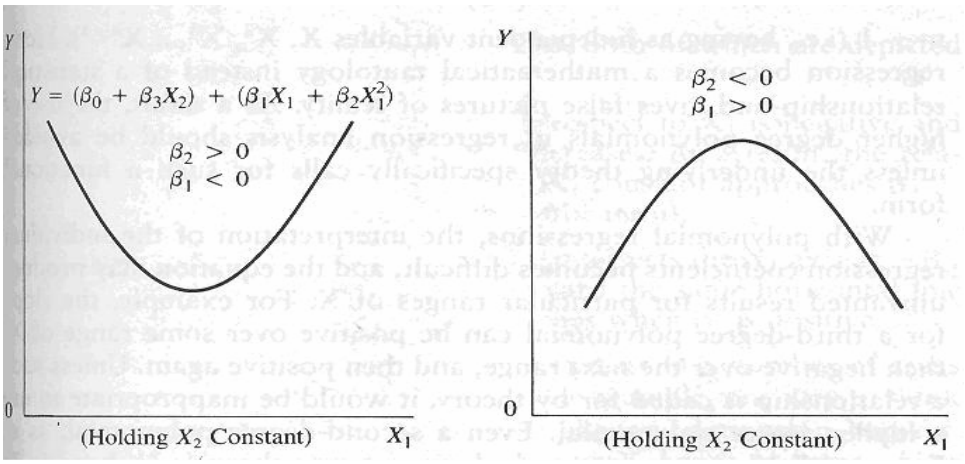
В экономике и реальном бизнесе полулогарифмическая функция встречается довольно часто. Например, большинство функций спроса при достижении определенного уровня дохода убывают с определенными темпами. Эти кривые Энгеля при существенном росте дохода имеют тенденцию к выпрямлению. В такой ситуации только небольшой процент дохода идет на потребление, а большая его часть сберегается, т.е. потребление растет убывающими темпами. Если Y описывает потребление определенного блага, а X_1 - располагаемый доход, (переменная X_2 будучи сохранена в уравнении вместо остальных независимых переменных), тогда использование полулогарифмической формы будет оправдано всякий раз, когда предполагается, что при росте располагаемого дохода темпы потребления определенного блага убывают.

Приведем пример, когда функциональная форма для переменной Y имеет представление отличное от остального уравнения и является модификацией

полулогарифмической функции, которая получается при логарифмировании зависимой переменной, при этом независимые переменные сохраняют линейное представление: $\ln Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$. В этой модели, как угловой коэффициент, так и эластичности не являются постоянными. При изменении X_1 на одну единицу, Y изменится на $\beta_1 * 100\%$, при этом X_2 остается постоянной.

12.2.4 Полиномиальная функциональная форма

Для большинства функций издержек их угловые коэффициенты меняются теми же темпами, что и объемы производства. Если предположить, что угловой коэффициент будет зависеть от величины вариации (например, становится круче с ростом объема производства), то правомерно рассматривать полиномиальную модель регрессии. Полиномиальная функциональная форма представляет Y как функцию от независимых переменных, некоторые из которых возведены в степень большую, чем единица. Например, полином второго порядка, называется квадратичным, если, по крайней мере, одна из независимых переменных возводится в квадрат: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_3 X_{2i} + \varepsilon_i$. Действительно, в такой модели, угловой коэффициент меняется одновременно с изменением независимой переменной. В рассматриваемом уравнении угловой коэффициент переменной Y по переменной X_1 имеет вид: $\frac{\Delta Y}{\Delta X_1} = \beta_1 + 2\beta_2 X_1$, а $\frac{\Delta Y}{\Delta X_2} = \beta_3$. Заметим, что первый угловой коэффициент зависит от значений переменной X_1, в то время как второй угловой коэффициент является постоянным. Если рассматривается функция стоимости, в которой переменная Y отражает среднюю цену продукции, а переменная X_1 представляет объем производства определенного блага, тогда вполне возможно, что β_1 будет отрицательным, а β_2 будет положительным, если функция стоимости имеет седловую точку.



В качестве другого примера рассмотрим модель доходов работников предприятия в зависимости от возраста каждого занятого и от ряда других показателей, которые стимулируют производительность, как, например, образование. Каким будет ожидаемое влияние возраста на заработок? С увеличением возраста, во многих случаях, заработки немолодых людей, вообще говоря, не имеют тенденции к росту, а с приближением пенсионного возраста, ожидается, что заработки вообще будут уменьшаться. И, как следствие, зависимость между величиной заработка и возрастом может быть представлена в виде графика, приведенного на верхнем рисунке в правой его части. Заработки будут расти до определенного уровня, а затем с возрастом будут убывать.

Такую теоретическую зависимость можно смоделировать с помощью квадратного многочлена: $Z_i = \beta_0 + \beta_1 V_i + \beta_2 V_i^2 + \cdots + \varepsilon_i$, какими должны быть ожидаемые знаки $\hat{\beta}_1, \hat{\beta}_2$? Чем старее работающий, тем разница между переменной V и ее квадратом V^2 будет расти катастрофически. И как следствие, коэффициент при переменной V будет более важным для молодых работников, чем для пожилых работников. И, наоборот, коэффициент при переменной V^2 будет иметь большее значение для работников пожилых. Поскольку ожидается, что влияние возраста сначала способствует росту заработка, а потом ведет к его уменьшению, то в этом случае можно предположить, что коэффициент $\hat{\beta}_1$ будет положительным, а коэффициент $\hat{\beta}_2$ будет отрицательным. Фактически, это является подтверждением того, что было теоретически обосновано многими исследователями в области экономики труда.

К сожалению, не все полиномы могут быть использованы в качестве кривых прогнозирования. В действительности, любые n наблюдений могут быть точно аппроксимированы, (все остаточные члены будут равны нулю) с помощью уравнения регрессии полиномиального вида степени $(n-1)$, имея в качестве независимых переменных $X, X^2, X^3, \cdots, X^{n-1}$. В этом случае регрессия превращается в математическое тождество, а не статистическую зависимость и предоставляет ложную картину исследуемой реальности. Необходимо избегать использования полиномов высокого порядка в регрессионном анализе, пока для такого рода функциональных форм не будет разработана соответствующая теория.

В полиномиальной регрессии истолкование частных коэффициентов становится затруднительным и уравнение может произвести не желаемый эффект для специальных областей изменения независимых переменных X. Например, коэффициент при полиноме третьего порядка может быть положительным для одной области значений X, а затем принимать отрицательные значения для последующей области значений, а затем опять принимать положительные значения. Такого рода взаимосвязи требуют специальных теоретических разработок, а потому использования полиномов высокого порядка в регрессионном анализе является не в полной мере обоснованным. Даже полином второго порядка в рассмотренном примере, предписывает симметричный угловой коэффициент, который в некоторых случаях может не быть разумным. Поэтому, каждый раз при использовании полиномиального уравнения регрессии, следует быть очень осторожным. В первую очередь необходимо убедиться, что данная функциональная форма в полной мере обеспечивает достижение поставленных исследователем целей.

12.2.5 Гиперболическая функциональная форма

Гиперболическая или обратная функциональная форма выражает Y как обратную функцию от одной или более независимых переменных, (в рассматриваемом примере от одной независимой переменной X_1): $Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_{1i}} \right) + \beta_2 X_{2i} + \varepsilon_i$.

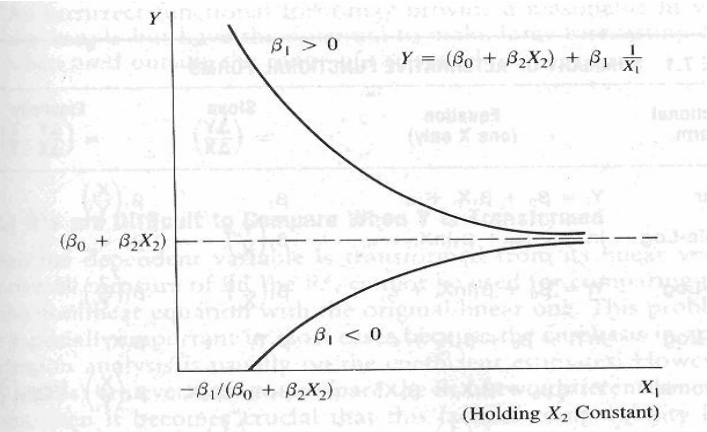
Обратная функциональная форма будет использоваться тогда, когда ожидаемое воздействие на единственную зависимую переменную близко к нулевому значению, в то время как независимая переменная существенно возрастает, стремясь к бесконечности. Отметим, что в таком случае, с ростом значений переменной X_1, ее воздействие на переменную Y будет убывать.

В рассматриваемом уравнении, переменная X_1 не может принимать нулевое значение, поскольку в этом случае переменная Y примет значение, равное бесконечности либо вообще неопределенное значение. Угловые коэффициенты принимают следующие

значения: $\frac{\Delta Y}{\Delta X_1} = -\frac{\beta_1}{X_1^2}; \frac{\Delta Y}{\Delta X_2} = \beta_2;$ угловой коэффициент при X_1 вписывается в две области, каждая из которых отражена на рисунке, который следует:

1. Когда $\beta_1 \succ 0$, угловой коэффициент по X_1, отрицателен и убывает по абсолютной величине одновременно с ростом X_1. Вследствие чего, зависимость Y от X_1 (X_2 будучи постоянным) с ростом X_1 стремится к виду $\beta_0 + \beta_2 X_2$.

2. Когда $\beta_1 \prec 0$, кривая пересекает ось X_1 в точке $-\frac{\beta_1}{(\beta_0 + \beta_2 X_2)}$ и угловой коэффициент возрастает, приближаясь к асимптотическому значению, близкому по значению к угловому коэффициенту для $\beta_1 \succ 0$.



Обратная функциональная форма находит широкое применение во многих разделах экономической теории и в реальной действительности. Если рассмотреть кривую Philips, представляющую обратную зависимость изменения заработной платы в процентах (W) от доли нетрудоустроенных U, то, наверняка, процентные изменения в заработной плате (W) окажут отрицательное воздействие на число безработных U, увеличив их количество, что, в свою очередь, будет содействовать уменьшению в будущем уровня заработной платы из-за институциональных и других причин. Такую гипотезу можно протестировать с помощью обратной функциональной формы. $W_t = \beta_0 + \beta_1\left(\frac{1}{U_t}\right) + \varepsilon_t$.

Оценка этого уравнения с помощью М.Н.К. дает следующий результат:

$$W_t = 0.00679 + 0.1842(1/U_t); \; R^2 = 0.397$$
$$(0.0590)$$
$$t = 3.20$$

12.2.6 Проблемы, возникающие при выборе некорректной функциональной формы

При выборе подходящей функциональной формы для модели регрессии, самый верный путь состоит в выборе спецификации, которая наилучшим образом соответствует теории, лежащей в основе исследуемого процесса или явления. В большинстве случаев, линейная функциональная форма является адекватной рассматриваемым процессам или явлениям, в остальных же случаях, исходя из знаний и опыта, подбирается одна из альтернативных функциональных форм, представленных в таблице.

Таблица функциональных форм.

Функциональная форма	Уравнение	Угловой коэффициент $=\left(\frac{\Delta Y}{\Delta X}\right)$	Эластичность $=\left(\frac{\Delta Y}{\Delta X}\cdot\frac{X}{Y}\right)$
Линейная	$Y_i=\beta_o+\beta_1 X_i+\varepsilon_i$	β_1	$\beta_1\left(\frac{X_i}{Y_i}\right)$
Полная логарифмическая	$\ln Y_i=\beta_o+\beta_1\ln X_i+\varepsilon_i$	$\beta_1\left(\frac{Y_i}{X_i}\right)$	β_1
Полулогаримическая ($\ln X$)	$Y_i=\beta_o+\beta_1\ln X_i+\varepsilon_i$	$\beta_1\left(\frac{1}{X_i}\right)$	$\beta_1\left(\frac{1}{Y_i}\right)$
Полулогаримическая ($\ln Y$)	$\ln Y_i=\beta_o+\beta_1 X_i+\varepsilon_i$	$\beta_1 Y_i$	$X_i\beta_1$
Полиномиальная	$Y_i=\beta_o+\beta_1 X_i+\beta_2 X_i^2+\varepsilon_i$	$\beta_1+\beta_2 X_i$	$\beta_1\left(\frac{X_i}{Y_i}\right)+2\beta_2\left(\frac{X_i^2}{Y_i}\right)$
Обратная	$Y_i=\beta_o+\beta_1\left(\frac{1}{X_i}\right)+\varepsilon_i$	$-\beta_1\left(\frac{1}{X_i^2}\right)$	$-\beta_1\left(\frac{1}{X_iY_i}\right)$

В некоторых случаях логические соображения констатируют, что модель является нелинейной по переменным, но определить вид этой формы сложно. В таком случае, линейная форма не является корректной, более того, из теоретических соображений, не может быть выбрана ни одна из альтернативных нелинейных функциональных форм. Именно в этих случаях, приходится расплачиваться (в терминах истинной зависимости) за выбор функциональной формы только на основе удачной аппроксимации. Рассмотрим различные ситуации, возникающие в этом случае.

– R^2 трудно сравнивать в случае преобразованных переменных.
– Неправильно подобранная функциональная форма может давать разумную аппроксимацию по наблюдениям, но имеет большие резервы по части увеличения статистической ошибки при прогнозе вне области наблюдаемых значений.
– Вызывает затруднение сравнение R^2 в случае, когда переменная Y преобразована.

В случае, когда зависимая переменная преобразуется от линейного вида к нелинейному виду, коэффициент детерминации R^2 не может быть использован для сравнения приближения нелинейного уравнения с оригинальным линейным уравнением. В большинстве случаев эта проблема не является особо важной, поскольку в регрессионном анализе акцент, как правило, ставится на оценку коэффициентов. Как бы то ни было, если R^2 (или $\overline{R}^2$) все-таки используются для сравнения двух аппроксимаций для двух различных функциональных форм, тогда это является важным. Например, предположим, что делается попытка сравнения линейного уравнения $Y_i=\beta_o+\beta_1 X_i+\varepsilon_i$ и его преобразованной версии $\ln Y_i=\beta_o+\beta_1\ln X_i+\varepsilon_i$. Отметим, что преобразование зависимой переменной отличает эти два уравнения. Причина, по которой коэффициент детерминации R^2 для исходного уравнения не может быть использован для сравнения

оценок, полученных по двум уравнениям, состоит в том, что, вычисленная общая вариация различна для этих двух уравнений. Коэффициент детерминации R^2 не может быть использован для сравнения, поскольку речь идет о различных уравнениях. Нет причин ожидать, что две различные зависимые переменные имеют одинаковое или сравнимое отклонение от среднего значения. А поскольку вариации различны, R^2 (или $\overline{R}^2$) не сравнимы между собой.

