

Введение

Данное пособие появилось как результат факультатива и спецкурса, прочитанных автором для студентов экономического факультета Новосибирского университета в 1996 г. Пособие состоит из двух самостоятельных разделов.

Раздел I основан на факультативе “Некоторые эконометрические методы” (совместно с Н. Ибрагимовым). Факультатив предназначался в основном для студентов 2-го курса, которые только начинали слушать вводный курс эконометрии. Поэтому для понимания раздела не требуется серьезного знакомства с эконометрической теорией. Исключение составляют дополнительные параграфы, посвященные популярной в настоящее время теме единичных корней и коинтеграции.

Раздел II представляет собой переработанный материал спецкурса “Метод максимального правдоподобия в эконометрии”. Это цельный продвинутый курс, рассчитанный на студентов, хорошо знакомых с классическими эконометрическими методами. Метод максимального правдоподобия составляет теоретическую основу большей части эконометрии. Знание его необходимо для понимания современной экономической литературы. В этом пособии продемонстрировано применение ММП к некоторым базовым видам моделей, что позволяет познакомиться с возможностями метода и научиться основным приемам. Главное внимание уделено теории тестирования и методов оценивания. Полученные навыки должны помочь, если такая необходимость возникнет в ходе исследований, самостоятельно разрабатывать методы оценивания и тестирования для моделей других видов.

Основой для изложения метода максимального правдоподобия и его применений послужила книга R. Davidson & J.G. MacKinnon, *Estimation and Inference in Econometrics*. Многие подходы и обозначения совпадают. Данное пособие, однако, не является простым переложением этой книги. Книга Дэвидсона и Мак-Киннона предназначена для изучения курса эконометрии в целом, а пособие делает акцент именно на одном этом методе. Кроме того, при написании пособия использованы другие учебники и оригинальные статьи из научных журналов. Поэтому рассматривается ряд тем и методов, которые отсутствуют у Дэвидсона и Мак-Киннона. Материал изложен так, как это было удобнее с точки зрения целей данного пособия. Доказательства основных свойств оценок ММП (состоятельности, асимптотической эффективности и асимптотической нормальности) не приводятся. Их можно найти в учебниках по математической статистике.

Хотя второй раздел ни в коем случае не претендует на математическую строгость, однако является гораздо менее простым, чем первый раздел. Широко используется аппарат матричной алгебры и матричного анализа. Используемые правила матричных операций особо выделены в тексте раздела.

В целом пособие дополняет имеющуюся на русском языке литературу по эконометрическим методам.

Оглавление

I. НЕКОТОРЫЕ ЭКОНОМЕТРИЧЕСКИЕ МЕТОДЫ	5
ФУНКЦИОНАЛЬНАЯ ФОРМА РЕГРЕССИОННОЙ МОДЕЛИ	6
• Тестирование правильности спецификации регрессионной модели	6
• Линейные и нелинейные модели	7
• Выбор между альтернативными функциональными формами	11
ФИКТИВНЫЕ ПЕРЕМЕННЫЕ КАК РЕГРЕССОРЫ	13
• Общие соображения	13
• Использование фиктивных переменных для проверки однородности наблюдений и прогнозирования	15
• Использование фиктивных переменных в моделях с временными рядами	17
• Спектральный анализ и регрессия	20
МОДЕЛИ С КАЧЕСТВЕННОЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ	20
• Модели с бинарной зависимой переменной	20
• Модель выбора. Пробит и логит	21
• Оценка качества модели и проверка гипотез	23
• Множественные модели с качественными зависимыми переменными	25
ДИНАМИЧЕСКАЯ СПЕЦИФИКАЦИЯ РЕГРЕССИОННОЙ МОДЕЛИ	27
• Модель распределенного лага	28
• Динамические регрессионные модели. Авторегрессионная модель с распределенным лагом	31
ИНТЕГРИРОВАННЫЕ ПРОЦЕССЫ, ЛОЖНАЯ РЕГРЕССИЯ И КОИНТЕГРАЦИЯ	34
• Стационарные и нестационарные случайные процессы	34
• Ложная регрессия	36
• Тестирование стационарности	38
• Коинтеграция. Регрессии с интегрированными переменными	43
• Оценивание коинтеграционной регрессии: подход Энгла-Грейнджера	45
• Коинтеграция в динамических системах: подход Йохансена	47
• Литература по единичным корням и коинтеграции	51

II. МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ В

ЭКОНОМЕТРИИ	53
БАЗОВЫЕ ПОНЯТИЯ.....	54
ХАРАКТЕРИСТИКА ММП.	59
СВЯЗЬ ММП С МНК. КВАЗИ-МП МЕТОДЫ.	60
СВЯЗЬ ГЕССИАНА И МАТРИЦЫ ВКЛАДОВ В ГРАДИЕНТ С ИНФОРМАЦИОННОЙ МАТРИЦЕЙ	61
• Гессиан и информационная матрица.....	61
• Матрица вкладов в градиент и информационная матрица	62
• Вычисление информационной матрицы	63
РАСПРЕДЕЛЕНИЕ ГРАДИЕНТА И ОЦЕНОК МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ.....	65
• Асимптотическое распределение градиента и оценок максимального правдоподобия.....	65
• Выборочная оценка распределения градиента и оценок максимального правдоподобия.....	66
ЧИСЛЕННЫЕ МЕТОДЫ НАХОЖДЕНИЯ ОЦЕНОК МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ	68
ММП И ПРОВЕРКА ГИПОТЕЗ.....	70
• Асимптотическое распределение и асимптотическая эквивалентность трех классических статистик	70
• Соотношения между статистиками	75
МОДЕЛИ С ДИСКРЕТНОЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ.....	78
• Модели с бинарной зависимой переменной (логит и пробит)	78
• Пуассонова регрессия	81
ОБОБЩЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ	83
РЕГРЕССИИ С НЕОДИНАКОВОЙ ДИСПЕРСИЕЙ И ТЕСТИРОВАНИЕ ГЕТЕРОСКЕДАСТИЧНОСТИ.....	86
• Взвешенный метод наименьших квадратов	86
• Проверка гипотезы о наличии гетероскедастичности известного вида	87
• Регрессия с мультипликативной гетероскедастичностью	89
НЕЛИНЕЙНАЯ РЕГРЕССИЯ. МЕТОД ГАУССА-НЬЮТОНА	91

ОЦЕНИВАНИЕ РЕГРЕССИИ С AR-ОШИБКОЙ	91
• Нелинейная регрессия с пропущенным первым наблюдением.....	92
• Оценивание регрессии с AR(1)-ошибкой полным методом максимального правдоподобия.....	93
РЕГРЕССИЯ С МА-ОШИБКОЙ	95
• Оценивание регрессии с MA(1)-процессом в ошибке полным методом максимального правдоподобия.....	95
• Оценивание регрессии с MA-ошибкой нелинейным МНК.....	97
РЕГРЕССИЯ С ARCH-ПРОЦЕССОМ В ОШИБКЕ	97
ЯКОБИАН ПРЕОБРАЗОВАНИЯ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ В ФУНКЦИИ ПРАВДОПОДОБИЯ	102
• Функция правдоподобия модели типа $\varepsilon = f(Y, \theta_1)$	102
• Преобразование зависимой переменной. Модель Бокса-Кокса	104
ТЕСТ НА НОРМАЛЬНОСТЬ	106
РЕГРЕССИЯ С ОШИБКАМИ ВО ВСЕХ ПЕРЕМЕННЫХ.....	109
ВНЕШНЕ НЕ СВЯЗАННЫЕ РЕГРЕССИОННЫЕ УРАВНЕНИЯ.....	112
СИСТЕМЫ ОДНОВРЕМЕННЫХ УРАВНЕНИЙ	116
• FIML	117
• LIML	120
ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА.....	122
ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ.....	124

I. Некоторые эконометрические методы

Функциональная форма регрессионной модели

Необходимость изменить функциональную форму модели возникает, если неверна одна из следующих гипотез, выполнение которых требуется для того, чтобы обычный метод наименьших квадратов (ОМНК) в применении к регрессионной модели $Y_i = X_i\beta + \varepsilon_i$ ($i = 1, \dots, N$) давал хорошие результаты:¹

1. Ошибки имеют нулевое математическое ожидание, или, что то же самое, мат. ожидание зависимой переменной является линейной комбинацией регрессоров:

$$E(\varepsilon_i) = 0, \quad E(Y_i) = X_i\beta.$$

2. Ошибки гомоскедастичны, т. е. имеют одинаковую дисперсию для всех наблюдений:

$$V(\varepsilon_i^2) = E(\varepsilon_i^2) = \sigma^2.$$

Тестирование правильности спецификации регрессионной модели

Если ошибка имеет ненулевое мат. ожидание, то оценки ОМНК окажутся смещенными. Другими словами, в ошибке осталась детерминированная (неслучайная) составляющая, которая может быть функцией входящих в модель

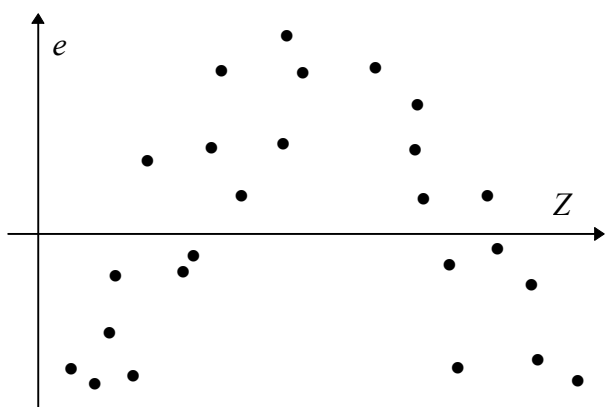


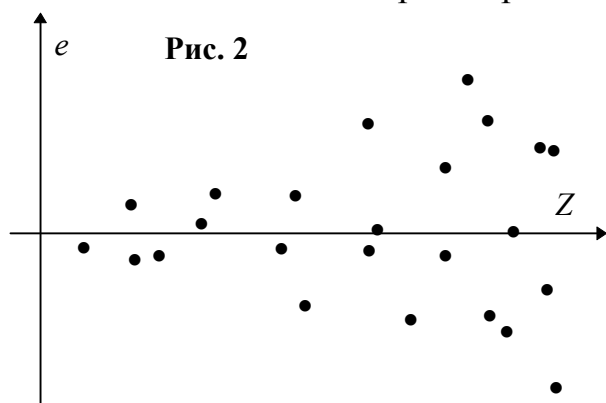
Рис. 1

регрессоров, что и означает, что функциональная форма выбрана неверно. Заметить эту ошибку спецификации можно на глаз с помощью графиков остатков по “подозрительным” переменным: регрессорам и их функциям (в т. ч. произведениям разных регрессоров), расчетным значениям и их функциям. Остатки дают представления об ошибках, поэтому они должны в правильно

заданной регрессии иметь везде нулевое среднее. Если остатки (e), например, для каких-то значений некоторой переменной Z в среднем больше нуля, а для каких-то — меньше, то это служит признаком неправильно специфицированной модели (см. Рис. 1).