Чтобы избежать этой проблемы, строится «квази -R^2» путем преобразования прогнозных значений нелинейной зависимой переменной к виду, который является непосредственно сравнимым с оригинальной зависимой переменной. Эта преобразованная зависимая переменная используется для вычисления «квази-R^2». По существу, «квази-R^2» является тем R^2, который позволяет сравнивать аппроксимации, полученные с помощью различных функциональных форм, преобразуя прогнозные значения одной зависимой переменной к функциональному виду другой зависимой переменной.

Для рассмотренного примера, это означало бы выполнение следующих шагов:

а) Оценка уравнения $\ln Y_i = \beta_o + \beta_1 \ln X_i + \varepsilon_i$ и определение $\ln \widehat{Y}_{pr}$

б) Преобразование $\ln \widehat{Y}_{pr}$ через антилогарифмы к виду $anti \ln(\ln \widehat{Y}_i) = Y_i$.

с) Вычисление «квази-R^2» или («квази-$\overline{R}^2$»), используя вновь вычисленные значения в качестве $\widehat{Y}_i$, для того чтобы получить остаточные переменные, необходимые для вычисления R^2. «квази-R^2» $=1-\dfrac{\sum\left[Y_i - anti\ln(\ln\widehat{Y}_i)\right]^2}{\sum\left[Y_i-\overline{Y}\right]^2}$. Полученное значение для «квази-R^2» для уравнения в логарифмах может быть напрямую сравнимо с традиционным R^2 для линейного уравнения.

12.2.7 Пример неправильного использования коэффициента_$\overline{R}^2$

Для включения в рассмотрение влияния изменений в количестве независимых переменных необходимо использовать коэффициент детерминации $\overline{R}^2$, который рассчитан относительно степеней свободы $\overline{R}^2 = 1-\dfrac{\sum u_i^2/(n-k-1)}{\sum\left(Y_i-\overline{Y}\right)/(n-1)}$, $\overline{R}^2 = 1-\left(1-R^2\right)\dfrac{(n-1)}{(n-k-1)}$, $\overline{R}^2$ будет возрастать, убывать либо останется неизменным тогда, когда в уравнение регрессии вводится дополнительная переменная, зависит от того, улучшается аппроксимация, обусловленная включением новых переменных, при преодолении проблемы путем потери степеней свободы. Вывод следующий: каждый раз необходимо помнить, что качество уравнения регрессии является только одним из критериев, обеспечивающих качество регрессии в целом. Как отмечалось выше, степень оценки коэффициентов в соответствии с экономической теорией и ожидаемые значения этих коэффициентов имеют ту же значимость, что и качество уравнения регрессии. Если оцененное уравнение регрессии обладает хорошими характеристиками качества, то теоретическое соответствие и полезность также играют важную роль.

Из сказанного делаем вывод, что лучшим с точки зрения оцененного уравнения является лучшее соответствие теоретическому уравнению. К сожалению, многие начинающие исследователи предполагают, что если ($\overline{R}^2, R^2, r$) хороши, тогда максимальное значение $\overline{R}^2$ является хорошим средством для улучшения качества уравнения. Однако, такое предположение является опасным, хорошее соответствие уравнения является только одним из критериев, определяющих качество уравнения.

Вероятно, примером злоупотребления одним критерием в противовес остальным является выявления качества уравнения регрессии посредством максимизации $\overline{R}^2$, не принимая во внимание экономический смысл и статистическую значимость уравнения.

13. Спецификация: выбор независимых переменных

13.1 Важные переменные, не включенные в уравнение регрессии (опущенные переменные)

Прежде чем перейти к оценке уравнения регрессии, его необходимо специфицировать. Спецификация эконометрического уравнения состоит из трех этапов: 1) правильный выбор независимых переменных; 2) корректный выбор функциональной формы; 3) правильный выбор члена стохастической ошибки.

Спецификация стохастической ошибки следует из неправильной спецификации одного из предыдущих этапов. Остановимся более детально на первом этапе – выборе независимых переменных.

Исследователь решает, какие независимые переменные будут включать в уравнение регрессии, и это является как слабым, так и сильным моментом в эконометрике. Сильный момент состоит в том, что уравнение может быть сформулировано так, чтобы прогнозировать конкретные индивидуальные потребности. А недостатки состоят в том, что исследователь в состоянии рассмотреть столько спецификаций, сколько необходимо для того, чтобы найти ту, которая прогнозирует ожидаемые результаты даже, если другие характеристики ей противоречат. Основная цель данной лекции состоит в демонстрации такого способа выбора независимых переменных для уравнения регрессии, чтобы избежать ошибок неправильной спецификации.

Первым аргументом, в принятии решения в пользу включения в уравнение регрессии независимой переменной, будет ее соответствующее теоретическое обоснование. Если ответом является не очень уверенное «да», тогда переменная будет включена в уравнение только при условии, что соответствующие статистики значимы. Оставить важную переменную вне уравнения регрессии равносильно увеличению смещений оценок оставшихся переменных, если же в уравнение включается переменная, не обоснованная теоретически, то это увеличивает вариации оцененных коэффициентов.

.Предположим, что, по той или иной причине, при первоначальной спецификации уравнения, в уравнение забыли ввести важную с теоретической точки зрения переменную. Либо, предположим, что не были найдены данные для одной из переменных, предполагаемых в качестве независимых в уравнении регрессии. В обоих случаях результат одинаков: не была включена в уравнение важная независимая переменная.

И каждый раз, когда одна из независимых переменных опущена, объяснение и использование уравнения становится подозрительным, как в случае цены в уравнении спроса, ее отсутствие не только затруднит оценку коэффициента регрессии для цены, но, как правило, вызовет смещения оценок коэффициентов оставшихся в уравнении регрессии переменных. Смещение, вызванное исключением независимой переменной из уравнения регрессии, называется смещением вследствие спецификации (или, реже, смещением опущенной переменной). В уравнении со многими переменными (более чем одной переменной), коэффициенты $\hat{\beta}_k$ представляют изменение в зависимой переменной Y, вызванное изменением на одну единицу в независимой переменной X_k, в то время как остальные переменные в уравнении не меняются. Пропуск переменной вызывает смещения: это может вызвать изменение ожидаемых значений, оцененных коэффициентов, по сравнению с истинными значениями коэффициентов.

13.1.1 Последствия не включения независимых переменных в уравнение регрессии

Предположим, что истинное уравнение регрессии имеет следующий вид:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \ , \qquad (13.1)$$

здесь ε_i - классический член стохастической ошибки. Если исследователь не включил в уравнение важную независимую переменную, уравнение принимает вид:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i^* \ , \ \varepsilon_i^* = X_{2i} + \varepsilon_i \ . \qquad (13.2)$$

Член стохастической ошибки ε_i^* зависит от экзогенной переменной X_{1i}, если переменные X_{1i} и X_{2i} коррелированны между собой, т. е. одновременно с изменением переменной X_{2i} меняется и переменная X_{1i}. Другими словами, нарушены классические гипотезы, утверждающие, что изменение экзогенных переменных не зависит от изменения члена стохастической ошибки, пока опущенная переменная не коррелированна ни с одной из экзогенных переменных, включенных в уравнение. Последнее утверждение является почти невероятным.

В общем случае, при нарушении некоторых классических гипотез, не имеет места теорема Гаусса-Маркова, и оценки перестают быть BLUE. Из чего следует, что линейные оценки больше не являются не смещенными и не обладают минимальной вариацией (это относится к вариациям всех линейных оценок) или одновременно не выполняются другие гипотезы.

Оценка уравнения (13.2) для истинного уравнения (13.1) вызывает смещения в оценках уравнения (2). А это означает, что

$$E(\hat{\beta}_1) \neq \beta_1 . \qquad (13.3)$$

Неравенство ожидаемого значения коэффициента $\hat{\beta}_1$ истинному значению будет результатом исключения переменной X_2 из уравнения. Если переменные X_1 и X_2 коррелированны между собой, а переменная X_2 опущена из уравнения, тогда М.Н.К. присвоит переменной X_1 вариацию, существенно обусловленную переменной X_2, откуда и будет следовать смещение оценки для коэффициента $\hat{\beta}_1$.

Чтобы убедиться в том, что переменная, не включенная в уравнение, может вызвать смещение в оценках, рассмотрим производственную функцию, которая выражает объем производства Y в зависимости от количества использованного труда (X_1) и капитала (X_2). Что произойдет, если по каким-либо соображениям отсутствуют данные для капитала и переменная (X_2) опущена из рассмотрения, не включена в модель. Это исключение, без сомнения, сместит оценки для коэффициента при переменной труд, поскольку очевидна корреляция между трудом и капиталом (возрастание капитала, как правило, привлекает, по крайней мере, несколько дополнительных рабочих рук и обратно). И как следствие, М.Н.К. присвоит труду больший объем роста производства, в действительности, обусловленный ростом капитала, (капитал и труд коррелированны). Тогда смещение будет функцией, зависящей от коэффициента β_2 и от коэффициента корреляции между капиталом и трудом

$$E\left(\hat{\beta}_1\right) = \beta_1 + \beta_2 f(r_{12}) \qquad (13.4)$$

Следовательно, получаем, что ожидаемое значение коэффициента ($\hat{\beta}_1$) при переменной, включенной в уравнение, при исключении из рассмотрения важной переменной (X_2), равно истинному значению плюс истинное значение коэффициента при исключенной переменной, умноженное на функцию от простого коэффициента парной регрессии между включенной и исключенной из уравнения регрессии переменными. Вывод, смещения отсутствуют пока:

а) истинное значение β_2 равно нулю (а это означает, что переменная X_2 не является важной для модели);

б) r_{12} принимает нулевое значение (т.е. переменные X_1 и X_2 совершенно не коррелированны)

Компонента $\beta_2 f(r_{12})$ является константой смещения вследствие спецификации, внесенной в оценку коэффициента β_1 исключением из уравнения регрессии важной переменной X_2. Если переменная, включенная в уравнение и переменная, исключенная из него, не коррелированны, тогда смещений не существует. Однако, в этом случае, почти каждый раз, когда в действительности существует небольшая корреляция между какими – либо переменными (пусть даже, стохастическая), тогда, почти каждый раз, появляются смещения, обусловленные исключением важной переменной.

<u>Пример смещений, вызванных спецификацией</u>.

$$\widehat{Y}_t = -0{,}605 - 0{,}45PC_t + 0{,}12PB_t + 12{,}2\ln YD_t$$

(0,07) (0,05) (11,2)

$t =$ -6,4 2,5 10,6

$\overline{R}^2 = 0{,}984; \quad n = 35 \quad (anuale \quad 1950 - 1984)$

Y_t – *consumul de pasare*; PC_t – *costul unui kg de carne de pasare*;

PB_t – *costul unui kg de carne de vit* ; $\ln YD_t$ – *logaritmul natural al venitului disponibil pe cap de locuitor*;

Если оценим это уравнение, исключив из него цену товара заменителя, получим:

$$\widehat{Y}_t = -80{,}7 - 0{,}34PC_t + 15{,}0\ln YD_t$$

(0,06) (0,42)

$t =$ -5,6 36,0

$\overline{R}^2 = 0{,}981; \quad n = 35$

13.1.2 Корректировка исключенных переменных

Теоретически проблема смещений, обусловленных спецификацией, сводится к включению опущенной переменной в уравнение. Однако, к сожалению, это гораздо легче произнести, чем выполнить.

Во-первых, смещения опущенных переменных очень тяжело выявлять. В то время как некоторые показатели смещений вследствие ошибок спецификации, очевидны, как, например знак, оцененного коэффициента, противоположен ожидаемому знаку, другие показатели не так очевидны. Лучшим показателем того, что опущенная переменная является важной, является теоретическое обоснование модели. Какую переменную необходимо включить в уравнение? Каков ожидаемый знак при этой переменной? Известна ли какая либо информация относительно интервала изменения коэффициентов? Возможно, случайно была исключена переменная, которую большинство исследователей считают важной? Самым лучшим способом избежать исключения из уравнения важных переменных состоит в том, чтобы потратить немного времени на тщательное осмысливание уравнения, прежде чем вводить данные в компьютер.

Вторая проблема состоит в выборе переменной, которая будет добавлена в уравнение, как только вы решили, что уравнение подвержено смещениям вследствие исключения важных переменных. В уравнение сразу могут быть добавлены все возможные важные переменные, но это ведет к потере точности оценок. Либо, могут быть тестированы несколько переменных, и будет включена та переменная, которая имеет лучшие статистики с точки зрения уменьшения смещений. Но данная техника – добавления переменных с целью фиксации хороших результатов регрессии является не удачной,

поскольку переменная, которая наилучшим способом изменяет смещения вследствие спецификации, может сделать это в большей степени путем модификации решения, но ничуть не путем получения истинного решения задачи. При таких обстоятельствах, полученное уравнение может предоставить превосходные статистические результаты для одного набора исследуемых данных, в то время как эти результаты становятся ужасными при применении к другим наборам данных, поскольку они не описывают характеристики истинной выборки.

Включение переменных не способствует решению проблемы смещений, возникших вследствие исключения переменных. Если знак опущенной переменной отличается от того, который предполагался, он не может быть изменен в нужном направлении путем удаления из уравнения переменной, имеющей меньшее (по абсолютной величине) значение t-теста для оцененного коэффициента, чем значение t-теста для коэффициента, имеющего противоположный знак. Более того, знак вообще не может быть изменен даже, если переменная, которая подлежит исключению имеет очень большое значение t-теста.

Если, значение оцененного коэффициента существенно отличается от ожидаемого значения (как по знаку, так и по амплитуде), тогда, наверняка, в модели существуют некоторые смещения, вызванные спецификацией переменных. Хотя, достоверно известно, что данные плохого качества, равно как и слабая теория, также могут предоставить неверные знаки и амплитуды переменных, а такие события в некоторых случаях могут быть ликвидированы.

Техника, применяемая при уменьшении количества опущенных независимых переменных, состоит в исследовании направления смещений, полученных в результате исключения из уравнения регрессии важной независимой переменной. Если будет показано, что знак ожидаемого смещения от исключения переменной противоположен наблюдаемому знаку, тогда переменная может быть исключена из уравнения.

$$E(\widehat{\beta}_1) = \beta_2 f(r_{12}), \quad E(\widehat{\beta}_{PB}) = \beta_{PB} f(r_{PC,PB}) = (+)\cdot(+) = (+); \; E(\widehat{\beta}_{PC}) = \beta_{PD} \cdot f(r_{PC,PD}) = (-)(+) = (-).$$

Эта техника будет хорошо работать тогда и только тогда, когда только одна единственная переменная исключена из уравнения. Когда из уравнения регрессии одновременно исключены несколько переменных, их влияние на коэффициенты трудно специфицировать.

13. 2. Не значимые переменные

Что происходит, если в уравнение включается переменная, которая не является важной с теоретической точки зрения? Такой случай, когда включены не важные переменные, противоположен не включенным переменным и может быть исследован, используя модель, разработанную в предыдущем параграфе. В данном случае, модель содержит больше переменных в оцениваемом уравнении, чем в истинном уравнении.

Включение переменной в уравнение, которому она не принадлежит, не вызывает смещений, но способствует росту вариаций включенных оцененных коэффициентов.

13.2.1 Воздействие не значимых переменных

Пусть, истинное уравнение имеет следующую спецификацию:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \qquad (13.7),$$

а исследователь, исходя из определенных соображений, включил дополнительную переменную:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i^{**}, \quad \varepsilon_i^{**} = \varepsilon_i - \beta_2 X_{2i}. \qquad (13.8)$$

Такого рода ошибка не вызовет смещений, если истинный коэффициент не значимой переменной равен нулю. В этом случае, $\varepsilon_i^{**} = \varepsilon_i$ и $\widehat{\beta}_i$ является не смещенной оценкой в (13.8), когда β_2=0.

Включение не значимой переменной увеличит вариацию оцененных коэффициентов и будет стремиться уменьшить абсолютную амплитуду t-теста. Также не значимые переменные будут содействовать уменьшению $\overline{R}^2, (\textit{но не } R^2)$. В модели для Y, как функции от X_1, X_2, вариации оценок коэффициентов, полученных по М.Н.К.: $VAR(\widehat{\beta}_1) = \frac{\widehat{\sigma}_u^2}{(1-r_{12}^2)} \sum (X_{1i} - \overline{X}_1)^2, \textit{если } r_{12} = 0, \textit{то} \quad VAR(\widehat{\beta}_1) = \widehat{\sigma}_u^2 \sum (X_{1i} - \overline{X}_1)^2$. Даже если не значимые переменные не вызывают смещений, они доставляют проблемы для регрессии, поскольку уменьшают точность регрессии.

Влияние на оцененный коэффициент	Опущенная переменная	Включенная не значимая переменная
Смещения	Да*	Нет
Рост или убывание вар.	Убывает*	Растет*

*Если $r_{12} \neq 0$.

13.2.2 Правильный выбор спецификации

Рассмотрим четыре критерия оценки, которые могут помочь в принятии решения о принадлежности переменной уравнению.

а) Теория: вводится ли переменная в уравнение регрессии без двусмысленностей и теоретически проверенной?

б) t-тест: существенно ли отличен от нуля оцененный коэффициент?

в) $\overline{R}^2$: улучшается ли аппроксимация уравнения (приведенная к одной степени свободы), когда одна переменная добавляется в уравнение регрессии?

г) Смещения: существенно ли меняются другие коэффициенты, когда дополнительная переменная включена в уравнение?

Если все эти условия выполнены, переменная принадлежит уравнению. Если же ни одно из условий не выполнено, то переменная не является значимой и, наверняка, может быть исключена из уравнения. Когда значимая переменная включена в уравнение, ее включение может содействовать росту $\overline{R}^2$ и изменению других коэффициентов, оставляя значимой неизменную величину t-теста. С другой стороны, если не значимая переменная включена в уравнение, она будет способствовать уменьшению $\overline{R}^2$, имея величину t-теста не значимой, и, как следствие, небольшое влияние на значение коэффициентов при остальных независимых переменных.

Во многих случаях, все четыре критерия не согласуются между собой. Это случается для переменных, которые имеют не значимый t-тест. В других случаях, переменная может быть относительно некоррелированной с переменными, входящими в уравнение, и, как следствие, иметь небольшое влияние на оцененные коэффициенты. Что делать в таких случаях?

Единственное и самое важное обоснование при определении важности переменной состоит в теоретическом обосновании. Не количество статистических данных докажет теоретическую необходимость «важности» переменной. Иногда исследователь вынужден оставить вне уравнения переменную, важную с теоретической точки зрения, в отсутствии лучшей альтернативы, но в этих случаях полезность уравнения ограничена.

13.3 Поиски спецификаций

Одна из слабых сторон эконометрики состоит в том, что исследователь может значительно манипулировать множеством данных, чтобы получить различные результаты, специфицируя различные уравнения регрессии, пока не будут получены оценки с искомыми качествами.

Имеет смысл сделать попытку минимизации количества оцененных уравнений, хотя эта задача не из легких, и, как можно чаще при выборе переменных, в большей степени основываться на теории, чем на статистической аппроксимации. Попробуем продемонстрировать это на трех, наиболее часто используемых, не корректных процедурах для спецификации уравнения регрессии.

12.3.1 Поиск данных в целях максимизации коэффициента детерминации

Наверняка, самой плохой является попытка сформулировать одновременно последовательность регрессий и выбрать уравнение, которое лучшим образом соответствует результатам, которые желает получить исследователь. В таком случае исследователь захочет оценить практически все возможные комбинации различных альтернативных независимых переменных и выбор среди них будет сделан на основе полученных результатов.

Такая практика одновременной оценки некоторой комбинации независимых переменных и выбор среди них лучшей не учитывается количество рассмотренных уравнений, включая последнее из них. При 95% уровне доверия для того, чтобы результаты регрессии не были случайными, а вы рассмотрели более 20 регрессий, каким будет доверие к полученным результатам? В то же время, оставив регрессию с большим значением t-теста и, игнорируя ту, у которой значение R^2 меньше, получим преувеличенный t-тест для оценки значимости коэффициентов.

Кроме того, такое «исследование данных» и «выуживание спешки» при получении необходимых статистик для финального уравнения регрессии, является процедурой, потенциально лишенной этики экспериментальных исследований. Такая процедура включает не только альтернативные комбинации независимых переменных, но и большое количество функциональных форм, лаговых структур и, предоставляемую на данном этапе изощренную и продвинутую технику оценки. Другими словами, если испробованы существенное количество альтернатив, чрезвычайно возрастает шанс получить необходимые результаты, в то время как финальный результат будет обесцененным (непригодным). Исследователь не ищет научного обоснования в поддержку первоначальных гипотез; более того, предыдущие ожидания навязаны данным таким способом, который является ошибочным по существу.

13.3.2 Процедура итеративной регрессии

Итеративная регрессия привлекает использование компьютерных программ для выбора независимой переменной, при оценке специфицированного уравнения. Компьютерная программа предоставляет нам перечень независимых переменных, а затем по этапам осуществляется их выбор. В первую очередь, выбирается независимая переменная, которая одна объясняет большую часть вариации зависимой переменной от среднего значения. В качестве второй переменной выбирается та, которая в большей степени определяет R^2, учитывая то, что первая переменная уже входит в уравнение. Итеративная процедура продолжается до тех пор, пока следующая переменная, которая должна быть включена в уравнение, не может способствовать, какому бы то ни было, росту R^2(который иногда называется «стертый» R^2), вызванному вводом в уравнение новой переменной.

К сожалению, любая корреляция между независимыми переменными, затрудняет использование этой процедуры. Трудно отделить влияние одной переменной от другой при измерении степени коррелированности переменных. В результате, при наличии мультиколлинеарности, невозможно определить индивидуальный вклад каждой специфицированной переменной, чтобы утверждать, что одна из них является самой важной и, следовательно, первой будет включена в уравнение. Еще хуже, что теряется необходимость в теоретическом обоснования выбранной комбинации специфицированных переменных.

В силу этих причин, многие исследователи избегают применения итеративной процедуры. Наибольшая опасность заключается в том, что полученные коэффициенты могут быть смещенными, вычисленные значения t-теста не следуют распределениям t-теста в будущем , важные переменные могут быть исключены вследствие проведения многочисленных попыток включения, и, наконец, знаки оцененных коэффициентов на каждом промежуточном или финальном этапе итеративной процедуры могут отличаться от ожидаемых знаков. Использование итеративной процедуры предполагает игнорирование интереса к порядку включения переменных.

13.3.3 Последовательный поиск спецификаций

Из непонятных соображений большинство специалистов из области эконометрики избегают использовать процедуру исследования данных и итеративную процедуру спецификации независимых переменных. Вместо этого стремясь специфицировать уравнение, оценив первоначальное уравнения, а затем, последовательно выводя или добавляя переменные (или изменяя функциональную форму) до тех пор, пока правдоподобное уравнение с «хорошими статистиками» не будет получено. По-видимому является общепринятой практикой прибегать к проверке коэффициента детерминации R^2 и значения t-теста для всех переменных (в обоих случаях до и после селекции), находясь в ситуации, когда теоретически обоснованы сведения о нескольких важных независимых переменных, не обладая при этом знанием того, что остальные переменные, включенные в уравнение, являются важными. Легко показать, что такой последовательный поиск является лучшим путем поиска «неправды». В то время, как существует огромная разница в подходе последовательного поиска и рекомендуемом подходе.

Последовательный поиск спецификации является процедурой, которая позволяет исследователю оценить секретное число регрессий, а затем предоставить финальный выбор, (который не основан на специфицированном множестве ожиданий по поводу знаков и значимости коэффициентов), представленный в качестве оцененной спецификации. Такой метод ошибочно фиксирует статистическую правдоподобность результатов регрессии по двум причинам:

а) статистическая значимость результатов переоценена, поскольку игнорируются результаты предыдущей регрессии,

б) множество ожиданий, используемых исследователем при выборе из результатов различных регрессий, в редких случаях или никогда, не является секретной. А тогда, читатель не имеет никакой возможности узнать: предлагают или нет другие результаты, проведенных регрессий, противоположные знаки или значимые коэффициенты для важных переменных.

К сожалению, не существует признанного универсального метода для управления последовательным поиском, в первую очередь, потому, что тест, подходящий к одному этапу процедуры, зависит от тестов, которые были выполнены перед этим, и еще потому, что эти тесты очень трудно придумать. Одна из возможностей состоит в уменьшении числа степеней свободы в «окончательном» уравнении регрессии на одну, для каждой

альтернативной попытки спецификации. Такая процедура далеко не точна, но налагает явный штраф на поиск спецификации.

В общем случае, рекомендуется поддерживать число оцененных регрессий на возможно меньшем уровне. Сконцентрироваться на теоретических соображениях тогда, когда выбираются независимые переменные, функциональные формы и раскрыть все исследованные спецификации. Т.е., рекомендуется комбинировать экономику (используя теорию и анализ при ограничении количества оцениваемых спецификаций) с объявлением всех оцененных уравнений.

И все-таки это другая сторона вопроса. Некоторые исследователи чувствуют, что истинная модель прямо покажет, что существует шанс получить лучшие статистические результаты (включая знаки и коэффициенты), наилучшим образом соответствующие истинным спецификациям. Проблема такой психологии состоит в том, что элемент шанса является очень важным элементом любого исследования. Дополнительно к этому, разумные личности часто не согласны с тем, как выглядит «истинная модель». Следовательно, различные исследователи будут анализировать один и тот же набор данных и «придут» с «лучшими» предельно отличающимися моделями. Поскольку это может произойти, то отличие между хорошим и плохим исследователем в области спецификации, вероятно, состоит в том, что он будут действовать разумным образом.

Уроки этого раздела прозрачны. Самые важные шаги при проведении спецификации уравнения регрессии будут проведены в самом начале до того, как будет предпринята любая попытка оценки уравнения регрессии с помощью компьютера. Поскольку не разумно рассчитывать на совершенного исследователя, будут периоды, когда дополнительные спецификации необходимо будет оценить. В любом случае, эти новые оценки необходимо основательно мотивировать теоретически и явно принимать во внимание тогда, когда тестируется значимость и обобщаются итоги. Только так может быть уменьшена опасность неверных статистических оценок.

13.3.4 Последствия последовательного поиска спецификаций

На приме покажем как исключение переменных из модели на основе значения t-теста только вносит систематические смещения в оцениваемое уравнение. Пусть воображаемая модель для одной независимой переменной имеет вид: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$. Предположим, что в будущем, основываясь на соответствующей теории, было установлено, что переменная X_1 принадлежит уравнению, а переменная X_2 не принадлежит ему. Поскольку критерий определенности требует ответа на вопрос: включать либо не включать в уравнение переменную X_2, то многие опытные исследователи используют только значения t-теста, который показывает, что коэффициент β_2 существенно отличается от нуля. А потому они оставляют переменную X_2 в уравнении, получая его вид, который приведен выше как окончательная форма уравнения регрессии. В противном случае, когда значения t-теста показывает не существенное отличие от нуля, этот исследователь решает исключить переменную X_2 из уравнения регрессии и рассматривает зависимость Y от одной независимой переменной X_1.