¹ Это, конечно, не все обычно принимаемые гипотезы, а только те, которые интересны с точки зрения рассматриваемой темы.

Похожим образом обнаруживается и гетероскедастичность (отсутствие гомоскедастичности). Она проявляется в том, что разброс остатков меняется в зависимости от некоторой переменной Z (см. Рис. 2)



Дисперсия ошибок может меняться в зависимости от регрессоров и их функций, расчетных значений и их функций. Формальный тест можно провести с помощью вспомогательной регрессии — регрессии квадратов остатков по “подозрительным” переменным и константе.

Соответствующая статистика — обычная F-статистика для гипотезы о равенстве нулю коэффициентов при всех переменных кроме константы, выдаваемая любым статистическим пакетом.

Ошибки в спецификации функциональной формы обнаруживаются также тестами на автокорреляцию остатков, такими как статистика Дарбина-Уотсона, если наблюдения упорядочены по какому-либо признаку, например, по порядку возрастания одного из регрессоров. Понятно, что это тест неформальный.

Линейные и нелинейные модели

Линейная форма модели в целом является более предпочтительной. Нелинейные модели оцениваются более простым методом наименьших квадратов. При выполнении некоторого набора гипотез оценки ОМНК для линейной модели обладают рядом хороших свойств, не выполняющихся для оценок нелинейной модели, это же относится к распределениям оценок и различным статистикам.

В линейной регрессионной модели мат. ожидание зависимой переменной — это линейная комбинация регрессоров с неизвестными коэффициентами, которые и являются оцениваемыми параметрами модели. Такая модель является **линейной по виду**. В матричной форме ее можно записать как $Y = X\beta + \varepsilon$. Не обязательно, чтобы влияющие на Y факторы входили в модель линейно. Регрессорами могут быть любые точно заданные (не содержащие неизвестных параметров) функции исходных факторов — это не меняет свойств ОМНК. Важно, чтобы модель была **линейной по параметрам**. Бывает, что модель записана в виде, который нелинеен по параметрам, но преобразовани-

ем уравнения регрессии и переобозначением параметров можно привести ее к линейному виду. Такую модель называют **внутренне линейной**.

Поясним введенные понятия на примерах. Модель $Y = \alpha + \beta X_1 X_2 + \varepsilon$ нелинейна по X_1 и X_2 , но линейна по параметрам, и можно сделать замену $X = X_1 X_2$, так что модель примет линейный вид: $Y = \alpha + \beta X + \varepsilon$. Модель $Y = \exp(\alpha + \beta x + \varepsilon)$ нелинейна по виду, но сводится к линейной логарифмированием обеих частей: $\ln Y = \alpha + \beta x + \varepsilon$. В этой новой модели зависимой переменной будет уже $\ln Y$. Модель $Y = (a - 1)(b + X) + \varepsilon$ нелинейна по параметрам a и b , но сводится к линейной заменой параметров $\alpha = (a - 1)b$ и $\beta = a - 1$. Тогда $Y = \alpha + \beta X + \varepsilon$.

Для применения метода наименьших квадратов важно, чтобы ошибка была **аддитивной**, то есть, чтобы зависимая переменная являлась суммой своего математического ожидания и ошибки. Об этом следует помнить, производя преобразования модели. Например, модель $Y = \alpha X^\beta + \varepsilon$ нельзя преобразовать в линейную по параметрам с аддитивной ошибкой. Аналогичную **модель с мультипликативной ошибкой** $Y = \alpha X^\beta \varepsilon$ можно преобразовать к виду $\ln Y = \ln \alpha + \beta \ln X + \ln \varepsilon$ или $\tilde{Y} = \tilde{\alpha} + \beta \tilde{X} + \tilde{\varepsilon}$ где $\tilde{Y} = \ln Y$, $\tilde{\alpha} = \ln \alpha$, $\tilde{X} = \ln X$, $\tilde{\varepsilon} = \ln \varepsilon$. Однако следует отметить, что вследствие преобразования распределение ошибки изменилось. Если $\tilde{\varepsilon}$ оказывается нормально распределенной, это значит, что ε имела логнормальное распределение.

Экономическая теория оперирует моделями разных типов. Некоторые из них дают регрессионные уравнения линейного вида, некоторые – нелинейного. Рассмотрим это на примере однородных производственных функций. Самая популярная производственная функция – **функция Кобба-Дугласа** – легко приводится к линейному виду логарифмированием:

$$Y = \alpha K^\beta L^{1-\beta} \Rightarrow \\ \ln Y - \ln L = \ln \alpha + \beta (\ln K - \ln L),$$

где Y – выпуск продукции, K — капитал, L — труд.

Функция с постоянной эластичностью замены (ПЭЗ) дает внутренне нелинейное уравнение регрессии:

$$Y = \alpha (\beta K^\rho + (1-\beta)L^\rho)^{1/\rho}.$$

Достаточно гладкую функцию вблизи некоторой точки можно разложить в ряд Тейлора, получив тем самым линейную форму модели. Так, при $\rho \rightarrow 0$ функция с постоянной эластичностью замены совпадает с функцией Кобба-Дугласа. Если же приблизить функцию ПЭЗ в точке $\rho = 0$ разложением в ряд

Тейлора до членов первого порядка, то получается так называемая **транслоговая производственная функция**:

$$\ln Y - \ln L = \ln \alpha + \beta (\ln K - \ln L) + \gamma (\ln K - \ln L)^2,$$

$$\text{где } \gamma = \frac{1}{2} \rho \beta (1 - \beta).$$

Разложение в ряд Тейлора дает **полиномиальную форму модели**. В полиномиальную регрессионную модель могут входить не только первые степени исходных переменных, но и их одночлены различных степеней: степени этих переменных и **члены взаимодействия** (произведения степеней двух или более различных переменных).

Может случиться: что “истинная” модель бывает настолько нелинейной, что полиномиальное приближение становится неудовлетворительным — количество оцениваемых параметров было бы слишком большим. Тогда приходится пожертвовать удобствами ОМНК и использовать нелинейный МНК или другие методы. Есть также много других причин, по которым предпочтительнее использовать внутренне нелинейную функциональную форму. Например, функция ПЭЗ, рассмотренная выше, включает в себя как частные случаи при разных значениях параметра ρ сразу несколько популярных видов производственных функций: функцию Кобба-Дугласа, линейную функцию (с полной взаимозаменяемостью факторов) и функцию леонтьевского типа (с полной взаимодополняемостью факторов). Оценив ее, можно сделать вывод о том, к какому из этих трех видов ближе “истинная” функция.

Кроме натуральных степеней исходных переменных можно использовать и другие функции от них. Это и уже встречавшиеся выше логарифмы и т. п.: $\ln X$, \sqrt{X} , $1/X$, e^X , $1/(1+e^{-X})$ (логиста) и др. Интересной функцией является **преобразование Бокса-Кокса**: $\frac{X^\alpha - 1}{\alpha}$. При $\alpha \rightarrow 0$ она стремится к $\ln X$. При других значениях это некоторая степень X (с точностью до линейного преобразования). В этом отношении преобразование Бокса-Кокса схоже с функцией ПЭЗ. Оно также похоже на нее в том отношении, что дает внутренне нелинейную модель. Обычно исследователь обладает достаточной свободой при выборе функциональной формы модели. Но важно, чтобы при этом не нарушались те условия, которые необходимы для хорошей работы применяемых методов оценивания. Нужно не забывать проводить проверку правильности спецификации модели и исправлять модель, когда получена плохая диагно-

стика, например, добавлять одночлены более высоких степеней в полиномиальную модель.

Рассмотрим, как может помочь изменение функциональной формы в борьбе с гетероскедастичностью. Многие экономические переменные таковы, что размер отклонений, с ними связанных, зависит от величины этих переменных (например, пропорционален), а величина эта в выборке колеблется в широких пределах (изменяется в несколько раз). Возникающая при этом гетероскедастичность снижает эффективность оценок параметров. Объяснить потерю эффективности можно следующим образом. В методе наименьших квадратов все наблюдения выступают в одинаковых "весовых категориях", и поэтому в оценках непропорционально мало используется информация от наблюдений с меньшей дисперсией. Тем самым происходит потеря информации. Поэтому, например, нехорошо в регрессию включать временные ряды для номинальных показателей, если в рассматриваемой стране высокая инфляция, или использовать непреобразованную модель в случае выборки стран, в которой есть и большие, и малые страны (США наряду с Исландией). Обычно применяют два вида преобразований. Рассмотрим их на примере функции потребительского спроса кейнсианского типа: $C = \alpha I + \beta X + \varepsilon$, где C — потребление, I — доход, X — символизирует прочие факторы. Разумно предположить, что среднеквадратическое отклонение ошибки прямо пропорционально I .

1) **Нормирование.** Пронормировать рассматриваемую модель можно, разделив ее на I :

$$C/I = \alpha + \beta X/I + \varepsilon/I.$$

Можно использовать для нормировки (взвешивания) и переменную, не входящую в модель. Обозначим ее N :

$$C/N = \alpha + \beta X/N + \varepsilon/N.$$

Нормирование равнозначно использованию взвешенного метода наименьших квадратов. Как веса для номинальных величин можно использовать уровень цен, получив тем самым реальные величины. Как веса для стран можно использовать население, получив тем самым среднедушевые показатели (потребление на душу населения и т. п.).

2) **Логарифмирование.** Прологарифмировав уравнение $C = \alpha I + \varepsilon$ при $\sqrt{V(\varepsilon)} \ll C$ можно получить следующее линейное приближение:

$$\ln C = \ln \alpha + \ln I + \varepsilon/\alpha I.$$

Вряд ли можно привести теоретические возражения и против того, чтобы сразу использовать **линейную в логарифмах модель** (эта форма модели сокращенно называется **логлинейной**), например,

$$\ln C = \alpha + \beta \ln I + \varepsilon.$$

"Кандидатами" на логарифмирование в первую очередь служат те переменные, которые заведомо могут принимать только положительные значения. Один из их признаков, это то, что, как правило, интересуются относительными приростами таких переменных, а не абсолютными приростами. В экономике это следующие величины: физические объемы благ, цены, стоимостные показатели, различные индексы.

Как итог, перечислим основные функциональные формы регрессионной модели (без учета ошибки) с примерами.

Функциональная форма	Пример
Линейная	$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$
Полиномиальная	$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3$
	$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_{11} X_1^2 + \alpha_{22} X_2^2 + \alpha_{12} X_1 X_2$
Логлинейная (линейная в логарифмах)	$\ln Y = \alpha_0 + \alpha_1 \ln X$
Мультипликативная	$Y = \alpha_0 X_1^{\alpha_1} X_2^{\alpha_2}$
Нормированная	$Y/N = \alpha_0 + \alpha_1 X/N$

Возможны различные комбинации этих форм. Например, часто встречается **полулогарифмическая форма**:

$$\ln Y = \alpha + \beta X, \text{ или } Y = \alpha + \beta \ln X,$$

$$\text{или } \ln Y = \alpha + \beta \ln X + \gamma Z.$$

Выбор между альтернативными функциональными формами

Самый распространенный способ выбора между альтернативными моделями — выбор на основе точности подбора. В качестве показателя точности подбора обычно используется коэффициент детерминации (R^2). Не следует забывать, что этот показатель можно использовать для сравнения только моделей с одной и той же зависимой переменной. Чтобы учитывать при выборе простоту модели, делают поправку на количество регрессоров. Это дает ко-

ээффициент детерминации скорректированный на количество степеней свободы (\tilde{R}^2).

Оценки метода наименьших квадратов являются одновременно и оценками метода максимального правдоподобия. Поэтому предлагается сравнивать модели на основе максимума логарифмической функции правдоподобия ($\hat{\ell}$). Если учесть при этом количество наблюдений (N) и ввести “штраф” за большое количество регрессоров (k), то получится **информационный критерий Акаике** (Akaike information criterion):

$$AIC = -2/N(\hat{\ell} - k).$$

Чем меньше AIC, тем лучшей считается модель.

Существует и другой подход к выбору между моделями. Одна из моделей предполагается истинной, т.е. принимается за нулевую гипотезу, и тестируется против некоторой альтернативной гипотезы, спецификация которой зависит от альтернативной модели. По сути дела, осуществляется тестирование функциональной формы “нулевой” модели.

Если одна из моделей является частным случаем другой модели (англ. nested), то в качестве “нулевой” берется более узкая модель, а альтернативой служит более широкая. В случае линейной регрессии применяется соответствующая F-статистика, а в случае нелинейной — одна из χ^2 -статистик: статистика Вальда, множителя Лагранжа или отношения правдоподобия. Если же модели не входят одна в другую (nonnested), то любая из них принимается за нулевую и дополняется за счет информации, содержащейся в другой модели, так, чтобы “нулевая” модель была частным случаем этой расширенной. Здесь уже можно применить один из вышеупомянутых тестов. Если нулевая гипотеза отвергается, то это означает, что альтернативная модель содержит какую-то информацию, не содержащуюся в “нулевой” модели.

Тестов такого рода предложено очень много. Опишем только концептуально наиболее простые.

Сначала рассмотрим случай, когда обе сравниваемые модели линейны и зависимая переменная одна и та же. **J-тест** заключается в том, что в “нулевую” модель добавляется в качестве еще одного регрессора расчетные значения из альтернативной модели. Проверяется гипотеза о равенстве коэффициента при дополнительном регрессоре нулю с помощью соответствующей t-статистики.

Похожий тест состоит в том, что в “нулевую” модель добавляют из альтернативной модели все те регрессоры, которые не содержатся в нулевой и проверяют гипотезу о равенстве коэффициентов при дополнительных регрессорах нулю с помощью соответствующей F-статистики. В этом тесте обе сравниваемые модели содержатся в расширенной модели.

Один из тестов для сравнения моделей с разными зависимыми переменными — **Р_E-тест**. Пусть две сравниваемые модели заданы следующими уравнениями:

$$Y_1 = f_1(Y) = X_1 \beta_1 + \varepsilon_1,$$

$$Y_2 = f_2(Y) = X_2 \beta_2 + \varepsilon_2.$$

Например, $f_1(Y) = Y$, $f_2(Y) = \ln Y$ или $f_2(Y) = Y/W$ (“взвешенная” зависимая переменная). В “нулевую” модель в Р_E-тесте добавляется регрессор, равный разности расчетных значений из альтернативной модели и приведенных к тому же виду расчетных значений из “нулевой” модели. Так, в первую модель нужно добавить

$$X_2 \hat{\beta}_2 - f_2(f_1^{-1}(X_1 \hat{\beta}_1)) = \hat{Y}_2 - f_2(f_1^{-1}(\hat{Y}_1)).$$

Пусть, к примеру, $f_1(Y) = Y$, а $f_2(Y) = \ln(Y)$. Тогда в первую модель добавляют $\hat{Y}_2 - \ln(\hat{Y}_1)$, а во вторую — $\hat{Y}_1 - \exp(\hat{Y}_2)$.

Если отвергаются обе модели, то это должно означать, что каждая из них содержит информацию, не содержащуюся в другой, и следует попытаться как-то соединить две модели в одну. Если обе модели не отвергаются, то это означает, что с точки зрения данного теста они эквивалентны.

Фиктивные переменные как регрессоры

Общие соображения

Термин “фиктивные переменные” используется как противоположность “значащим” переменным, показывающим уровень количественного показателя, принимающего значения из непрерывного интервала. Как правило, фиктивная переменная — это индикаторная переменная, отражающая качественную характеристику. Чаще всего применяются бинарные фиктивные переменные, принимающие два значения, 0 и 1, в зависимости от определенного условия. Например, в результате опроса группы людей 0 может означать, что опрашиваемый — мужчина, а 1 — женщина. К фиктивным переменным иногда относят регрессор, состоящий из одних единиц (т.е. константу, свободный член), а также временной тренд.

Фиктивные переменные, будучи экзогенными, не создают каких-либо трудностей при применении ОМНК. Фиктивные переменные являются эффективным инструментом построения регрессионных моделей и проверки гипотез.

Пример. (Проверка гипотезы о равенстве средних в двух выборках в предположении равенства дисперсий)

Нулевая гипотеза состоит в том, что случайные величины в двух выборках имеют одинаковое математическое ожидание. Альтернативная гипотеза состоит в том, что математические ожидания равны только в пределах выборок, но не между выборками. Предполагается, что величины нормально распределены и дисперсии одинаковы для всех наблюдений. Объединим две выборки в одну. Пусть Y_i — вектор наблюдений для данной величины, D_i — фиктивная переменная принимающая значение 0 для первой выборки и 1 для второй выборки. Тогда для проверки гипотезы оценим регрессионную модель:

$$Y_i = \alpha + \beta D_i + \varepsilon_i.$$

Нулевая гипотеза: $\beta=0$. Альтернативная гипотеза: $\beta \neq 0$. Такую гипотезу можно проверить с помощью t-статистики для коэффициента β . $\hat{\alpha}$ будет оценкой мат. ожидания для первой выборки, $\hat{\alpha} + \hat{\beta}$ для второй.

— — —

Предположим, что математическое ожидание зависимой переменной в регрессии увеличивается на некоторую фиксированную величину, если выполняется определенное условие. Пусть, например для выборки предприятий одной отрасли оценивается производственная функция Кобба-Дугласа. Есть гипотеза, что для частных предприятий в этой отрасли производство при тех же труде и капитале выше, чем для государственных. Введем переменную D_i , которая принимает значение 0 для государственных предприятий и 1 для частных. Регрессионное уравнение будет иметь вид:

$$\ln Y_i - \ln L_i = \alpha_0 + \alpha_1 D_i + \beta (\ln K_i - \ln L_i).$$

Если коэффициент α_1 значимо положителен, то гипотезу нельзя отвергнуть.

Еще одна область применения фиктивных переменных — когда предполагается, что коэффициенты при “значащих” переменных меняются в зависимости от некоторого условия.

Пусть в приведенной модели $\beta = \beta_0$ для гос. предприятий и $\beta = \beta_1$ для частных. Тогда модель запишется в виде:

$$\ln Y_i - \ln L_i = \alpha_0 + \alpha_1 D_i + \beta_0 (\ln K_i - \ln L_i) + (\beta_1 - \beta_0) D_i (\ln K_i - \ln L_i).$$

Заменяя параметры, получаем линейную относительно параметров модель.

В регрессионное уравнение может войти несколько фиктивных переменных. Рассмотрим два условия: А и В. Пусть D_i^A — индикатор условия А ($D_i^A = 1$, если выполнено условие А, и 0 — если нет), D_i^B — индикатор условия В. Тогда $D_i^{AB} = D_i^A D_i^B$ — индикатор одновременного выполнения условий А и В. Эти три переменные будут точно описывать, в каком состоянии находится “мир” для данного наблюдения. Следует отметить, что четвертая фиктивная переменная (индикатор того, что одновременно не выполнены условия А и В) будет излишней в регрессии, уже включающей константу. Если ее добавить в регрессию, то матрица регрессоров будет вырожденной.

Дисперсионный анализ с фиксированными эффектами может быть сведен к регрессионному анализу с фиктивными регрессорами. Проверке гипотез с помощью ковариационного анализа будет соответствовать проверка гипотезы о равенстве нулю коэффициентов при соответствующей группе фиктивных переменных.

Использование фиктивных переменных для проверки однородности наблюдений и прогнозирования

Приведенную выше модель для двух типов предприятий

$$\ln Y_i - \ln L_i = \alpha_0 + \alpha_1 D_i + \beta_0 (\ln K_i - \ln L_i) + (\beta_1 - \beta_0) D_i (\ln K_i - \ln L_i).$$

можно использовать для проверки гипотезы о том, что коэффициенты регрессии разные для гос. предприятий и для частных. Гипотеза проверяется с помощью F-теста на добавление переменных D_i и $D_i (\ln K_i - \ln L_i)$.

В общем случае пусть наблюдения разбиты на две группы — I^1 и I^2 . Матрица регрессоров X распадается на две матрицы регрессоров X^1 и X^2 соответственно, а зависимая переменная Y — на Y^1 и Y^2 соответственно. Нулевая гипотеза состоит в том, что наблюдения порождены моделью $Y = X\beta + \varepsilon$. Альтернативная гипотеза состоит в том, что первая группа наблюдений порождена моделью $Y^1 = X^1\beta^1 + \varepsilon^1$, а вторая группа наблюдений — моделью $Y^2 = X^2\beta^2 + \varepsilon^2$, причем $\beta^1 \neq \beta^2$.