При использовании такого подхода могут возникнуть два типа ошибок. Первый, переменная X_2 может быть сохранена в уравнении, в то время как она не принадлежит ему, и такая ошибка не изменит ожидаемого значения коэффициента β_1. Второй, переменная X_2 может быть исключена из уравнения, хотя должна принадлежать ему. Тогда, оцененный коэффициент β_1 для переменной X_1 будет смещен на величину β_2, в зависимости от степени корреляции переменных X_1 и X_2. Другими словами,

коэффициент β_1 будет смещен столько времени, сколько времени будет отсутствовать в уравнении переменная X_2, которая должна принадлежать ему. Т.е. переменная X_2 будет исключена из уравнения каждый раз, когда оцененный коэффициент не существенно отличается от нуля, и в этом случае ожидаемое значение $\widehat{\beta}_1$ не будет равно β_1, и будет иметь в рассматриваемом уравнении систематические смещения: $E(\widehat{\beta}_1) = \beta_1 + \beta_2 \cdot f(r_{X_1/X_2})P \neq \beta_1$. Здесь P указывает на вероятность значимости t-теста. Это как раз тот случай, когда t-тест для $\widehat{\beta}_1$ больше не соответствует "t" распределению. Другими словами, t-тест смещен вследствие последовательно поиска спецификации.

Поскольку, большинство исследователей доверять перед тем как получить окончательный вид модели рассматривают различное количество переменных, обладающих значением t-теста, которому можно, то тем самым систематически приходится сталкиваться и преодолевать эту проблему. Практика исключения потенциально независимых переменных только по причине того, что индивидуальные t-тесты их оцененных коэффициентов не отличается существенно от нуля, приводит к систематическим смещениям оцененных коэффициентов оставшихся переменных, а также их t-тестов.

14. Экономические данные

Для проведения количественного анализа необходимо провести сбор, обработку и ввод данных в компьютер. Как правило, это неблагодарная работа, которая занимает много времени, поскольку данные трудно найти, теоретические данные отличаются от эмпирических данных, и велика вероятность появления типографских ошибок и ошибок набора. А потому исследователь, который знаком с источниками информации и методами ее определения, с меньшей вероятностью может допустить ошибки при использовании или интерпретации результатов регрессии, выиграв время за счет определения природы данных и их сбора.

14.1 Искомые данные

Когда выбирается тема для исследований, в первую очередь необходимо быть уверенным, что данные для зависимой переменной и для всех относящихся к делу независимых переменных, возможно найти. В любом случае, проверка доступности данных означает первый шаг на пути спецификации переменных, которые примут участие в исследовании. Следует отметить, что половина времени, используемая начинающими исследователями на сбор информации, тратится на поиск некорректных переменных из ошибочных источников. Несколько минут, потраченные на обдумывание природы данных, равносильны многим часам, потраченным впустую в будущем.

К примеру, если зависимая переменная определяет количество телевизоров, реализованных за один год, тогда и большинство независимых переменных также будут годовыми переменными, но ни в коем случае квартальными или месячными. Было бы неподходящим или попросту ошибочным рассматривать цены на TV как месячные цены. Средняя цена в течение одного года, определяемая как взвешенная цена по количеству ежемесячно продаваемых телевизоров, будет более понятной. Если зависимая переменная включает все проданные телевизоры, независимо от марки, тогда наиболее подходящей ценой будет агрегированная цена, основанная на ценах телевизоров всех марок. И тем не менее, вычисление такого рода переменной не является подходящим.

Статистические данные, используемые в регрессионном анализе, представляют связующее звено между теоретическим экономическим моделированием и реальной

экономикой, которая должна быть понята с помощью этого звена. Большинство моделируемых экономических процессов и явлений можно с достаточной точностью измерить и представить в виде числовых значений. Количественные данные можно получить из Статистических ежегодников, некоторые из них можно использовать сразу, другие же требуют дополнительной обработки.

Экономические и социальные науки очень часто имеют дело с качественными феноменами дихотомического типа. Все случаи, которые характеризуются дихотомическим феноменом, могут быть представлены с помощью логических переменных или переменных dummy.

14.2 Статистические данные могут иметь следующую природу:

Временные данные, в которых зафиксирован объект исследования, рассматриваемый в различные моменты времени.

Перекрестные данные, для которых при фиксированном моменте времени исследуются различные объекты.

Панельные данные, которые представляют данные опросов, опытные данные и являются наиболее достоверными данными.

Данные отсутствуют или не укомплектованы:

Cross-section или перекрестные данные, отсутствуют некоторые данные из одного или более наблюдений, эти наблюдения могут быть удалены.

Time-series или временные ряды, отсутствуют данные для некоторого периода времени, тогда оцениваются пропущенные данные путем интерполирования либо, используя среднее значение для смежных данных.

Нужны квартальные данные, а имеются только годовые данные, тогда итерполируются квартальные данные. В любом случае, процедура интерполяции может быть оправдана только в том случае, когда изменения в переменных плавные и гладкие. Если же данные полностью отсутствуют, то исключение переменной создает проблемы, поскольку удаление переменной из рассмотрения приводит к смещениям оценок.

В общем случае, удачный регрессионный проект может быть приостановлен из-за неадекватных данных. Во многих случаях даже простая регрессионная техника не может быть применена, поскольку информация является искаженной. Иногда она измерена с таким количеством ошибок, что пользователь в состоянии построить таблицы и графики и сделать соответствующие выводы. Между прочим, эти таблицы и графики служат полезным вспомогательным материалом при обосновании уравнения регрессии.

Источники экономических данных

Статистические ежегодники
Бюллетени Национального Банка Молдовы
Экономические тенденции в экономике Молдовы
International Financial Statistics
www.IREX.RU/PUBLICATIONS/POLEMICA/5/ARTY.HTM
www.WEFA.com
www.CIRS-md.org
www.met.dnt.md
http://ecfor.rssi.ru

14.3.1 Переменные «proxy» или делегированные переменные

Переменные «proxy» используются в целях замены необходимых теоретических переменных тогда, когда данные для соответствующих переменных либо отсутствуют, либо являются неполными. Например, значения «чистых» инвестиций является значением переменной, которая не подлежит оценке в большом числе стран. Следовательно, исследователь может использовать значения инвестиций «брутто» как значение переменной «proxy» в предположении, что значения инвестиций «брутто» прямо пропорциональны значениям чистых инвестиций. Пропорциональность – это все, что необходимо, поскольку регрессионный анализ является, в первую очередь, зависимостью между изменением в переменных, между абсолютным уровнем переменных.

В общем случае переменная «proxy» – это «хорошая» делегированная переменная тогда, когда ее относительное изменение сравнительно хорошо соотносится с изменениями в корректной теоретической переменной.

14.3.2 Использование временных лагов в экономике и в эконометрике

Большинство исследуемых регрессий являются одновременными по своей природе. Другими словами, они включают значения зависимой и независимой переменных в один и тот же период времени: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$, индекс t фиксирует точки для разных моментов времени; тогда когда все переменные имеют один и тот же индекс, уравнение является одновременным по времени. Однако, не все процессы в экономике или бизнесе отражают такую одновременную зависимость во времени между зависимой и независимыми переменными. Во многих случаях необходимо предоставить возможность, чтобы между изменениями зависимой переменной и независимыми переменными прошло некоторое время. Длительность этого промежутка во времени между причиной и эффектом называется лагом (lag). Многие эконометрические уравнения включают одну или более независимых переменных с запаздыванием во времени, как, например, X_{t-1}, где индекс $t-1$указывает на то, что наблюдение X_{t-1} относится к периоду времени, предшествующему моменту времени t, как, к примеру, в следующем уравнении: $Y_t = \beta_o + \beta_1 X_{1t-1} + \beta_2 X_{2t} + \varepsilon_t$. Любое изменения на сельскохозяйственном рынке, как то рост цен, которое фермер может заработать при продвижении определенного продукта, имеет запаздывающий эффект по предложению этого продукта.

Аналогично, многие макроэкономические теории имеют неявную запаздывающую структуру, входящую в их состав. Длительность во времени между решениями относительно таких макроэкономических политик (как государственные затраты или рост денежного предложения) и последствиями воздействия этих политик на ВВП, занятость или цены, как правило, измеряется в годах. Рост денежного предложения стимулирует рост ВВП, как следствие роста инвестиций. Однако инвестиции не могут измениться за ночь, поскольку должны быть приняты решения относительно их роста. А эти решения должны быть воплощены в спроектированные планы, должны быть приняты на работу дополнительные сотрудники и т. д. В действительности, как отметил экономист Milton Friedman, для полной реализации изменений в денежной политике необходимо от 6 до 30 месяцев.

Если сформулирована гипотеза относительно простого лага, возникают сложности использования этого лага в эконометрических уравнениях. Переменная с запаздыванием вводится в эконометрическое уравнение, как и любая другая переменная. Например, уравнение для предложения хлопка имеет вид:

$C_t = F(PC_{t-1}, PF_t) = \beta_0 + \beta_1 PC_{t-1} + \beta_2 PF_t + \varepsilon_t$, где

C_t – объем спроса на хлопок в году t;

PC_{t-1} - цена хлопка в году $t-1$;

PF_t - расходы на оплату труда фермера в году t.

Смысл коэффициентов регрессии для переменных с запаздыванием отличается от смысла коэффициентов регрессии для переменных без запаздывания. Оцененный коэффициент при запаздывающей переменной измеряет изменения, происходящие в зависимой переменной Y в текущем году при изменении на одну единицу в значении независимой переменной X в предыдущем году, при этом значения оставшихся независимых переменных X_s в уравнении принимают постоянные значения. Следовательно, коэффициент β_1 в рассматриваемом уравнении измеряет количество единиц хлопка, которые будут произведены дополнительно в текущем году в результате изменении на единицу, произошедшем в ценах на хлопок в предыдущем году, при этом расходы на оплату труда фермера остаются неизменными.

14.3.3 Переменные «dummy»

Некоторые переменные (как например, род) могут быть объяснены только на качественном уровне. Такие переменные в большинстве случаев квалифицируются как двоичные переменные или как переменные «dummy». Переменные «dummy» принимают значения 0 или 1 в зависимости от выполнения или невыполнения некоторого условия. Основное их использование будет представлено далее.

Чтобы продемонстрировать это предположим, что Y_i задает зарплату школьного учителя i, при этом уровень зарплаты зависит, в первую очередь, от полученной степени и от опыта преподавания. Все учителя имеют степени $B.A.$, а некоторые из них имеют степень $M.A.$ Уравнение, которое представляет зависимость между заработком и полученной степенью, может иметь вид:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, где

$$X_{1i} = \begin{cases} 1, & \text{если учитель "i" имеет } M.A. \\ 0, & \text{в противном случае} \end{cases}$$

X_{2i} - стаж работы учителем для i-ого лектора.

Переменная X_1 принимает только два значения 0 и 1, и, следовательно, переменная X_1 называется переменной «dummy» или попросту dummy. В данном примере переменная dummy представляет условие наличия $M.A.$.

Коэффициенты регрессии для этого уравнения интерпретируются следующим образом.

а) если учитель имеет степень $B.A.$, только $X_1 = 0$ и $E\left(Y_i / X_i\right) = \beta_0 + \beta_2 X_{2i}$

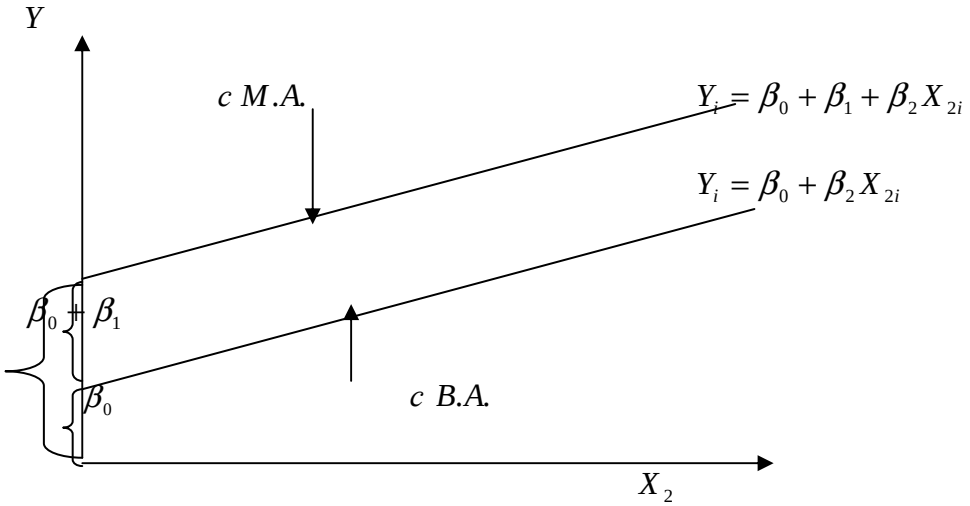
б) если учитель имеет степень $M.A.$, $X_1 = 1$ и $E\left(Y_i / X_i\right) = \beta_0 + \beta_1 + \beta_2 X_{2i}$.

Сравнивая эти два уравнения, приходим к выводу, что β_1 представляет дополнительный средний заработок, полученный учителем, имеющим степень мастера, по сравнению с учителем, имеющим степень бакалавра, при одном и том же стаже работы в качестве учителя. Возможность постулирования знака коэффициента в уравнении регрессии является важным моментом в регрессионном анализе. Так, например, переменная dummy называется переменной пресечения, поскольку она действительно меняет точку пересечения уравнения регрессии в зависимости от того, имеет учитель степень мастера или нет.

Альтернативная формулировка модели регрессии получается, если определить X_{1i} как:

$$X_{1i} = \begin{cases} 0, & \text{если учитель "i" имеет } M.A. \\ 1, & \text{в противном случае} \end{cases}$$

Эти условия меняют политику. В этом случае коэффициент β_1 будет интерпретирован как разность между средним заработком преподавателя при наличии *В.А.* и средним заработком преподавателя при наличии *М.А.*, при этом знак предполагается отрицательным. Несмотря на то, что коэффициент



β_1 при переопределенной переменной будет иметь противоположный знак, его абсолютная амплитуда будет такой же, как и для β_1 при оригинальной переменной. Это объясняется тем, что эти два коэффициента β_s измеряют одни и те же события, но в противоположных направлениях. В этом смысле определение переменных dummy является произвольным. Как только они определены, может быть дано единственное представление.

Надо отметить, что в данном примере была использована только одна переменная dummy, хотя существуют два условия. Это происходит из-за того, что условия строятся на основе одной переменной. Событие не представлено явно переменной dummy, опущенное условие формирует базу, по которой включенное условие сравнивается. Так, в двойственных ситуациях (типа *М.А.* и *В.А.*), включается только одна независимая переменная dummy; коэффициент интерпретируется как эффект от включенного условия по отношению к опущенному условию. Если третье условие (наличие Ph.D) будет включено, только тогда будут использованы две переменные.