Введем фиктивную переменную D , такую что $D_i=0$ при $i \in I^1$ и $D_i=1$ при $i \in I^2$. Если все ошибки имеют одинаковую дисперсию, то гипотезу можно проверить с помощью регрессии Y по $Z = [X \ X^*D]$. Здесь X^*D обозначает прямое произведение матрицы X на D , так что i -я строка матрицы Z равна $Z_i = [X_i, D_i X_i]$.

Тест на равенство коэффициентов регрессии в двух выборках, называют **тестом Чоу**. Нулевая гипотеза проверяется с помощью F-статистики для гипотезы о том, что коэффициенты при всех добавленных переменных равны нулю.

Еще одно использование фиктивных переменных — проверка гипотезы о том, что некоторое наблюдение принадлежит к той же выборке, что и все остальные наблюдения. Пусть i^* — номер этого наблюдения. Введем фиктивную переменную D , такую что $D_i=0$ при $i \neq i^*$ и $D_{i^*}=1$. Добавим эту переменную в исходную регрессию. Нужной нам статистикой будет F- или t-статистика для гипотезы о том, что коэффициент при добавленной переменной равен нулю. Если нулевая гипотеза отвергается, то соответствующее наблюдение следует считать **выбросом**. Назовем этот тест тестом для выбросов.

Тот же тест можно провести для небольшой группы наблюдений I^* . Требуется добавить регрессию по одной фиктивной переменной описанного вида для каждого из наблюдений $i \in I^*$. Нужной нам статистикой будет F-статистика для гипотезы о том, что коэффициенты при всех добавленных переменных одновременно равны нулю.

Фиктивные переменные, которые равны нулю для всех наблюдений кроме одного, обладают тем свойством, что при добавлении их в регрессию соответствующий остаток зануляется.

Если в тесте Чоу одна из двух выборок содержит мало наблюдений (не больше количества регрессоров), то остатки в этой выборке должны зануляться при применении ОМНК. В этом случае тест Чоу совпадает с описанным только что тестом для выбросов.

Рассмотрим теперь использование фиктивных переменных для прогнозирования. Пусть мы оценили некоторую регрессию ($Y = X\beta + \varepsilon$) и у нас имеются дополнительные наблюдения, для которых известна матрица регрессоров (X^*), но неизвестны значения зависимой переменной (Y^*). Предсказания

находятся по формуле $X^* \hat{\beta}$, где $\hat{\beta}$ — оценки ОМНК из регрессии Y по X . Эти предсказания можно найти с помощью следующей регрессионной модели:

$$\begin{bmatrix} Y \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} X & \mathbf{O} \\ X^* & \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \beta^* \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \varepsilon^* \end{bmatrix}.$$

Вместо неизвестной зависимой переменной здесь стоят нули, и добавлены фиктивные переменные, каждая из которых равна нулю для соответственного добавочного наблюдения. Оценки β будут совпадать с $\hat{\beta}$, а оценки β^* будут равны $\hat{\beta}^* = -X^* \hat{\beta}$, то есть будут равны предсказаниям со знаком минус. Стандартные ошибки предсказаний будут равны стандартным ошибкам оценок $\hat{\beta}^*$, полученным из той же регрессии.

Пусть теперь Y^* становятся известными. Интересно было бы проверить, насколько фактические значения отличаются от предсказанных. Оказывается, можно воспользоваться аналогичной регрессией, в которой слева вместо нулей стоят Y^* :

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} = \begin{bmatrix} X & \mathbf{O} \\ X^* & \mathbf{I} \end{bmatrix} \begin{bmatrix} \beta \\ \beta^* \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \varepsilon^* \end{bmatrix}.$$

Оценки коэффициентов при фиктивных переменных в этом случае будут равны ошибкам предсказаний $\hat{\beta}^* = Y^* - X^* \hat{\beta}$. **Тест на адекватность предсказаний** проводится как тест на одновременное равенство коэффициентов при фиктивных переменных нулю: $\beta^* = 0$. Очевидно, что этот тест совпадает с тестом для выбросов.

Использование фиктивных переменных в моделях с временными рядами

В регрессионных моделях с временными рядами используется три основных вида фиктивных переменных:

1) Переменные-индикаторы принадлежности наблюдения к определенному периоду — для моделирования скачкообразных структурных сдвигов. Границы периода (моменты “скачков”) должны быть установлены из априорных соображений. Например, 1, если наблюдение принадлежит периоду 1941-45 гг. и 0 в противном случае. Это пример использования для моделирования временного структурного сдвига. Постоянный структурный сдвиг моделируется переменной равной 0 до определенного момента времени и 1 для всех наблюдений после этого момента времени.

2) Сезонные переменные — для моделирования сезонности. Сезонные переменные принимают разные значения в зависимости от того, какому месяцу или кварталу года или какому дню недели соответствует наблюдение.

3) Линейный временной тренд — для моделирования постепенных плавных структурных сдвигов. Эта фиктивная переменная показывает, какой промежуток времени прошел от некоторого “нулевого” момента времени до того момента, к которому относится данное наблюдение (координаты данного наблюдения на временной шкале). Если промежутки времени между последовательными наблюдениями одинаковы, то временной тренд можно составить из номеров наблюдений.

Фиктивные переменные помогают отразить тот факт, что коэффициенты линейной регрессии могут меняться во времени. В простейшем случае изменяется константа, а тем самым и мат. ожидание зависимой переменной.

Пусть исходная модель имеет вид $Y_t = \alpha + \beta X_t + \varepsilon_t$ и предполагается, что α линейно зависит от фиктивной переменной F_t : $\alpha_t = \alpha_0 + \alpha_1 F_t$. Тогда уравнение изменится следующим образом: $Y_t = \alpha_0 + \alpha_1 F_t + \beta X_t + \varepsilon_t$, оставаясь линейным по параметрам.

Коэффициенты при значащих переменных тоже могут быть подвержены изменениям. Проинтерпретировать это можно так, что сила их влияния на независимую переменную меняется со временем.

Например, в рассмотренном уравнении может быть $\beta_t = \beta_0 + \beta_1 F_t$. Тогда $Y_t = \alpha + \beta_0 X_t + \beta_1 F_t X_t + \varepsilon$. Эта модель также остается линейной по параметрам. Коэффициент β_1 показывает, как исходный коэффициент β зависит от времени. С помощью соответствующей t-статистики можно проверить гипотезу, что $\beta_1 = 0$ (β не меняется со временем).

Можно предложить следующий тест на стабильность коэффициентов модели во времени. Для его проведения нужно добавить в уравнения произведения всех исходных регрессоров и фиктивной переменной. Например, в модель $Y_t = \alpha + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon$ следует добавить регрессоры F_t , $X_{t1}F_t$ и $X_{t2}F_t$. Если коэффициенты при добавочных переменных значимы в совокупности (применяем F-статистику), то нельзя отвергнуть гипотезу о том, что коэффициенты изменяются со временем.

Тест Чоу представляет собой частный случай описанного теста. Для временных рядов тест Чоу — это тест на то, что в определенный момент времени произошло скачкообразное изменение коэффициентов регрессии.

Временной тренд отличается от бинарных фиктивных переменных тем, что имеет смысл использовать его степени: t^2 , t^3 и т. д. Они помогают моделировать гладкий, но нелинейный тренд. (Бинарную переменную нет смысла возводить в степень, потому что в результате получится та же самая переменная.)

Можно также комбинировать три указанных вида фиктивных переменных, создавая переменные “взаимодействия” соответствующих эффектов. Пусть Y — квартальные данные по некоторому показателю. Его поведение можно смоделировать, представляя мат. ожидание как комбинацию линейного тренда и сезонности.

$$Y_t = \alpha_0 + \alpha_1 t + \beta_1 Q_{t1} + \beta_2 Q_{t2} + \beta_3 Q_{t3} + \gamma_1 Q_{t1}t + \gamma_2 Q_{t2}t + \gamma_3 Q_{t3}t + \varepsilon_t,$$

где t — тренд, Q_i — квартальные сезонные переменные

$$Q_{ij} = \begin{cases} 1 & \text{если } j\text{-й квартал} \\ 0 & \text{иначе} \end{cases}$$

Q_{t4} не нужно вводить в эту регрессию, так как есть константа, а $Q_{t4}t$ не нужно вводить в регрессию, так как есть временной тренд t .

Если все $\gamma_j \neq 0$, то это означает, что структура сезонности линейно изменяется со временем.

Комбинация рассмотренных фиктивных переменных позволяет моделировать еще один эффект — изменение наклона тренда с определенного момента. Помимо тренда в регрессию следует тогда ввести следующую переменную: в начале выборки до некоторого момента времени она равна 0, а вторая ее часть представляет собой временной тренд (1, 2, 3 и т. д. в случае одинаковых интервалов между наблюдениями).

Регрессионные модели с фиктивными переменными являются альтернативой ARIMA-моделям и регрессионным моделям с AR- или MA-процессом в ошибке. В первом случае изменение мат. ожидания во времени можно назвать детерминированным трендом, во втором — стохастическим (строго говоря термин “стохастический тренд” употребляют только по отношению к нестационарным процессам). Решить, какой вид модели применять, сложно. Дело в том, что трудно отличить (в случае малых выборок), когда случайная величина имеет линейный детерминированный тренд со стационарными отклонениями от него, а когда она формируется нестационарным авторегрессионным процессом. То же самое верно для выбора способа моделирования сезонности.

Использование фиктивных переменных имеет следующие преимущества:

- 1) Интервалы между наблюдениями не обязательно должны быть одинаковыми. В выборке могут быть пропущенные наблюдения.
- 2) Коэффициенты при фиктивных переменных легко интерпретировать, они наглядно представляют структуру динамического процесса.
- 3) Для оценивания модели не приходится выходить за рамки классического метода наименьших квадратов.

Спектральный анализ и регрессия

Спектральный анализ можно осуществить с помощью гармонических фиктивных переменных (гармонического тренда). Пусть рассматриваются помесечные данные. Тогда сезонные колебания можно моделировать, используя следующий набор фиктивных переменных:

$$S_{tk} = \sin(2\pi t/k), \quad C_{tk} = \cos(2\pi t/k), \quad k=1, \dots, 6.$$

При $k=1$ период колебаний равен 12 месяцам, при $k=2$ — 6 месяцам, при $k=3$ — 4 месяцам, при $k=5$ — 2,4 месяцам, при $k=6$ — 2 месяцам.