Начинающие исследователи очень часто совершают ошибку, включая столько переменных dummy сколько есть условий, и такая модель становится бесполезной, поскольку переменные dummy добавляют константы, которые коллинеарны со свободным членом, уже включенным в уравнение. Поскольку совершенная мультиколлинеарность определяется как линейная зависимость между частью или всеми независимыми переменными, то это происходит из-за того, что сумма переменных dummy равна единице для всех произведенных наблюдений. В этом случае регрессионные программы, используемые на компьютерах, не выдадут никакого результата.

Переменная dummy, которая только для одного наблюдения принимает значение 1, а для остальных наблюдений - значение 0, не будет рассматриваться. Такое действие, как один раз dummy просто-напросто исключает это наблюдение из множества данных, тем самым искусственно улучшая структуру данных, определяя коэффициент при

переменной dummy равным остаточной переменной для этого наблюдения. То же самое может быть получено для остальных коэффициентов, если наблюдение будет исключено, т.е. удаление наблюдения является каждый раз более подходящей процедурой.

Иногда переменные dummy используются при проведении расчетов сезонных вариаций по данным для моделей с временными рядами. Например, если

$$X_{1t} = \begin{cases} 1 & \text{в I квартале} \\ 0 & \text{в остальных} \end{cases}$$

$$X_{2t} = \begin{cases} 1 & \text{в II квартале} \\ 0 & \text{в остальных} \end{cases},$$

$$X_{3t} = \begin{cases} 1 & \text{в III квартале} \\ 0 & \text{в остальных} \end{cases}$$

тогда $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t$ где переменная X_4 является независимой переменной, отличной от dummy, а t индекс квартальных наблюдений. Отметим, что только три переменные dummy необходимы для представления четырех времен года. В такой формулировке β_1 измеряет насколько ожидаемое значение Y в первом квартале отличается от его ожидаемого значения в четвертом квартале. β_2 *и* β_3 интерпретируются аналогично. Данная процедура может быть использована только тогда, когда к переменным Y *и* X_4 не было применено сезонное сглаживание до того, как было произведено оценивание. Включение переменных dummy в соответствии со временами года делает переменную Y не зависящей от сезона, равно как и другие независимые переменные, которые не сглажены по времени года.

14.4 Этапы при проведении прикладной регрессии

– Изучение литературы по рассматриваемой проблеме
– Спецификация модели: выбор независимых переменных и функциональной формы
– Выдвижение гипотез относительно ожидаемых знаков при коэффициентах уравнения регрессии
– Сбор данных
– Вычисление коэффициентов уравнения регрессии и его оценка
– Оформление результатов

15. Системы эконометрических уравнений

15.1 Общие понятия о системах уравнений, используемых в эконометрике

Объектом статистического изучения в социальных науках являются сложные системы. Измерение тесноты связей между переменными, построенное на исследовании изолированных уравнений регрессии недостаточно для описания таких систем и объяснения механизма их функционирования. При использовании отдельных уравнений регрессии, например, для экономических расчетов, в большинстве случаев предполагается, что аргументы (факторы) можно изменять независимо друг от друга. Однако это предположение является очень грубым, поскольку практически изменение одной переменной, как правило, не может происходить при абсолютной неизменности других. Ее изменение повлечет за собой изменения во всей системе взаимосвязанных признаков. Следовательно, отдельно взятое уравнение множественной регрессии не может характеризовать истинное влияние отдельных признаков на вариацию результирующей переменной. Именно поэтому в последние десятилетия в экономических, биометрических и социологических исследованиях важное место заняла

проблема описания структуры связей между переменными системой так называемых одновременных уравнений, называемых также структурными уравнениями. Так, если изучается модель спроса как соотношение цен и количества потребляемых товаров, то одновременно для прогнозирования спроса необходима модель предложения товаров, в которой также рассматривается взаимосвязь между количеством и ценой предлагаемых благ. Это позволяет достичь равновесия между спросом и предложением. При оценке эффективности производства нельзя руководствоваться только моделью рентабельности. Она должна быть дополнена моделью производительности труда, а также моделью себестоимости единицы продукции.

В еще большей степени возрастает потребность в использовании взаимосвязанных уравнений, если мы переходим от исследований на микро уровне к макроэкономическим расчетам. Модель национальной экономики включает в себя систему уравнений: функции потребления, инвестиций, заработной платы, а также тождество доходов и т.д. Это связано с тем, что макроэкономические показатели, являясь обобщающими показателями, чаще всего взаимозависимы. Так, расходы на конечное потребление в экономике зависят от валового национального дохода. В то время как величина валового национального дохода рассматривается как функция инвестиций.

Системы уравнений в эконометрических исследованиях могут быть построены по-разному. Возможна система независимых уравнений, когда каждая зависимая переменная y рассматривается как функция одного и того же набора факторов x:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Набор факторов x_i в каждом уравнении может варьировать. Так, модель вида

$$y_1 = f(x_1, x_2, x_3, x_4, x_5)$$
$$y_2 = f(x_1, x_3, x_4, x_5)$$
$$y_3 = f(x_2, x_3, x_5)$$
$$y_4 = f(x_3, x_4, x_5)$$

также является системой независимых уравнений с тем лишь отличием, что в ней набор факторов видоизменяется в уравнениях, входящих в систему. Отсутствие того или иного фактора в уравнении системы может быть следствием как экономической нецелесообразности его включения в модель, так и несущественности его воздействия на результативный признак (незначимо значение t критерия или частного F - критерия для данного фактора).

Каждое уравнение системы независимых уравнений может рассматриваться самостоятельно. Для нахождения его параметров используется метод наименьших квадратов. По существу каждое уравнение этой системы является уравнением регрессии. Поскольку никогда нет уверенности, что факторы полностью объясняют зависимые переменные, то в уравнениях присутствует свободный член a_0. Так как фактические значения зависимой переменной отличаются от теоретических значений на величину случайной ошибки, то в каждом уравнении присутствует величина случайной ошибки.

В результате чего система независимых уравнений при трех зависимых переменных и четырех факторах примет вид:

$$\begin{cases} y_1 = a_{01} + a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + \varepsilon_1, \\ y_2 = a_{02} + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 + \varepsilon_2, \\ y_3 = a_{03} + a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 + \varepsilon_3. \end{cases}$$

Однако если зависимая переменная y одного уравнения выступает в виде фактора x в другом уравнении, то возможно построить модель в виде системы рекурсивных уравнений:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + \ldots + a_{3m}x_m + \varepsilon_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nm-1}y_m + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

В данной системе зависимая переменная y включает в каждое последующее уравнение в качестве факторов все зависимые переменные предшествующих уравнений наряду с набором собственно факторов x. Примером такой системы может служить модель производительности труда и фондоотдачи вида:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2, \end{cases}$$

Где y_1 – производительность труда; y_2 – фондоотдача; x_1 - фондовооруженность труда; x_2 - энерговооруженность труда; x_3 - квалификация рабочих.

Как и в предыдущей системе, каждое уравнение может рассматриваться самостоятельно, и его параметры определяются М.Н.К.

Наибольшее распространение в эконометрических исследованиях получила система совместных уравнений. В ней одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других уравнениях – в правую часть системы:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + \ldots + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + \ldots + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nm-1}y_m + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Система взаимозависимых уравнений получила название системы совместных или одновременных уравнений. Тем самым подчеркивается, что в системе одни и те же переменные y одновременно рассматриваются как зависимые в одних уравнениях и как независимые в других. В эконометрике эта система уравнений называется также структурной формой модели. В отличие от предыдущих систем каждое уравнение системы одновременных уравнений не может рассматриваться самостоятельно, и для нахождения его параметров традиционный М.Н.К. неприменим. С этой целью используются специальные приемы оценивания.

Примером системы одновременных уравнений может служить модель динамики цены и заработной платы вида

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2, \end{cases}$$

где y_1 – темп изменения месячной заработной платы; y_2 – темп изменения цен; x_1 - процент безработных; x_2 - темп изменения постоянного капитала; x_3 - темп изменения цен на импорт сырья.

15.2 Структурная и приведенная форма модели

Система совместных, одновременных уравнений (или структурная форма модели) обычно содержит эндогенные и экзогенные переменные. Эндогенные переменные обозначены в приведенной ранее системе уравнений как y. Это зависимые переменные, число которых равно числу уравнений системы. Экзогенные переменные обозначаются обычно как x. Это предопределенные переменные, влияющие на эндогенные переменные, но не зависящие от них.

Простейшая структурная форма модели имеет вид:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + \varepsilon_2, \end{cases}$$

где y - эндогенные переменные; x - экзогенные переменные.

Классификация переменных на эндогенные и экзогенные переменные зависит от теоретической концепции принятой модели. Экономические переменные могут выступать в одних моделях как эндогенные переменные, а в других моделях как экзогенные переменные. В качестве экзогенных переменных могут рассматриваться значения эндогенных переменных за предшествующий период времени (лаговые переменные). Так, потребление текущего года y_t может зависеть не только от ряда экономических факторов, но и от уровня потребления в предыдущем году y_{t-1}. Не экономические переменные (например, климатические условия) входят в систему как экзогенные переменные.

Структурная форма модели позволяет увидеть влияние любой экзогенной переменной на значения эндогенной переменной. Целесообразно в качестве экзогенных переменных выбирать такие переменные, которые могут быть объектом регулирования. Меняя их и управляя ими, можно определить заранее заданные целевые значения эндогенных переменных. Структурная форма модели в правой части содержит при эндогенных и экзогенных переменных коэффициенты a_{is} и b_{ik}, которые называются структурными коэффициентами модели. Все переменные модели выражены в отклонениях от среднего уровня, т.е. под переменной x подразумевается $x - \bar{x}$, а под переменной y - соответственно $y - \bar{y}$, поэтому свободный член в каждом уравнении системы отсутствует.

Использование М.Н.К. для оценивания структурных коэффициентов модели дает, как принято считать в теории, смещенные и несостоятельные оценки. Поэтому, обычно для определения коэффициентов в таких моделях, структурная форма модели преобразуется в приведенную форму модели.

Приведенная форма модели представляет собой систему нелинейных относительно коэффициентов функций эндогенных переменных от экзогенных переменных:

$$\begin{cases} \hat{y}_1 = \delta_{11}x_1 + \delta_{12}x_2 + \ldots + \delta_{1m}x_m, \\ \hat{y}_2 = \delta_{21}x_1 + \delta_{22}x_2 + \ldots + \delta_{2m}x_m, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{y}_n = \delta_{n1}x_1 + \delta_{n2}x_2 + \ldots + \delta_{nm}x_m, \end{cases}$$

где δ_{ij} - коэффициенты приведенной формы модели.

По своему виду приведенная форма модели ничем не отличается от обычной системы независимых уравнений, параметры которой оцениваются традиционным М.Н.К.. Применяя к приведенной системе М.Н.К., можно оценить δ, а затем выразить значения эндогенных переменных через экзогенные переменные.

Коэффициенты приведенной формы модели представляют собой нелинейные функции от коэффициентов структурной формы модели. Рассмотрим это положение на примере простейшей структурной модели, выразив коэффициенты приведенной формы модели δ_{ij} через коэффициенты структурной модели a_{is} и b_{ik}. Для упрощения в модель не введены случайные переменные. Для структурной модели вида:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 \\ y_2 = b_{21}y_1 + a_{22}x_2, \end{cases}$$ приведенная форма имеет вид: $$\begin{cases} \hat{y}_1 = \delta_{11}x_1 + \delta_{12}x_2, \\ \hat{y}_2 = \delta_{21}x_1 + \delta_{22}x_2, \end{cases}$$

в которой y_2 можно выразить из первого уравнения структурной модели следующим образом: $y_2 = \dfrac{y_1 - a_{11}x_1}{b_{12}}$. Тогда система одновременных уравнений будет представлена как

$$\begin{cases} y_2 = y_1 - a_{11}x_1/b_{12} \\ y_2 = b_{21}y_1 + a_{22}x_2, \end{cases}$$ Отсюда имеем равенство $\dfrac{y_1 - a_{11}x_1}{b_{12}} = b_{21}y_1 + a_{22}x_2$, тогда $y_1 - b_{12}b_{21}y_1 = a_{11}x_1 + b_{12}a_{22}x_2$ или $y_1 = a_{11}x_1/(1 - b_{12}b_{21}) + b_{12}a_{22}x_2/(1 - b_{12}b_{21})$.

Таким образом, первое уравнение структурной формы модели представлено в виде уравнения приведенной формы модели:

$y_1 = \delta_{11}x_1 + \delta_{12}x_2$. Из данного уравнения следует, что коэффициенты приведенной формы модели представляют собой нелинейные соотношения коэффициентов структурной формы модели, т.е.

$\delta_{11} = a_{11}/(1 - b_{12}b_{21})$, $\delta_{12} = b_{12}a_{22}/(1 - b_{12}b_{21})$.

Аналогично можно показать, что коэффициенты приведенной (δ_{21}, δ_{22}) формы модели второго уравнения системы также нелинейно связаны с коэффициентами структурной формы модели. Для этого выразим переменную y_1 из второго структурного уравнения модели как $y_1 = \dfrac{y_2 - a_{22}x_2}{b_{21}}$ или, записав это выражение в левой части первого уравнения структурной формы модели, получим $(y_2 - a_{22}x_2)/b_{21} = b_{12}y_2 + a_{11}x_1$. Отсюда получаем $y_2 = \dfrac{a_{11}b_{21}}{1 - b_{21}b_{12}}x_1 + \dfrac{a_{22}}{1 - b_{21}b_{12}}x_2$, что соответствует уравнению приведенной формы модели: $y_2 = \delta_{21}x_1 + \delta_{22}x_2$ т.е. $\delta_{21} = a_{11}b_{21}/(1 - b_{21}b_{12})$, $\delta_{22} = a_{22}/(1 - b_{21}b_{12})$.

Эконометрические модели обычно включают в систему не только уравнения, отражающие взаимосвязи между отдельными переменными, но и выражения тенденции развития явления, а также разного рода тождества.