Включение в регрессию полного набора ($k=1, \dots, 6$) рассматриваемых переменных эквивалентно включению набора месячных бинарных фиктивных переменных ($M_{jt} = 1$, если j -й месяц и 0 в противном случае). Гармонические переменные следует применять в том случае, если предполагается, что сезонность может быть гладкой. В этом случае высокочастотные гармоники (с коротким периодом) не включают в регрессию, например, берут только $k=1, 2$.

Одна из возможных содержательных интерпретаций такого подхода состоит в том, что гармоники с более длинным периодом моделируют долгосрочные (перманентные) эффекты, а с коротким — краткосрочные.

Модели с качественной зависимой переменной

Модели с качественной зависимой переменной как правило возникают, когда экономика рассматривается на очень дезагрегированном уровне. Обычно это ситуация, когда некоторая экономическая единица (субъект), делает выбор между двумя и более возможными альтернативами. В качестве примера можно привести выбор предприятия: внедрять какую-то новую технологию или нет.

Модели с бинарной зависимой переменной

В бинарную модель входит зависимая переменная Y , принимающая два значения (обычно 0 и 1), а также регрессоры X , которые содержат факторы, определяющие выбор. Обычная линейная регрессионная модель не подходит

для описания этой ситуации. Она предполагает, что зависимая переменная имеет непрерывное распределение, а здесь необходимо, чтобы она имела дискретное распределение.

Пусть, например, рассматривается выбор группы людей: быть безработным или работать. Построенная линейная регрессия будет предсказывать абсурдные значения Y – дробные, отрицательные и большие единицы. Математическое ожидание остатков при этом будет даже асимптотически зависеть от X .

Вообще говоря, предсказывать результат выбора можно было бы и по результатам линейной регрессии: если \hat{Y} (расчетное значение Y) больше $1/2$, то берем 1, если \hat{Y} меньше $1/2$, то берем 0. Хотя это соображение и не улучшает модель, оно подсказывает, какой может быть более адекватная модель.

С формальной точки зрения требуется найти модель, которая порождала бы дискретное распределение, зависящее от X , которое бы хорошо описывало данные. Поскольку для бинарной зависимости переменной распределение будет бинарным, то оно полностью определяется вероятностью получения единицы (как функцией X), которая совпадает с математическим ожиданием, если переменная принимает значения 0 и 1:

$$E(Y | X) = \text{Prob}(Y = 1 | X) \cdot 1 + \text{Prob}(Y = 0 | X) \cdot 0 = \\ = \text{Prob}(Y = 1 | X).$$

Геометрически задача состоит в том, чтобы найти гиперплоскость, которая бы в определенном смысле наилучшим образом разделяла две группы наблюдений (соответствующие 0 и 1) в пространстве регрессоров. Может случиться, что такая плоскость будет не единственной; это происходит при идеальном разделении.

Модель выбора. Пробит и логит

Предлагается два вида моделей выбора, которые могли бы породить интересующее нас распределение зависимой переменной: *пороговая модель* и *модель, основанная на полезности альтернатив*. Идея пороговой модели уже обрисована выше. Предполагается, что в основе выбора лежит ненаблюдаемая переменная \tilde{Y} , математическое ожидание которой является линейной комбинацией набора регрессоров X : $\tilde{Y} = X\beta + \varepsilon$. Наблюдается только дискретная величина Y , которая связана с \tilde{Y} следующим образом: если \tilde{Y} больше некоторой пороговой величины C , то $Y = 1$, если меньше, то $Y = 0$. Как обыч-

но предполагается, что ошибки ε_i имеют нулевое математическое ожидание, одинаково распределены и независимы. Величину C можно принять равной нулю.

Другая модель предполагает, что выбор осуществляется на основе ненаблюдаемой полезности альтернатив $u(Y, X)$. Если $u(1, X) > u(0, X)$, то выбираем 1, если $u(0, X) < u(1, X)$, то выбираем 0. В простейшем случае полезность является линейной функцией регрессоров:

$$u(1, X) = u_1 = X\beta_1, \quad u(0, X) = u_0 = X\beta_0.$$

Чтобы модель была вероятностной, предполагается, что есть отклоняющие факторы, так что $u_1 = X\beta_1 + \varepsilon_1$, $u_0 = X\beta_0 + \varepsilon_0$. Эта модель сводится к пороговой, если взять $\tilde{Y} = u_1 - u_0 = X(\beta_1 - \beta_0) + \varepsilon_1 - \varepsilon_0 = X\beta + \varepsilon$, а в качестве порога — ноль. Выведем теперь из распределения ε распределение \tilde{Y} , а из распределения \tilde{Y} — распределение Y .²

Есть два удобных вида распределения, которые обычно используют для описания отклонения ε .

1. Логистическое распределение.³

Плотность логистического распределения равна $\frac{e^z}{(1 + e^z)^2}$ (см. Рис. 3), а функция распределения равна $\frac{1}{1 + e^{-z}}$ (ее называют логистой). Модель с бинарной зависимой переменной с логистически распределенным отклонением называют **ЛОГИТ**. Для логита $E(Y | X) = 1 - \frac{1}{1 + e^{X\beta}} = \frac{1}{1 + e^{-X\beta}}$.

² Для симметричного распределения $1 - F(-X) = F(X)$.

³ Если ε_1 и ε_0 имеют распределение Вейбулла и независимы, то ε имеет логистическое распределение. Если ε_1 и ε_0 имеют нормальное распределение и независимы, то ε тоже имеет нормальное распределение.

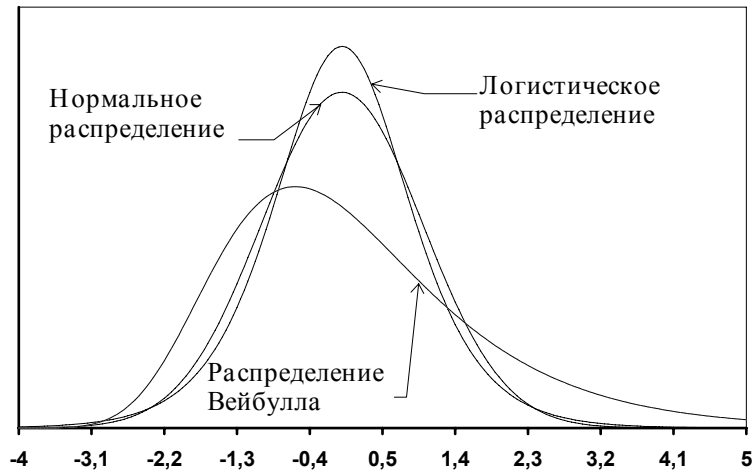


Рис. 3

2. Нормальное распределение.

Модель с нормально распределенным отклонением называют **пробит**.

Для пробита

$$E(Y | X) = \int_{-\infty}^{X\beta} \frac{1}{2\pi} e^{-\frac{t^2}{2}} dt.$$

Логистическое распределение очень похоже на нормальное. Различить, когда следует применять логит, а когда — пробит, в малых выборках невозможно. Оценки коэффициентов β отличаются множителем, который практически постоянен.

Оценка качества модели и проверка гипотез

Пробит и логит обычно оценивают методом максимального правдоподобия. Существуют также упрощенные методы, использующие сгруппированные наблюдения. Предположим, что методом максимального правдоподобия получен вектор оценок $\hat{\beta}$. Как в этом случае можно судить о качестве модели и проверять гипотезы?

Величину $\hat{Y} = X\hat{\beta}$ можно назвать по аналогии с линейной регрессией расчетными значениями. Она является оценкой математического ожидания ненаблюдаемой величины \tilde{Y} , сравнивая которую с нулем делают выбор между 0 и 1. Уравнение $\hat{Y}(X) = 0$ задает ту гиперплоскость, которой разделяются две группы точек — те точки, для которых предсказано $Y_i = 0$, и те точки, для которых предсказано $Y_i = 1$ (с помощью критерия $\hat{Y}_i < 0 \Rightarrow 0$, а $\hat{Y}_i > 0 \Rightarrow 1$). Поэтому наглядно о качестве модели можно судить по диаграмме соответ-

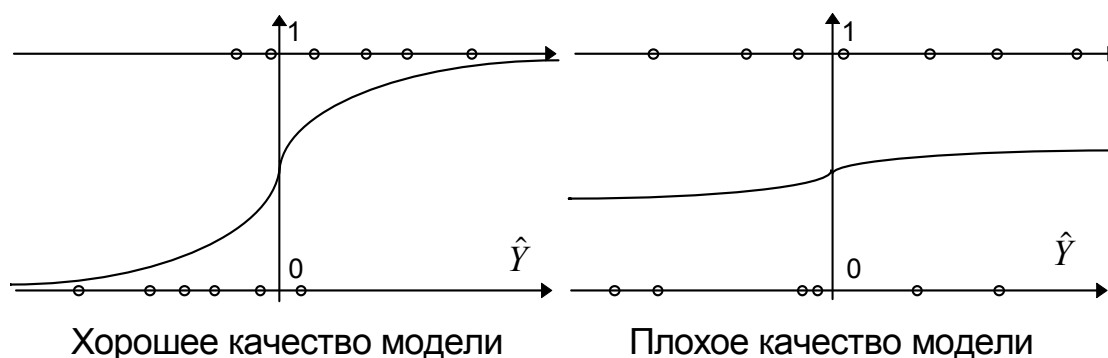


Рис. 4

вующих точек по \hat{Y} : чем лучше разделены две группы точек, тем более качественна модель. О качестве модели можно судить также по графику оценки $E(Y)$ по \hat{Y} ($\frac{1}{1+e^{-\hat{Y}}}$ по \hat{Y}). Этот график в случае “хорошей” модели должен быть

“крутой” в нуле. (См. Рис. 4)

На этих двух графиках слева внизу и справа вверху расположены правильно предсказанные точки, а слева вверху и справа внизу — неправильно. То же самое можно представить таблицей:

		Предсказано		
		0	1	Сумма
На самом деле	0	×	×	×
	1	×	×	×
	Сумма	×	×	×

Понятно, что “хорошая” модель должна давать высокий процент правильных предсказаний.

Для проверки набора ограничений на параметры удобно использовать статистику отношения правдоподобия $LR = 2 (\ell^{\square}(\hat{\beta}) - \ell^{\square}(\hat{\beta}_R))$, где

$$\ell = \sum_{i=1}^N [Y_i \ln P_i + (1 - Y_i) \ln (1 - P_i)]$$

— логарифмическая функция правдоподобия,

$\hat{\beta}$ — оценка методом максимума правдоподобия без ограничений,

$\hat{\beta}_R$ — оценка при ограничениях.