Так, в 1947 году, исследуя линейную зависимость потребления c от дохода y, Т. Хавельмо предложил одновременно учитывать тождество дохода. В этом случае модель имеет вид: $\begin{cases} c = a + by \\ y = c + x, \end{cases}$ где x - инвестиции в основной капитал и в запасы экспорта и импорта, a, b - параметры линейной зависимости эндогенной переменной c от эндогенной переменной y. Их оценки должны учитывать тождество дохода в отличие от параметров обычной линейной регрессии.

В этой модели две эндогенные переменные c, y и одна экзогенная переменная x. Систему приведенных уравнений представим следующим образом:

$$\begin{cases} c = A_0 + A_1x \\ y = B_0 + B_1x \end{cases}.$$

$x=(y-B_0)/B_1, c=A_0+A_1(y-B_0)/B_1$ $\quad c=A_0-A_1B_0/B_1+A_1y/B_1=a+by$, $\quad a=A_0-A_1B_0/B_1$, $b=A_1/B_1$. Она позволяет получить значения эндогенной переменной c через переменную x. Рассчитав коэффициенты модели (A_0,A_1,B_0,B_1) можно перейти к коэффициентам структурной модели a,b, подставляя в первое уравнение приведенной формы выражение переменной x из второго уравнения приведенной формы модели. Приведенная форма модели, хотя и позволяет получить значения эндогенной переменной через значения экзогенных переменных, аналитически уступает структурной форме модели, так как в ней отсутствуют оценки взаимосвязи между эндогенными переменными.

16. Проблема идентификации

При переходе от приведенной формы модели к структурной возникает проблема идентификации. *Идентификация – это единственность соответствия между приведенной и структурной формами модели.*

Рассмотрим проблему идентификации для случая с двумя эндогенными переменными. Пусть структурная модель имеет вид:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m, \end{cases}$$

где y_1 и y_2 - совместные зависимые переменные.

Из второго уравнения можно выразить y_1 следующей формулой $y_1=\frac{y_2}{b_{21}}-\frac{a_{21}}{b_{21}}x_1-\ldots-\frac{a_{2m}}{b_{21}}x_m$. Тогда в системе имеем два уравнения для эндогенной переменной y_1 с одним и тем же набором экзогенных переменных, но с разными коэффициентами при них:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_{m1}, \\ y_1 = y_2/b_{21} - a_{21}x_1/b_{21} - a_{22}x_2/b_{21} - \ldots - a_{2m}x_{m2}/b_{21} \end{cases}.$$

Наличие двух вариантов для расчета структурных коэффициентов одной и той же модели связано с неполной ее идентификацией. Структурная модель в полном виде, состоящая в каждом уравнении системы из n эндогенных и m экзогенных переменных, содержит $n(n-1+m)$ параметров. Так, если $n=2$ и $m=3$, полный вид структурной модели примет вид:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + a_{13}x_3, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3, \end{cases}.$$

Как видно, модель содержит восемь структурных коэффициентов, что соответствует выражению $n(n-1+m)$.

Приведенная форма модели в полном виде содержит nm параметров. Для последнего примера это означает наличие шести коэффициентов приведенной формы модели. В этом можно убедиться, обратившись к приведенной форме модели, которая будет иметь вид:

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3, \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3 \end{cases}.$$

Действительно, она включает в себя шесть коэффициентов δ_{ij}. На основе шести коэффициентов приведенной формы модели требуется определить восемь структурных коэффициентов рассматриваемой структурной модели, что, естественно, не может привести к единственному решению. В полном виде структурная модель содержит большее число параметров, чем приведенная форма модели. Соответственно $n(n-1+m)$

параметров структурной модели не могут быть однозначно определены из nm параметров приведенной формы модели.

Чтобы получить единственно возможное решение для структурной модели, необходимо предположить, что некоторые из структурных коэффициентов модели ввиду слабой взаимосвязи признаков с эндогенной переменной из левой части системы равны нулю. Тем самым уменьшится число структурных коэффициентов модели. Так, если предположить, что в нашей модели $a_{13} = 0$, $a_{21} = 0$, то структурная модель примет вид:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 \end{cases}.$$

В такой модели число структурных коэффициентов не превышает число коэффициентов приведенной модели, которое равно 6. Уменьшение числа структурных коэффициентов модели возможно и другим путем: например, путем приравнивания некоторых коэффициентов друг к другу, т.е. путем предположения, что их воздействие на формируемую эндогенную переменную одинаково. На структурные коэффициенты могут накладываться, например, ограничения вида $b_{ij} + a_{ij} = 0$.

С позиции идентифицируемости структурные модели можно подразделить на три вида:

идентифицируемые,

неидентифицируемые,

сверхидентифицируемые.

Модель идентифицируема, если все структурные ее коэффициенты определяются однозначно, единственным образом по коэффициентам приведенной формы модели, т.е. число параметров структурной модели равно числу параметров приведенной формы модели. В этом случае структурные коэффициенты модели оцениваются через параметры приведенной формы модели и модель идентифицируема. Рассмотренная выше структурная модель с двумя эндогенными и тремя экзогенными (предопределенными) переменными, содержащая шесть структурных коэффициентов, представляет собой идентифицируемую модель.

Модель неидентифицируема, если число приведенных коэффициентов меньше числа структурных коэффициентов, и в результате структурные коэффициенты не могут быть оценены через коэффициенты приведенной формы модели. Структурная модель в полном виде, содержащая n эндогенных и m предопределенных переменных в каждом уравнении системы, всегда неидентифицируема.

Модель сверхидентифицируема, если число приведенных коэффициентов больше числа структурных коэффициентов. В этом случае на основе коэффициентов приведенной формы модели можно получить два или более значений одного структурного коэффициента. В этой модели число структурных коэффициентов меньше числа коэффициентов приведенной формы. Так, если в структурной модели полного типа предположить нулевые значения коэффициентов $a_{13} = 0$, $a_{21} = 0$, и $a_{22} = 0$, тогда система уравнений станет сверхидентифицируемой:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{23}x_3 \end{cases}.$$

В этой системе пять структурных коэффициентов не могут быть однозначно определены из шести коэффициентов приведенной формы модели. Сверхидентифицируемая модель в отличие от неидентифицируемой модели практически решаема, но требует для этого специальных методов исчисления параметров.

Структурная модель всегда представляет собой систему совместных уравнений, каждое их которых требуется проверить на идентификацию. Модель считается идентифицируемой, если каждое уравнение системы идентифицируемо. Если хотя бы

одно из уравнений системы неидентифицируемо, то и вся модель считается неидентифицируемой. Сверхидентифицированная модель содержит хотя бы одно сверхидентифицируемое уравнение.

Выполнение условия идентифицируемости модели проверяется для каждого уравнения системы. *Чтобы уравнение было идентифицируемо, необходимо чтобы число предопределенных переменных, отсутствующих в данном уравнении, но присутствующих в системе, было равно числу эндогенных переменных в данном уравнении без одного.*

Если обозначить число эндогенных переменных в уравнении j через H, а число экзогенных (предопределенных) переменных, которые содержатся в системе, но не входят в данное уравнение, через D, то условие идентификации модели может быть записано в виде следующего правила:

$D+1=H$, *уравнение идентифицировано;*

$D+1<H$, *уравнение неидентифицировано;*

$D+1>H$, *уравнение сверхидентифицировано.*

Предположим, что рассматривается следующая система одновременных уравнений:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{33}x_3 + a_{34}x_4. \end{cases}$$

Первое уравнение этой системы точно идентифицировано, поскольку в нем присутствуют три эндогенные переменные - y_1, y_2, y_3, т.е. $H=3$, и две экзогенные переменные - x_1, x_2, число отсутствующих экзогенных переменных равно двум - x_3 и x_4, $D=2$. Тогда имеем равенство: $D+1=H$, т.е. $2+1=3$, что означает наличие идентифицируемого уравнения.

Во втором уравнении системы имеем, $H=2$ т.е. в уравнении присутствуют эндогенные переменные y_1 и y_2, а поскольку экзогенная переменная x_4 не входит в уравнение то, $D=1$, и тогда по счетному правилу, $D+1=H$ т.е. $1+1=2$, $D+1=H$, т.е. рассматриваемое уравнение идентифицируемо.

В третьем уравнении системы имеем, $H=3$ т.е. в уравнении присутствуют эндогенные переменные y_1, y_2, y_3; а поскольку экзогенные переменные x_1, x_2 не входят в уравнение, то $D=2$, и тогда по счетному правилу $D+1=H$, т.е. рассматриваемое уравнение идентифицируемо. Таким образом, идентифицируема и система в целом.

Теперь предположим, что в рассматриваемой модели $a_{21}=0$, $a_{33}=0$, тогда система примет вид:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{34}x_4. \end{cases}$$

Первое уравнение этой системы не изменилось. Система по-прежнему содержит три эндогенные и четыре экзогенные переменные, поэтому для него $D=2$ и $H=3$, и оно, как и в предыдущей системе, идентифицируемо. Второе уравнение имеет $H=2$ и $D=2$; переменные x_1, x_4 отсутствуют, следовательно $2+1>2$. Данное уравнение сверхидентифицируемо. Также сверхидентифицируемым оказывается и третье уравнение системы, в котором $H=3$; y_1, y_2, y_3, $D=3$; отсутствуют переменные x_1, x_2, x_3 т.е. счетное правило составляет неравенство: $3+1>3$ или $D+1>H$. Таким образом модель в целом является сверхидентифицируемой.

Допустим, что последнее уравнение системы имеет вид: $y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + a_{34}x_4$. В отличие от предыдущего уравнения в него включены еще две экзогенные переменные x_1, x_2, входящие в систему. В этом случае уравнение становится неидентифицируемым, ибо $H = 3$ и $D = 1$, $D + 1 < H$, а $1 + 1 < 3$. Итак, несмотря на то, что первое уравнение идентифицировано, поскольку второе уравнение сверхидентифицировано, а третье уравнение неидентифицировано, то и вся модель считается неидентифицируемой и не имеет статистического решения.

Для того, чтобы оценить параметры структурной модели, система должна быть идентифицируема или сверхидентифицируема.

Рассмотренное счетное правило является необходимым, но не достаточным условием идентифицируемости. Более точно условие идентифицируемости определяется, если накладывать ограничения на коэффициенты матриц параметров структурной модели. *Уравнение идентифицируемо, если по отсутствующим в нем переменным (эндогенным и экзогенным) можно из коэффициентов при других уравнениях системы получить матрицу, определитель которой не равен нулю, а ранг матрицы не меньше, чем число эндогенных переменных в системе без одного*.

Целесообразность проверки условия идентификации модели через определитель матрицы коэффициентов, отсутствующих в данном уравнении, но присутствующих в других, объясняется тем, что возможна ситуация, когда для каждого уравнения системы выполнено счетное правило, а определитель матрицы названных коэффициентов равен нулю. В этом случае соблюдается лишь необходимое условие идентификации, в то время как не выполнено достаточное условие идентификации.

Рассмотрим следующую структурную модель:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \end{cases}$$

Проверим каждое уравнение системы на необходимое и достаточное условие идентифицируемости. Для первого уравнения имеем $H = 3$; y_1, y_2, y_3, и $D = 2$; переменные x_3, x_4 отсутствуют, т.е. $D + 1 = H$ и необходимое условие идентификации выполнено, а потому уравнение точно идентифицируемо. Для проверки достаточного условия идентифицируемости, заполним следующую таблицу коэффициентов при отсутствующих в первом уравнении переменных. Определитель матрицы коэффициентов которой равен нулю.

Уравнения	Переменные	
	x_3	x_4
2	a_{23}	a_{24}
3	0	0

Для второго уравнения $H = 2$; y_1, y_2 и $D = 1$ (x_1 отсутствует), счетное правило дает утвердительный ответ: уравнение идентифицируемо $D + 1 = H$. Выполняется и достаточное условие идентификации. Коэффициенты при отсутствующих во втором уравнении переменных составят:

Уравнения	Переменные	
	y_3	x_1

2	b_{13}	a_{11}
3	-1	a_{31}

Согласно таблице, $\det A \neq 0$, ранг матрицы равен 2, что соответствует следующему критерию: ранг матрица коэффициентов должен быть не меньше чем число эндогенных переменных в системе без одного, следовательно, второе уравнение точно идентифицируемо.

Для третьего уравнения системы $H = 3$ и $D = 2$, т.е. по необходимому условию идентификации оно точно идентифицируемо $D + 1 = H$. Тем не менее, достаточное условия идентификации, при проверке его на выполнимость, дает отрицательный ответ. Составим по отсутствующим в третьем уравнении переменным таблицу коэффициентов, из которой следует, что $\det A \neq 0$:

Уравнения	Переменные	
	x_3	x_4
2	0	0
3	a_{23}	a_{24}

Из таблицы видно, что достаточное условие идентификации не выполняется, следовательно, уравнение не идентифицируемо. А потому, и рассматриваемая структурная модель в целом не идентифицируема, поскольку, будучи выполнено необходимое условие идентификации, нарушено достаточное условие.

В эконометрических моделях часто наряду с уравнениями, параметры которых должны быть статистически оценены, используются балансовые тождества переменных, коэффициенты при которых равны ± 1. В этом случае, хотя само тождество и не требует проверки на идентификацию, ибо коэффициенты при переменных в тождестве известны, в проверке на идентификацию собственно структурных уравнений системы тождества участвуют.

Например, рассмотрим эконометрическую модель экономики страны:

$$\begin{cases} y_1 = A_{01} + b_{13}y_3 + b_{14}y_4 + \varepsilon_1 \\ y_2 = A_{02} + b_{23}y_3 + a_{21}x_1 + \varepsilon_2, \\ y_3 = A_{03} + b_{34}y_4 + a_{31}x_1 + \varepsilon_3 \\ y_4 = y_1 + y_2 + x_2. \end{cases},$$

где y_1 - расходы на конечное потребление данного года; y_2 - валовые инвестиции в текущем году; y_3 - расходы на заработную плату в текущем году; y_4 - валовой доход за текущий год; x_1 - валовой доход предыдущего года; x_2 - государственные расходы текущего года; A_{0i} - свободный член i- ого уравнения; ε_i - случайная ошибка i- ого уравнения. В этой модели четыре эндогенные переменные y_1, y_2, y_3, y_4, причем переменная y_4 задана тождеством. Поэтому практически статистическое решение необходимо только для первых трех уравнений системы, которые необходимо проверить на идентификацию. Модель содержит две предопределенные переменные – одну временную переменную x_2 и одну лаговую переменную x_1.