Эту же статистику можно использовать для построения показателя качества модели, аналогичного F-статистике для линейной регрессии. Это статисти-

стика для проверки гипотезы о том, что коэффициенты при всех регрессорах, кроме константы, равны одновременно нулю. Соответствующая статистика отношения правдоподобия равна $LR_0 = 2 (\ell(\hat{\beta}) - \ell_0)$, где ℓ_0 – максимум логарифмической функции правдоподобия для константы. Она распределена асимптотически как χ^2 с $k-1$ степенями свободы, где k – количество параметров в исходной модели, включая константу. Величина l получается следующим образом. Пусть N – общее количество наблюдений, n_0 – количество наблюдений, для которых $Y_i = 0$, n_1 – количество наблюдений, для которых $Y_i = 1$. Тогда предсказанная вероятность появления $Y_i = 1$ в модели с одной константой будет равна для всех наблюдений n_1/N . Отсюда $\ell_0 = n_0 \ln n_0 + n_1 \ln n_1 - N \ln N$. Еще один показатель качества модели, основанный на максимуме функции правдоподобия — информационный критерий Акаике:

$$AIC = -\frac{2}{N} (\ell(\hat{\beta}) - k).$$

Для моделей с бинарной зависимой переменной можно сконструировать и некий аналог коэффициента детерминации — псевдо- R^2 :

$$R^2 = \frac{\sum_{i=1}^N (\hat{Y}_i - \ddot{Y})^2}{\sum_{i=1}^N (\hat{Y}_i - \ddot{Y})^2 + N\sigma^2},$$

где \ddot{Y} — среднее \hat{Y}_i , σ^2 — дисперсия ошибки ε , которая равна 1 для пробита и $\frac{\pi^2}{3}$ для логита.

Множественные модели с качественными зависимыми переменными

В этом подразделе будет говориться о логите, хотя это верно и для пробита. Множественный логит является логическим продолжением бинарного. Он возникает, когда рассматривается выбор между более чем двумя альтернативами. Существует два основных типа множественных моделей: упорядоченный логит и собственно множественный логит. Упорядоченный логит развивает пороговую модель, а собственно множественный логит — модель выбора по полезности.

Упорядоченный логит имеет дело с альтернативами, которые можно расположить в определенном порядке. Например, это могут быть оценки, полученные на экзамене, или качество товара, которое может характеризоваться

сортом от "высшего" до "третьего". Будем предполагать, что альтернативы пронумерованы от 0 до S . Переменная Y принимает значение s , если выбрана альтернатива s . Предполагается, что в основе выбора лежит ненаблюдаемая величина $\tilde{Y} = X\beta + \varepsilon$. $Y=0$ выбирается, если \tilde{Y} меньше нижнего (первого) порогового значения, $Y=1$, если \tilde{Y} попадает в промежуток от первого до второго порогового значения и т. д.; $Y=S$ выбирается, если \tilde{Y} превышает верхнее пороговое значение:

$$Y_i = \begin{cases} 0 & \tilde{Y} \leq \gamma_1 \\ & \gamma_1 < \tilde{Y} \leq \gamma_2 \\ & \vdots \\ S & \tilde{Y} > \gamma_S \end{cases}$$

Если ε имеет логистическое распределение, то логарифмическая функция распределения равна

$$\begin{aligned} \ell &= \sum_{i \in I_0} \ln(\text{Prob}(Y_i = 0)) + \sum_{i \in I_1} \ln(\text{Prob}(Y_i = 1)) + \dots + \\ &+ \sum_{i \in I_S} \ln(\text{Prob}(Y_i = S)) = \\ &= \sum_{i \in I_0} \ln\left(\frac{1}{1 + e^{X_i\beta}}\right) + \sum_{i \in I_1} \ln\left(\frac{1}{1 + e^{X_i\beta - \gamma_2}} - \frac{1}{1 + e^{X_i\beta}}\right) + \dots + \\ &+ \sum_{i \in I_S} \ln\left(\frac{1}{1 + e^{\gamma_S - X_i\beta}}\right). \end{aligned}$$

Эту величину следует максимизировать по β и γ . В результате получается оценка максимума правдоподобия.

Если альтернативы не упорядочены, то предполагается, что выбор делается на основе функции полезности $u(Y, Z)$. Обозначим $u_s(Z) = u(s, Z)$. В линейной модели $u_s = Z_s \beta_s + \varepsilon_s$, где Z_s – матрица регрессоров, β_s – неизвестные параметры. Обычно делают одно из двух упрощающих допущений: либо что регрессоры для всех альтернатив одни и те же: $u_s = Z \beta_s + \varepsilon_s$, либо что функция имеет один и тот же вид, а меняются только факторы, определяющие выбор, т.е. $u_s = Z_s \beta + \varepsilon_s$. Y_i выбирается равным s , если $u_s(Z_i) > u_t(Z_i) \forall s \neq t$. В

множественном логите принимается, что ошибки ε_s имеют распределение Вейбулла. Распределение Вейбулла⁴ в стандартной форме имеет функцию распределения $F(X) = e^{-e^{-X}}$ (см. Рис. 3). Распределение Вейбулла обладает следующими важными для рассматриваемой модели свойствами: максимум нескольких величин, распределенных по Вейбуллу, также распределен по Вейбуллу, а разность двух величин, распределенных по Вейбуллу, имеет логистическое распределение. Используя эти свойства, можно вывести, что в многомерном логите

$$\text{Prob}(Y_i = s) = P_i^s = \frac{e^{Z_s \beta_s}}{\sum_{t=0}^S e^{Z_t \beta_t}}.$$

Вероятности не изменятся, если числитель и знаменатель нормировать, разделив на $e^{Z_0 \beta_0}$:

$$P_i^s = \frac{e^{Z_s \beta_s - Z_0 \beta_0}}{1 + \sum_{t=1}^S e^{Z_t \beta_t - Z_0 \beta_0}}$$

Если принимается, что $\beta_s = \beta \forall s$, то удобно обозначить $Z_s - Z_0 = X_s$ ($s = 1, \dots, S$), а если $Z_s = X \forall s$, то $\beta_s - \beta_0$ можно заменить на β_s .

Динамическая спецификация регрессионной модели

В этом разделе рассматривается, как можно построить модель, в которой переменными являются временные ряды. Основным понятием, употребляемым при разговоре о регрессионной модели для временных рядов, является понятие **лага**. В буквальном смысле по-английски lag — запаздывание. Под лагом некоторой переменной понимают ее значение в предыдущие периоды времени. Например, для переменной Y_t лагом в τ периодов будет $Y_{t-\tau}$. В векторном виде лаг переменной Y принято записывать как $Y_{-\tau}$. В терминологии имеется некоторая неоднозначность. Часто лагом называют величину $t - \tau$. Кроме того, лагом называют структуру, т.е. форму, в которой входят в модель лаги некоторой переменной.

⁴ Распределение Вейбулла также иногда называют распределением экстремального значения первого рода. Кроме того, именем Вейбулла называют и другие распределения, поэтому может возникнуть путаница.

Другой способ обозначения лага — с помощью **лагового оператора**. Его обозначают буквой L (иногда B). Лаговый оператор — это линейный оператор. С ним можно обращаться как с переменной, но он должен стоять *перед* той переменной, к которой применяется. LX обозначает X_{-1} , $L^r X = X_{-r}$. Если применить многочлен от лага $f(L) = a_n L^n + \dots + a_1 L + a_0$ к переменной X , то получится

$$f(L)X = \left(\sum_{i=0}^n a_i L^i \right) X = \sum_{i=0}^n a_i (L^i X) = \sum_{i=0}^n a_i X_{-i}.$$

Другой постоянно используемый оператор — оператор разности или абсолютного прироста Δ , который определяется как $1 - L$, так что $\Delta X = X - X_{-1}$. Вторая разность — дважды взятый оператор Δ : $\Delta^2 = (1 - L)^2 = 1 - 2L + L^2$ и т. д.

Модель распределенного лага

Часто при моделировании экономических процессов на зависимую переменную влияют не только текущие значения объясняющего фактора, но и его лаги. Типичным примером являются капиталовложения: они всегда дают результат с некоторым лагом.

Модель **распределенного лага** можно записать следующим образом:

$$Y = \alpha + \sum_{\tau=0}^q \beta_{\tau} X_{-\tau} + \varepsilon = \alpha f(L) X + \varepsilon,$$

где q — величина наибольшего лага, $f(z) = \sum \beta_{\tau} z^{\tau}$ — многочлен. Коэффициенты β_{τ} показывают структуру лага и называются весами. Оценивание этой модели может быть затруднено проблемой мультиколлинеарности. Такое случается, если величина X_t мало меняется со временем (если X_t — случайный процесс, то это означает автокорреляцию). При этом невозможно точно оценить структуру лага; хотя возможно точно оценить сумму весов β_{τ} . Последнюю можно вычленивать из модели следующим образом:

$$Y = \alpha + \beta_{\Sigma} X + \sum_{\tau=1}^q \beta_{\tau} (X_{-\tau} - X) + \varepsilon, \quad \text{где } \beta_{\Sigma} = \sum_{\tau=0}^q \beta_{\tau}.$$

Это пример преобразования формы регрессионной модели с временными рядами.

В случае мультиколлинеарности лаговых переменных обычно на лаговую структуру накладывают какое-нибудь ограничение, чтобы уменьшить

количество оцениваемых коэффициентов. Одна из возможных структур лага — это **полиномиальный лаг**, веса которого задаются полиномом от величины лага τ :

$$\beta_\tau = \gamma_0 + \gamma_1\tau + \gamma_2\tau^2 + \dots + \gamma_p\tau^p = \sum_{s=0}^p \gamma_s \tau^s, \quad \tau = 0, \dots, q.$$

где p — степень многочлена. Простейший полиномиальный лаг — линейный. Для него $\beta_\tau = \gamma_0 + \gamma_1\tau$. Его структуру можно представить на следующей диаграмме (Рис. 5).

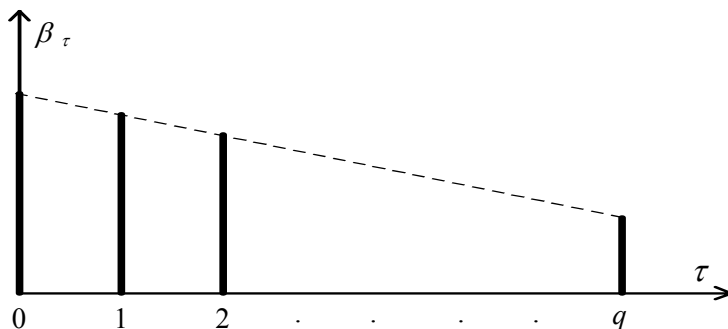


Рис. 5

Полиномиальный лаг накладывают на модель $q - p$ линейных ограничений. Понятно при этом, что если модель была линейной, то она и останется линейной. Рассмотрим, каким образом ее можно оценить.

Подставим выражения для β_τ в исходную модель.

$$\sum_{\tau=0}^q \beta_\tau X_{-\tau} = \sum_{\tau=0}^q \left(\sum_{s=0}^p \gamma_s \tau^s \right) X_{-\tau} = \sum_{s=0}^p \gamma_s \sum_{\tau=0}^q \tau^s X_{-\tau} = \sum_{s=0}^p \gamma_s Z_s.$$

Получим новую модель

$$Y = \alpha + \sum_{s=0}^p \gamma_s Z_s + \varepsilon$$

с преобразованными регрессорами $Z_s = \sum_{\tau=0}^q \tau^s X_{-\tau}$. Оценив γ_s надо подставить

их в формулу для весов β_τ .

При оценивании модели с ограничениями на структуру лага, нужно проверить, правильно ли наложены ограничения. С помощью соответствующей F-статистики можно сравнить ее с исходной, неограниченной, моделью, поскольку она является ее частным случаем. Модель

$$Y = \alpha + \sum_{s=0}^q \gamma_s Z_s + \varepsilon$$

эквивалентна исходной модели с точностью до линейных преобразований, поэтому достаточно проверить гипотезу о том, что последние $q - p$ коэффициентов в ней ($\gamma_{p+1}, \dots, \gamma_q$) равны нулю.

Часто принимают, что веса на концах полиномиальной лаговой структуры равны нулю. Это требование накладывает на модель дополнительные ограничения.

Еще один популярный вид структуры лага — **экспоненциальный (геометрический) лаг**. Его веса задаются следующими соотношениями:

$$\beta_\tau = \beta_0 \delta^\tau, \quad \tau = 0, \dots, \infty, \quad \text{где } 0 < \delta < 1.$$

Веса геометрического лага убывают экспоненциально с увеличением лага (Рис. 6).

Сумма весов в этой модели равна

$$\beta_\Sigma = \sum_{\tau=0}^{\infty} \beta_\tau = \sum_{\tau=0}^{\infty} \beta_0 \delta^\tau = \frac{\beta_0}{1-\delta}.$$

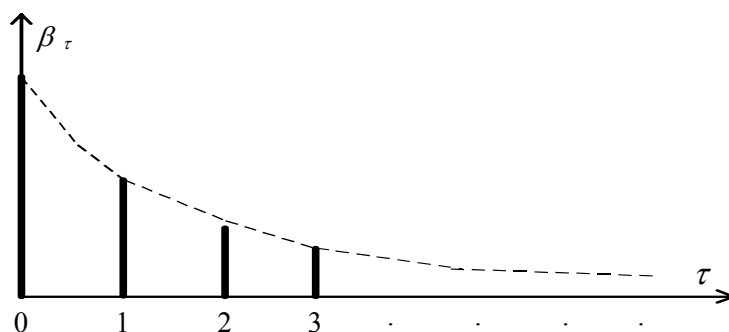


Рис. 6

К модели с геометрическим лагом можно применить **преобразование Койка** (Koyck transformation). Проведем его с использованием лаговых операторов.

$$Y = \sum_{\tau=0}^{\infty} \beta_0 \delta^\tau L^\tau X + \varepsilon = \beta_0 \sum_{\tau=0}^{\infty} (\delta L)^\tau X + \varepsilon = \beta_0 \frac{1}{1-\delta L} X + \varepsilon.$$

Отсюда $(1-\delta L)Y = \beta X + (1-\delta L)\varepsilon$ или, по определению лагового оператора,

$$Y - \delta Y_{-1} = \beta X + \varepsilon - \delta \varepsilon_{-1}.$$

Еще одна проблема, возникающая при оценивании модели распределенного лага, — найти величину наибольшего лага. Самый простой способ — взять неограниченную модель с достаточно большим лагом и проверять гипотезы по “отсечению хвоста” с помощью t и F -статистик.

Динамические регрессионные модели. Авторегрессионная модель с распределенным лагом

Динамическая регрессия — это такая регрессия, в которой в качестве регрессоров используются лаги зависимой переменной. Рассмотрим достаточно общую модель с одной независимой переменной — **авторегрессионную модель с распределенным лагом**. Ее можно записать в следующем виде:

$$Y = \alpha + \sum_{k=1}^p \beta_k Y_{-k} + \sum_{L=0}^q \gamma_L X_{-L} + \varepsilon,$$

где первая сумма представляет собой авторегрессионный член — распределенный лаг зависимой переменной, вторая сумма — распределенный лаг независимой переменной. Сокращенно эту модель обозначают $ADL(p, q)$ (от английского autoregressive distributed lag).

В операторной форме:

$$Y = \alpha + Lf(L)Y + g(L)X + \varepsilon, \text{ где } f(.) \text{ и } g(.) \text{ — многочлены,}$$

или

$$h(L)Y = \alpha + g(L)X + \varepsilon, \text{ где } h(L) = 1 - Lf(L).$$

В частности, $ADL(1, 1)$ имеет вид

$$Y = \alpha + \beta_1 Y_{-1} + \gamma_0 X + \gamma_1 X_{-1} + \varepsilon.$$

Рассмотрим некоторые часто встречающиеся динамические модели, являющиеся частными случаями ADL -модели.

Модель $ADL(0, q)$ — это модель распределенного лага, рассмотренная в предыдущем параграфе, так что в правой части нет лагов зависимой переменной.

Модель геометрического распределенного лага после преобразования Койка — это $ADL(1, 0)$ с $MA(1)$ -ошибкой и ограничением, что коэффициент при Y_{-1} равен параметру MA -процесса (δ) с обратным знаком:

$$Y = (1 - \delta)\alpha + \delta Y_{-1} + \beta_0 X + (\varepsilon - \delta \varepsilon_{-1}).$$

Авторегрессионную модель $AR(p)$ можно считать $ADL(p, 0)$ с ограничением $\beta_0 = 0$. В этой модели переменная в левой части зависит только от своих собственных лагов:

$$Y = \alpha + \sum_{k=1}^p \beta_k Y_{-k} + \varepsilon.$$

В экономике субъекты не сразу могут приспособиться к меняющимся условиям — это происходит постепенно. Нужно время на обучение, переход на новые технологии, изменение условий долгосрочных контрактов и т.д. Эти процессы можно моделировать с помощью **модели частичного приспособления**

$$Y^D = b_0 + b_1 X + \varepsilon,$$

$$\Delta Y = Y - Y_{-1} = \mu(Y^D - Y_{-1}),$$

где Y^D — желаемый уровень величины Y , μ — скорость приспособления ($0 < \mu \leq 1$). Если $\mu = 1$, то приспособление происходит мгновенно и всегда $Y^D = Y$.

Исключив ненаблюдаемую переменную Y^D , модель приводят к виду, удобному для оценивания:

$$Y = \mu b_0 + (1 - \mu)Y_{-1} + \mu b_1 X_{-1} + \mu \varepsilon.$$

Это ADL(1, 1) с коэффициентом при текущем значении X равным нулю.

Чтобы ввести в экономические модели ожидания экономических субъектов в простейшем случае используют модель **адаптивных ожиданий**. Адаптивные ожидания некоторой величины формируются только на основе прошлых значений этой величины. Например, пусть Y зависит от ожиданий величины X (X^E):

$$Y = \alpha_0 + \alpha_1 X^E + \varepsilon.$$

Ошибка в ожиданиях в предыдущий период приводит к корректировке ожиданий:

$$\Delta X^E = X^E - X_{-1}^E = \theta(X - X_{-1}^E).$$

Здесь θ — скорость приспособления ожиданий ($0 < \theta \leq 1$). Если $\theta = 1$, то ожидания всегда равны действительной величине X : $X^E = X$.

Решить разностное уравнение для ожиданий проще всего с использованием лагового оператора. Схему корректировки ожиданий можно записать как

$$(1 - (1 - \theta)L) X^E = \theta X, \text{ откуда}$$

$$X^E = \frac{\theta}{1 - (1 - \theta)L}$$

$$X = \theta \sum_{\tau=0}^{\infty} (1 - \theta)^\tau X_{-\tau}.$$

Исключив ненаблюдаемые ожидания X^E , получим модель с геометрическим распределенным лагом.

Преобразование Койка дает другую форму модели адаптивных ожиданий — ADL(1, 0) с MA(1)-ошибкой и ограничением на коэффициенты:

$$(1 - (1 - \theta)L)Y = \theta\alpha_0 + \alpha_1\theta X + (1 - (1 - \theta)L)\varepsilon.$$

В динамических регрессионных моделях важно различие между долгосрочной и краткосрочной динамикой (англ. Long-run и short-run). Рассмотрим в долгосрочном аспекте модель ADL(1,1):

$$Y = \alpha + \beta_1 Y_{-1} + \gamma_0 X + \gamma_1 X_{-1} + \varepsilon.$$

Пусть установились стационарные уровни X и Y . Обозначим их X^* и Y^* . Тогда

$$Y^* = \alpha + \beta_1 Y^* + \gamma_0 X^* + \gamma_1 X^*.$$

Уравнение

$$Y^* = \frac{\alpha}{1 - \beta_1} + \frac{\gamma_0 + \gamma_1}{1 - \beta_1} X^* = \alpha' + \lambda X^*$$

описывает **долгосрочное стационарное состояние** экономического процесса. Здесь $\lambda = \frac{\gamma_0 + \gamma_1}{1 - \beta_1}$ — коэффициент долгосрочного влияния X на Y . Если Y и X — логарифмы исходных переменных, то λ — долгосрочная эластичность.

Модель ADL(1,1) можно привести к виду, который отражает краткосрочную динамику экономической системы. В этом виде модель называется **моделью исправления ошибок**, сокращенно ЕСМ (англ. error-correction model):

$$\begin{aligned} \Delta Y &= \alpha - (1 - \beta_1) Y_{-1} + \gamma_0 \Delta X + (\gamma_0 + \gamma_1) X_{-1} + \varepsilon \\ &= \alpha + \gamma_0 \Delta X - (1 - \beta_1) (Y_{-1} - \lambda X_{-1}) + \varepsilon. \end{aligned}$$

Предполагается, что если в предыдущий период переменная Y отклонилась от своего долгосрочного значения $\alpha' + \lambda X$, то член $Y_{-1} - \lambda X_{-1}$ корректирует динамику в нужном направлении. Для того, чтобы это происходило, необходимо выполнение условия $\beta_1 < 1$.