При практическом решении задачи на основе статистической информации за ряд лет или по совокупности регионов за один год в уравнениях для эндогенных переменных y_1, y_2, y_3 обычно содержится свободный член A_{0i}, $i = \overline{1,3}$, значение которого аккумулирует влияние неучтенных в уравнении факторов и не влияет на определение идентифицируемости модели.

Поскольку фактические данные об эндогенных переменных y_1, y_2, y_3 могут отличаться от теоретических постулируемых моделью, то принято включать в модель случайную составляющую для каждого уравнения системы, за исключением тождеств. Случайные составляющие (возмущения), обозначенные как ε_i не влияют на решение проблемы идентифицируемости модели.

В рассматриваемой эконометрической модели первое уравнение системы идентифицируемо, ибо $H = 3$ и $D = 2$, и выполняется необходимое условие идентификации $D + 1 = H$. Кроме того, выполняется и достаточное условие идентификации, т.е. ранг матрицы равен 3, а определитель ее не равен 0: $\det A \neq 0$

Уравнения	y_2	x_1	x_2
2	-1	a_{21}	0
3	0	a_{31}	0
4	1	0	1

Второе уравнение системы так же точно идентифицируемо: $H = 2$ и $D = 1$, т.е. выполнено счетное правило $D + 1 = H$, также выполнено достаточное условие идентификации, поскольку ранг матрицы равен 3, а определитель ее не равен нулю $\det A = -b_{34}$.

Уравнения	y_1	y_4	x_2
1	-1	b_{14}	0
3	0	b_{34}	0
4	1	-1	1

Аналогично идентифицируемо третье уравнение системы, поскольку $H = 2$ и $D = 1$, т.е. выполнено счетное правило $D + 1 = H$, также выполнено достаточное условие идентификации, поскольку. ранг матрицы равен 3, а определитель ее не равен 0: $\det A = 1$.

Уравнения	y_1	y_2	x_2
1	-1	0	0
2	0	-1	0
4	1	1	1

Идентификация уравнений достаточно сложна и не ограничивается тем, что было изложено ранее. На структурные коэффициенты модели могут накладываться и другие ограничения, например, в производственной функции сумма эластичностей может быть равна единице по предположению. Могут накладываться ограничения на дисперсии и ковариации остаточных величин.

Коэффициенты структурной модели могут быть оценены разными способами в зависимости от вида системы одновременных уравнений. Наибольшее распространение в литературе получили следующие методы оценивания коэффициентов структурной модели:

– косвенный метод наименьших квадратов;
– двух шаговый метод наименьших квадратов;
– трех шаговый метод наименьших квадратов;
– метод максимального правдоподобия с полной информацией;
– метод максимального правдоподобия при ограниченной информации.

Косвенный метод наименьших квадратов (КМНК) применяется для идентифицированной системы совместных уравнений, а двух шаговый метод наименьших квадратов (ДМНК) используется для оценки коэффициентов сверхидентифицированной модели.

Перечисленные методы оценивания используются и для сверхидентифицированной системы уравнений.

Метод максимального правдоподобия рассматривается как наиболее общий метод оценивания, результаты которого при нормальном распределении признаков, совпадают с М.Н.К.. Однако при большом числе уравнений системы этот метод приводит к достаточно сложным вычислительным процедурам. Поэтому в качестве модификации используется метод максимального правдоподобия при ограниченной информации (метод наименьшего дисперсионного отношения).

В отличие от метода максимального правдоподобия в этом методе сняты ограничения на параметры, связанные с функционированием системы в целом. Это делает решение более простым, но трудоемкость вычислений остается достаточно высокой. Несмотря на его значительную популярность к середине 60-х годов он был практически вытеснен двухшаговым методом наименьших квадратов (Д.М.Н.К.) в связи с гораздо большей его простотой.

Дальнейшим развитием Д.М.Н.К. является трехшаговый метод наименьших квадратов (Т.М.Н.К.). Этот метод оценивания пригоден для всех видов уравнений структурной модели. Однако при некоторых ограничениях на параметры более эффективным оказывается Д.М.Н.К.

Б И Б Л И О Г Р А Ф И Я

1. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики: Учебник. ЮНИТИ, Москва, 1998.
2. Джонстон Дж. Эконометрические методы. Статистика, Москва, 1980.
3. Эконометрика. Под редакцией члена-корреспондента Российской Академии Наук И. И. Елисеевой. Финансы и статистика, Москва, 2001, 343 с.
4. dr. Nicos Economou. An Estimation of the Potential Output and the Output Gap of the Moldovan Economy.Chisinau, MER, TACIS. Moldovan Economic Trends, 2002, q.4., pp.85-93.
5. Маленво Э. Статистические методы эконометрии. Статистика, Москва, 1976.
6. Моделирование экономических процессов. Под редакцией доктора экономических наук, профессора М. В. Грачевой, кандидата физико-математических наук, доцента Л. Н. Фадеевой, доктора экономических наук, профессора Ю. Н. Черемных. ЮНИТИ, Москва, 2005, 353 с.
7. dr. Apostolos Papaphilippou. An Econometric Estimation of the Import Demand in Moldova. Chisinau, MER, TACIS. Moldovan Economic Trends, 2001, q.3., pp.93-99.
8. Ion Pârţachi, Alexandru Brailă, Natalia Şişcanu. Econometrie Aplicată. A.S.E.M., Chişinău, 1999, 172 p.
9. Pecican E., Econometrie. Editura All, Bucureşti, 1994.
10. Schatteles T. Metode econometrice moderne, Universitas, Chişinău, 1992.
11. A. H. Studenmund. Using Econometrics (second edition). Washington. HarperCollins Publishers Inc. 1992, 662 p.
12. Титнер Г. Введение в эконометрию. Статистика, Москва, 1978.
13. Четыркин Е. М. Статистические методы прогнозирования. Статистика, Москва, 1977.
14. Zaman C. Econometrie, Bucureşti, 1998.

П Р И Л О Ж Е Н И Я

ЛАБОРАТОРНАЯ РАБОТА № 1
ЭКОНОМЕТРИЧЕСКАЯ ОЦЕНКА ФУНКЦИИ СПРОСА НА ИМПОРТ

Этап I. *Статическая регресия*

Оценить коэффициенты регрессии для функции спроса на импорт, представленной в виде полной логарифмической функции:

$\boldsymbol{lnM_t = \beta_0 + \beta_1 lnY_t + \beta_2 lnP_t}$, unde (1),

M_t– квартальные данные для реальных объемов импорта, представленные в $ SUA,

Y_t– квартальные данные для реального ВВП, представленные в $ SUA,

P_t – относительная квартальная цена, являющаяся отношением индекса мировых цен к индексу внутренних потребительских $P_t=e_t*P_t^{*M}/IPC_t$.

P_t^{*M} – индекс мировых цен на импорт, квартальные данные;

e_t - обменный курс, нормированный по первому кварталу 1996;

IPC_t – индекс потребительских цен, квартальные данные.

Предположим, что:

a) проконсультирована теория относительно аналитического представления функции спроса на импорт;
b) специфицирована модель;
c) исследованы знаки при коэффициентах специфицированной модели;
d) собрана необходимая информация для выполнения регрессионного анализа;
e) был использован метод наименьших квадратов для получения коэффициентов и оценки предложенной (статистики Stiudent, коэффициент детерминации, статистика Fişer);
f) составлен отчет.

Введем двоичную переменную $Dummy_t$ которая учтет влияние финансового кризиса в России, 1998 года и которая имеет вид:

$$\boldsymbol{Dummy_t = 1 \begin{cases} IIItr.\ 1998,\ IVtr.\ 1998,\ Itr.\ 1999,\ IItr.\ 1999, \\ 0 \ \text{pentru trimestrele rămase.} \end{cases}}$$

Далее, исследуем в качестве функции спроса на импорт полулогарифмическую форму следующего вида:

$\boldsymbol{lnM_t = \beta_0 + \beta_1 lnYt + \beta_2 lnP_t + \beta_3 Dummy_t}$ (2),

Необходимо выполнить этапы a)-f).

Этап II. *Динамическая регрессия*

Оценить коэффициенты регрессии для функции спроса на импорт, представленную полулогарифмической формой с лаговыми переменными:

$\boldsymbol{lnM_t = \beta 0 + \beta 1 lnYt + \beta 2 lnPt + \beta 3 Mt\text{-}1}$ (3),

Необходимо выполнить этапы a)-f).

И, наконец, необходимо оценить коэффициенты регрессии для функции спроса на импорт, заданной полулогарифмической формой с запаздывающей переменной и двоичной переменной $Dummy_t$:

$\boldsymbol{lnMt = \beta_0 + \beta_1 lnY_t + \beta_2 lnPt + \beta_3 M_{t\text{-}1} + \beta_3 Dummy_t}$ (4).

Первый квартал 1996 года выбирается в качестве базового. Используя обменный курс, рассчитается реальный Валовой Внутренний Продукт. При вычислении индекса потребительских цен будут использованы данные по квартальной инфляции. Значения реальных показателей вычисляются с использованием ИПЦ. $M_t\$real=M_t\$nom./IPC$; $Y_t\$real=Y_t\$nom./IPC$. $IPC_t=\prod_{k=tr.b}^{n} IPC_k*(1+infl._{k+1}/100)$.

Необходимая информация может быть найдена на www.statistica.md .

ЛАБОРАТОРНАЯ РАБОТА № 2
ОЦЕНКА ВАЛОВОГО ВНУТРЕННЕГО ПРОДУКТА МОЛДОВЫ СО СТОРОНЫ СПРОСА

Этап I. *Оценка потребления как функции от ВВП*

Оценить потребление по следующей формуле:

$C_t = cY_{t-1} + C_0$ (1),

где C_t - потребление в году t , Y_{t-1} - валовой внутренний продукт в году t , C_0 - начальное потребление, а c – предельная склонность к потреблению, $c = (\sum_{1995}^{2003} C_t / Y_t)/9$.

Валовой Внутренний Продукт определяется по формуле

$Y_t = C_t + I_t = cY_{t-1} + C_0 + I_t$ (2),

здесь I_t - валовые инвестиции в году t, предположим, что $I_t = I_0$.

1. Используя статистическую информацию за последние годы оцените коэффициент c - предельную склонность к потреблению.
2. Задать I_0

Информацию можно найти на сайте : http://www.statistica.md.

Этап II. *Вычислить ВВП в году* t

1. Оценить ВВП со стороны спроса по формуле:

$$Y_t = Y_{t_0} c^t + \overline{A} \sum_{l=t_0}^{l=t-1} c^l$$

здесь $\overline{A} = C_0 + I_0$ - независимые затраты. В качестве $C_0 = 0$, $I_0 = I_{t_0}$, $t_0 = 2003$ год, а $t_k = 2008$, $t_k = 2012$.

2. Взяв за базу 2001 год, пересчитать Y_t в постоянных ценах 2001 года, предположив, что, начиная с 2005, года инфляция будет принимать значение 7,5% ежегодно, а в 2004 она равна, предположительно 7,5%

Таб.1 Исторические данные

Годы	1995	1996	1997	1998	1999	2000	2001	2002	2003
ВВП$t_{nomlefec}$	**6480**	**7798**	**8917**	**9122**	**12322**	**16020**	**19052**	**22556**	**27297**
ВВП$t_{реалефект}$									
Инфляция$_t$	0,3	0,24	0,12	0,08	0,39	0,31	0,10	0,05	0,012
Дефл.ВВП$_{2001}$									
Годы	2004	2005	2006	2007	2008	2009	2010	2011	2012
ВВП$t_{номиналпрогноз}$	**6480**	**7798**	**8917**	**9122**	**12322**	**16020**	**19052**	**22556**	**27297**
Инфляция$_t$	0,075	0,075	0,075	0,07	0,065	0,06	0,055	0,05	0,05
ВВП$t_{реалефект}$									

ЛАБОРАТОРНАЯ РАБОТА № 3
ОЦЕНКА ЭКОНОМИЧЕСКОГО ПОТЕНЦИАЛА МОЛДОВЫ

Этап I.. *Использование фильтра Hodric-Prescot*

Оценить тренд потенциального реального ВВП используя фильтр Hodrick-Prescott (HP). Эту оценку можно получить, если определить тренд соответствующий потенциальному реальному ВВП, который минимизирует одновременно средневзвешенную разность между наблюдаемыми значениями реального ВВП и оцененными и разностью между оцененным ВВП для любого момента времени. Это равносильно определению минимума целевой функции следующего вида:

$$\sum\left(\ln Y_t - \ln Y_t^*\right)^2 + \lambda\sum\left[\left(\ln Y_{t+1}^* - \ln Y_t^*\right) - \left(\ln Y_t^* - \ln Y_{t-1}^*\right)\right]^2 \qquad (1),$$

здесь $\ln Y_t$ и $\ln Y_t^*$ являются логарифмами реального ВВП и оцененного ВВП соответственно. $\sum\left(\ln Y_t - \ln Y_t^*\right)^2$ определяет сумму квадратов отклонений между логарифмами реального наблюдаемого ВВП - $\ln Y_t$ и тренда соответствующего оцененному ВВП $\ln Y_t^*$. $\lambda\sum\left[\left(\ln Y_{t+1}^* - \ln Y_t^*\right) - \left(\ln Y_t^* - \ln Y_{t-1}^*\right)\right]^2$ представляет штрафную функцию, определяющую штраф за отклонение суммы квадратов отклонений от темпа прироста соответствующих компонент тренда.

Используя процедуру Eviews оценить тренд потенциального реального ВВП для следующих значений λ=10;30;100.

Y_t– годовой воловий внутренний продукт, выраженный в MDL для периода 1995-2002гг.

Информацию можете найти на сайте: http://www.statistica.md.

Результаты должны быть представлены как в виде графиков, так и в виде таблиц.

Этап *II*. *Формулировка проблемы с помощью производственной функции*

Оценить коэффициенты регрессии для производственной функции типа Кобба-Дугласа, зависящей от двух производственных факторов – капитала K_t и труда, при этом подразумевается, что процесс характеризуется постоянной фондоотдачей $Y_t = K^{1-\alpha}N^{\alpha}$ или в логарифмической форме $\ln Y_t = \beta_1 \ln K_t + \beta_2 \ln N_t$

1. Оценить производственную функцию, заданную в виде полной логарифмической формы для значений эффективного реального ВВП для значений потенциального реального ВВП.
2. Все показатели, участвующие в расчетах , должны быть представлены в постоянных ценах.
3. Используя оцененные производственные функции выполнить прогноз на 2011-2015гг., будучи заданы годовые темпы прироста капитала 5%,10%,15%,20% и 25%, в то время как годовой темп прироста трудовых ресурсов составит 10% .
4. Результаты представить как в графической, так и в табличной форме.

Tab.1 Исторические данные по исследуемой проблеме

Годы	1995	1996	1997	1998	1999	2000	2001	2002	2003
ВВП$t_{номинюэфект}$	**6480**	**7798**	**8917**	**9122**	**12322**	**16020**	**19052**	**22556**	**27297**
ВВП$t_{реальнюэфект}$									
Инфляция$_t$	0,3	0,24	0,12	0,08	0,39	0,31	0,10	0,05	0,012
Дефл.ВВП$_{1995}$									
Kt$_{номин.}$	**14450**	**16138**	**14743**	**24702**	**30926**	**33598**	**34325**	**35827**	**37782**
Kt$_{реальн.}$									
Lt$_{номинальн.}$	**2800**	**3623**	**4153**	**4689**	**5207**	**7108**	**9322**	**12729**	**18508**
Lt$_{реальн.}$									
PIBt$_{потенциальн.}$	**6574**	**7879**	**8026**	**11126**	**13150**	**15941**	**18336**	**21802**	**26824**
PIBt$_{пот.\ реальн.}$									

ЛАБОРАТОРНАЯ РАБОТА № 4
ОЦЕНКА ВАЛОВОГО ВНУТРЕННЕГО ПРОДУКТА МОЛДОВЫ СО СТОРОНЫ СПРОСА

Этап I. *Оценка потребления как функции от ВВП*

Оценить потребление по следующей формуле:

$C_t = cY_{t-1} + C_0$ (1),

где C_t - потребление в году t , Y_{t-1} - валовой внутренний продукт в году t , C_0 - начальное потребление, а c – предельная склонность к потреблению, $c = (\sum_{1995}^{2003} C_t / Y_t)/9$.

Валовой Внутренний Продукт определяется по формуле

$Y_t = C_t + I_t = cY_{t-1} + C_0 + I_t$ (2),

здесь I_t - валовые инвестиции в году t, предположим, что $I_t = I_0$.

3. Используя статистическую информацию за последние годы оцените коэффициент c - предельную склонность к потреблению.
4. Задать I_0

Информацию можно найти на сайте : http://www.statistica.md.

Этап II. *Вычислить ВВП в году* t

5. Оценить ВВП со стороны спроса по формуле:

$$Y_t = Y_{t_0} c^t + \overline{A} \sum_{l=t_0}^{l=t-1} c^l$$

здесь $\overline{A} = C_0 + I_0$ - независимые затраты. В качестве $C_0 = 0$, $I_0 = I_{t_0}$, $t_0 = 2003$ год, а $t_k = 2008$, $t_k = 2012$.

6. Взяв за базу 2001 год, пересчитать Y_t в постоянных ценах 2001 года, предположив, что, начиная с 2005, года инфляция будет принимать значение 7,5% ежегодно, а в 2004 она равна, предположительно 7,5%

Таб.1 Исторические данные

Годы	1995	1996	1997	1998	1999	2000	2001	2002	2003
ВВП$t_{nomlefec}$	**6480**	**7798**	**8917**	**9122**	**12322**	**16020**	**19052**	**22556**	**27297**
ВВП$t_{реалефект}$									
Инфляция$_t$	0,3	0,24	0,12	0,08	0,39	0,31	0,10	0,05	0,012
Дефл.ВВП$_{2001}$									
Годы	2004	2005	2006	2007	2008	2009	2010	2011	2012
ВВП$t_{номиналпрогноз}$									
Инфляция$_t$	0,075	0,075	0,075	0,07	0,065	0,06	0,055	0,05	0,05
ВВП$t_{реалефект}$									

Этап III. *Оценка экономического роста по модели Domar и Harrod-Domar*

1. Вычислить ВВП на перспективу 2004-2012 по следующей формуле:

$$Y_t = (1+\sigma \times s)^t Y_{t_0}$$

где σ прирост производства на одну единицу инвестиций, а s - норма накоплений, $s = 1 - c$, здесь s - предельная склонность к потреблению, Y_{t_0} - валовой внутренний продукт в году t_0. Используйте значение c, определенное ранее при вычислении s; вычислите $\sigma = \max_t (I_t / Y_t)$ за период 1995-2002.

2. Вычислите ВВП на перспективу 2004-2012 в соответствии с формулой:

$$Y_t = (1+\rho_g)^t Y_{t_0}$$

в которой гарантированный темп роста ВВП определяется как $\rho_g = \dfrac{s}{v-s}$. Здесь v определяется по принципу ускорения, а, именно, $I_t = v \times (Y_t - Y_{t-1})$, Y_t - валовой внутренний продукт в году t, а Y_{t-1} - валовой внутренний продукт в году $t-1$, s - норма накопления, $s = 1 - c$, сейчас, как и ранее, c определяет предельную склонность к потреблению, Y_{t_0} - валовой внутренний продукт в году t_0.

Вычислите $v = \max_t I_t / (Y_t - Y_{t-1})$ за период 1995-2002.

Этап IV. *Оценка номинальной и реальной заработной платы, используя производственную функцию Cobb-Douglas.*

1. Имея коэффициенты производственной функции $Y_t = K^{1-\alpha} L^{\alpha}$, оцененные в предыдущих работах, вычислите реальную заработную плату в виде $w_t = \partial Y_t / \partial L_t = \alpha \times K^{1-\alpha} L^{\alpha-1}$, а номинальную заработную плату - $w_t nom = w_t Defl.PIB_t$, используя дефлятор ВВП - $Defl.PIB_t$, определенный на предыдущих этапах.
2. Сохраняя темпы прироста прогнозного ВВП, капитала и труда, указанные в предыдущей лабораторной работе, определите прогнозные значения для номинальной и реальной заработной платы.

Таблица 2. Исторические данные

ОСНОВНЫЕ МАКРОЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ								
	1995	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*
Валовой Внутренний Продукт (ВВП)								
млн.лей	6480	7798	8917	9122	12322	16020	19052	22556
Конечное Потребление (С)								
млн.лей	5371	7356	8681	9203	11090	16503	19263	23289
Валовое Накопление Капитала								
млн.лей	1612	1891	2123	2360	2820	3836	4436	4886
Инвестиции в Основной Капитал								
млн.лей	844,8	987,4	1202,2	1444,4	1591,8	1759,3	2315,1	2804,2

ЛАБОРАТОРНАЯ РАБОТА № 5
ЛИКВИДАЦИЯ ГЕТЕРОСКЕДАСТИЧНОСТИ

Nr.	Цена Y_i (lei)	Объем X_i	$Y_{iоцен.}$	u_i	u_i*u_i	$\ln(u_i*u_i)$	$\ln(Z_i)$	Y_i/X_i
1	4,6	2						
2	4,6	2						
3	4,7	2						
4	5	4						
5	5,2	4						
6	5,4	4						
7	5,6	6						
8	5,8	6						
9	5,9	6						
10	6,3	8						
11	6,5	8						
12	6,8	8						
13	7,4	10						
14	7,6	10						
15	7,1	10						
16	4,8	2						
17	5	2						
18	5,1	2						
19	5,4	4						
20	5,5	4						
21	5,6	4						
22	6,3	6						
23	6,3	6						
24	6,4	6						
25	7,2	8						
26	7,4	8						
27	7,5	8						
28	8,2	10						
29	8,4	10						
30	8,8	10						
Total								
Media								

1. Запустить уравнение регрессии. Переменную X объявить фактором пропорциональности

Применить тест Park для выявления гетеросчедастичности.

2. Вычислить $\ln((u_i)^2)$ и $\ln(Zi)$ d_L и запустить новое уравнение регрессии.

3. Сравнить t-статистику переменной Z с $t_{табл(30-2;,0,05)}$.

a) $t_{выч}$ > ttабл(30-2;,0,05), остаточный член гетеросчедастичен.

Выполнить преобразования переменных $У*_i=Y_i/X_i$; X*i=1/Xi.

4. Оценить преобразованное уравнение регрессии.

Проверить значимость коеффициентов: детерминации, кореляции, t-статистики, F-статистика.

b) tвыч < ttабл(30-2;,0,05), остаточный член не является гетеросчедастичным.

Запуск исходного уравнения регрессии

Y = bo + b1*X

Запуск преобразованного уравнения регрессии

Y* = Y/X = bo* 1/X + b1

Y = bo + b1*X

ЛАБОРАТОРНАЯ РАБОТА № 6
ЛИКВИДАЦИЯ АВТОКОРЕЛЛЯЦИИ ОСТАТКОВ

Год	Потребление Y_t	Доход X_t	$Y_{тоцен}$	u_t	u_t*u_t	u_t-u_{t-1}	$ut*u_{t-1}$	$u_{t-1}*u_{t-1}$	Y_t	X_t	$Y_{тоцен}$	u_t	$(u*_t)^2$	u_t-u_{t-1}
1	84,4	88,0												
2	91,9	94,0												
3	99,2	100,0												
4	104,0	106,0												
5	109,0	110,0												
6	117,8	119,0												
7	122,9	127,0												
8	130,0	135,0												
9	138,7	143,0												
10	149,1	155,0												
11	158,0	167,0												
12	167,5	177,0												
13	177,8	186,0												
14	186,6	197,0												
15	195,7	211,0												
16	208,6	228,0												
17	221,5	239,0												
18	232,1	252,0												
Total														
Media														

$DW*_2$	d_U	d_L
1,39	1,16	

Y_t-ro*Y_{t-1}=b0+b1*(X_t-ro*X_{t-1})
Y_t=b0+b1*X_t+ro*Y_{t-1}-b1*ro*X_{t-1}

1. Запустить уравнение регрессии. Вычислить статистику DW =Sum(u_t-u_{t-1})^2/Sum(u_t)^2. Обратить внимание, что сумма с запаздывающим членом содержит на одно слагаемое меньше.
2. Определить табличные значения d_U = DW(n;k;0,05) и d_L = DW(n;k;0,05), (N-количество наблюдений, k-количество независимых переменных, k-количество независимых переменных) и сравнить с рассчитанной статистикой DW.
a) DW < d_U, остаточный член автокорелирован.
Вычислить ro=Sum(u_t*u_{t-1})/Sum((u_{t-1}))^2
Преобразовать переменные по формулам $У*_1=(1-ro)^{1/2}*У_1$; $У*_t=Y_t-Y_{t-1}*ro$;
$X*_1=(1-ro)^{1/2}*X_1$; $X*_t=X_t-X_{t-1}*ro$

3. Оценить преобразованное уравнение регрессии.
Пересчитать статистику DW =Sum(u_t-u_{t-1})^2/Sum(u_t)^2.
Выполнить этап 2.

b) DW < d_U автокореляция остаточного члена отсутствует.

ЛАБОРАТОРНАЯ РАБОТА № 7
ЛИКВИДАЦИЯ МУЛЬТИКОЛЛИНЕАРНОСТИ

Nr.	Y	X_1	X_2	X_3	X_4	SumX_i
1	74,3	1,0	29,0	15,0	52,0	
2	72,5	1,0	31,0	22,0	44,0	
3	83,8	1,0	40,0	23,0	34,0	
4	93,1	2,0	54,0	18,0	22,0	
5	102,7	3,0	71,0	17,0	6,0	
6	78,5	7,0	26,0	6,0	60,0	
7	95,9	7,0	52,0	6,0	33,0	
8	109,4	10,0	68,0	8,0	12,0	
9	104,3	11,0	56,0	8,0	20,0	
10	87,6	11,0	31,0	8,0	47,0	
11	109,2	11,0	55,0	9,0	22,0	
12	113,3	11,0	66,0	9,0	12,0	
13	115,9	21,0	47,0	4,0	26,0	

1. Убедиться, что уравнение регрессии подвержено мультиколлинеарности.
2. Вычислить суму всех 4-х переменных для каждого наблюдения.
3. Запустить уравнение регрессии. Вычислить t-статистики bi=bi/sigma(bi).
   Вычислить значение частного критерия $Fi=(t_{bi})^2$
4. Определить наименьшее значение F_i, равное $F_L=\min(F_i)$, и сравнить с табличным значением F(1;n-m-1;alfa) (n-количество наблюдений; m-количество независимых переменных; alfa-заданный уровень вероятности - 0,05).
   Перейти к 5-ому этапу. Возможны варианты:
   a) $F_L < F_0$. Независимая переменная исключается из уравнения..
   b) $F_L > F_0$. В этом случае, полученная модель является искомой.
5. Оценить уравнение регрессии по оставшимся независимым переменным.
   Выполнить этапы 1-4.

Y	X_1	X_2	X_4	**$Y = bo + b1*X_1 + b2*X_2 + b3*X_3 - b4*X_4$**					
74,3	1,0	29,0	52,0	**sigma(bi)**					
72,5	1,0	31,0	44,0	t_{bi}					
83,8	1,0	40,0	34,0	F_i	i =	F_i	<	F_{itabel}	5,32
93,1	2,0	54,0	22,0	F_L					
102,7	3,0	71,0	6,0						
78,5	7,0	26,0	60,0						
95,9	7,0	52,0	33,0						
109,4	10,0	68,0	12,0	**$Y = bo + b1*X_1 + b2*X_2 - b3*X_4$**					
104,3	11,0	56,0	20,0	**sigma(bi)**					
87,6	11,0	31,0	47,0	t_{bi}					
109,2	11,0	55,0	22,0	F_i	i =	F_i	<	F_{itabel}	5,12
113,3	11,0	66,0	12,0	F_L					
115,9	21,0	47,0	26,0						
				$Y = b0 + b1*X_1 + b2*X_2$					
				sigma(b_i)					
				t_{bi}					
				F_i		F_i	<	F_{itabel}	4,96
				F_L					