Бывает, что из теории явления известно, что $\lambda = 1$, тогда $\beta_1 + \gamma_0 + \gamma_1 = 1$. Часто именно такую модель называют ЕСМ.

Модели частичного приспособления и адаптивных ожиданий являются частными случаями модели исправления ошибок — не только формально математически, но и по экономическому содержанию. Например, модель частичного приспособления в форме ЕСМ выглядит как

$$\Delta Y = \mu b_0 - \mu (Y_{-1} - b_1 X_{-1}).$$

Интегрированные процессы, ложная регрессия и коинтеграция

Стационарные и нестационарные случайные процессы.

Чтобы проиллюстрировать различие между стационарными и нестационарными случайными процессами, рассмотрим авторегрессию первого порядка (AR(1)), т.е. авторегрессию, содержащую один лаг зависимой переменной:

$$Y_t = \mu + \rho Y_{t-1} + \varepsilon_t, \quad t = (-\infty, \dots, 0, 1, \dots, +\infty)$$

(предполагаем, что $\varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2)$ — независимые одинаково распределенные случайные величины с нулевым мат. ожиданием и дисперсией σ_ε^2).

Слабое определение **стационарности** требует, чтобы математическое ожидание Y_t было постоянным (или нулевым), а ковариации не зависели от времени, только от лага:

$$Y_t = \text{const} (= 0), \quad \text{var}(Y_t) = \sigma_Y^2 = \text{const}, \quad \text{cov}(Y_t, Y_{t-\tau}) = c_\tau.$$

Покажем, что если $|\rho| < 1$, то процесс AR(1) будет стационарным. Решая уравнение авторегрессионной модели, получим

$$Y = \frac{\mu}{1-\rho} + \sum_{i=0}^{\infty} \rho^i \varepsilon_{-i}.$$

Итак, математическое ожидание Y равно $\frac{\mu}{1-\rho}$. Второй член — это взвешенная сумма ошибок (геометрический распределенный лаг). Условие $|\rho| < 1$ гарантирует, что дисперсия этой суммы, а следовательно, и дисперсия Y конечна:

$$\sigma_Y^2 = \sum_{i=0}^{\infty} \rho^{2i} \text{var}(\varepsilon) = \sigma_\varepsilon^2 \frac{1}{1-\rho^2}.$$

Найдем также автоковариации процесса:

$$\text{cov}(Y, Y_{-\tau}) = \sum_{i=0}^{\infty} \rho^{-\tau} \rho^{2i} \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \frac{\rho^{-\tau}}{1-\rho^2}.$$

Таким образом, рассматриваемый процесс слабо стационарен. На самом деле, поскольку ошибки ε_t одинаково распределены, то он стационарен и в сильном смысле.

Вывод изменится, если рассмотреть процесс с определенного момента времени, например, с $t = 1$. Предположим, что Y_0 — детерминированная ве-

личина. В этом случае процесс AR(1) не будет стационарным по данному выше определению. Дисперсия Y и автоковариации будут зависеть от t :

$$\text{var}(Y_t) = \sigma_{Y_t}^2 \quad \text{cov}(Y_t, Y_{t-\tau}) = c_{\tau t}.$$

Однако со временем такой процесс (если только $|\rho| < 1$) все больше приближается к стационарному. Его можно назвать асимптотически стационарным.

При $|\rho| > 1$ это будет “**взрывной**” процесс. Влияние прошлых ошибок в нем не угасает, и все более усиливается со временем. Мы не будем рассматривать такие процессы.

Авторегрессионный процесс первого порядка при $\rho = 1$ называют **случайным блужданием**. Если $\mu = 0$, то это случайное блуждание в собственном смысле слова, а при $\mu \neq 0$ это **случайное блуждание с дрейфом**.

Нет смысла рассматривать случайное блуждание, начавшееся бесконечно давно, поскольку за бесконечное время процесс “уходит в бесконечность”, его дисперсия становится бесконечной.

Для процесса, начавшегося в момент $t = 1$ имеем:

$$Y_t = \mu t + \sum_{i=1}^t \varepsilon_i + Y_0, \quad E(Y_t) = \alpha t + Y.$$

Таким образом, константа (“дрейф”) в авторегрессионной записи процесса приводит к появлению *линейного тренда* в Y_t . Дисперсия равна

$$\text{var}(Y_t) = t\delta_\varepsilon^2.$$

Она возрастает бесконечно со временем.

Случайное блуждание является примером авторегрессионного процесса с **единичным корнем**. Он называется так по следующей причине. Запишем AR(1) с помощью лагового оператора:

$$(1 - \rho L)Y_t = \mu + \varepsilon_t.$$

В левой части этого уравнения первый множитель — многочлен первой степени от лага. Корень этого многочлена равен $1/\rho$. При $\rho=1$ корень многочлена равен 1.

В случае авторегрессионного процесса произвольного порядка имеем

$$f(L)Y_t = \mu + \varepsilon_t.$$

Если все корни многочлена $f(\cdot)$ по модулю больше 1, то есть лежат за пределами единичного круга на комплексной плоскости, то процесс стационарен. Если один из корней лежит в пределах единичного круга, то процесс “взрывной”. Если же $k > 0$ корней лежат на единичной окружности, а осталь-

ные — за ее пределами, то процесс нестационарный, но не “взрывной” и о нем говорят, что он имеет k единичных корней.

Первые разности ΔY_t авторегрессионного процесса первого порядка с $\rho=1$ есть просто ошибки ε_t , т.е. первые разности стационарны. Нестационарный процесс, первые разности которого стационарны называют **интегрированным** первого порядка и обозначают $I(1)$. Стационарный процесс обозначают $I(0)$. Если k -е разности случайного процесса стационарны, то его называют интегрированным k -го порядка и обозначают $I(k)$.

Рассмотрим, например, процесс

$$z_t = \sum_{i=1}^t Y_i, \text{ где } Y_t = Y_{t-1} + \varepsilon_t.$$

Он будет $I(2)$, то есть вторые разности ($\Delta^2 z_t$) стационарны.

Ложная регрессия

Очень часто экономические процессы бывают нестационарными. В качестве примера можно привести объем производства, уровень цен. Уровень безработицы как процент трудоспособного населения это, с другой стороны, пример стационарной переменной. В данном случае термин “стационарность” употреблен не в строгом смысле. Скорее подразумевается, что дисперсия процесса ограничена.

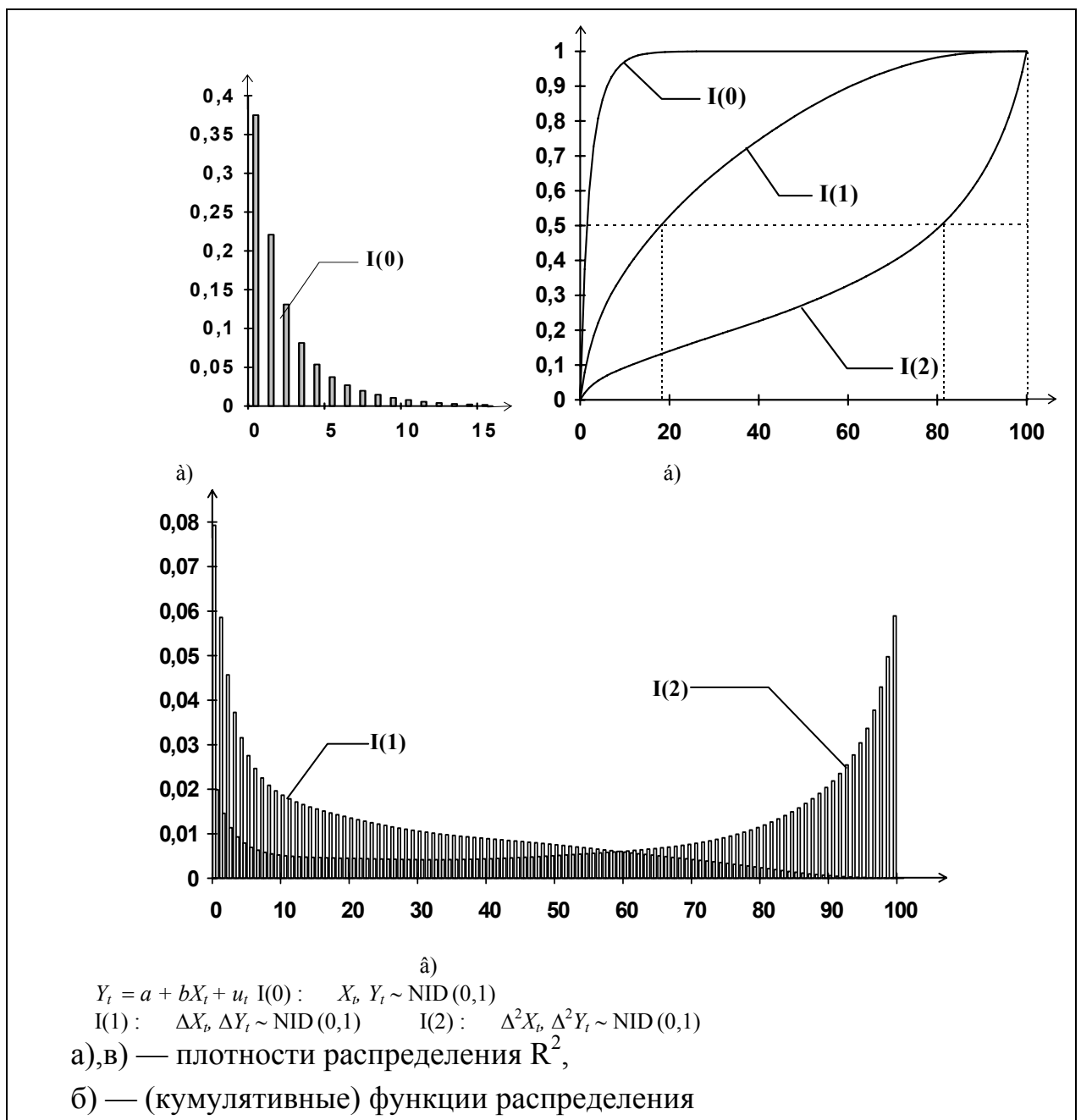
Стационарность регрессоров является очень важным условием при оценивании регрессионных моделей. Если модель неверно специфицирована, и некоторые из переменных, которые в нее неправильно включены, являются $I(1)$, то полученные оценки будут очень плохими. Они не будут обладать свойством состоятельности, то есть не будут сходиться по вероятности к истинным значениям параметров по мере увеличения размеров выборки. Привычные показатели, такие как коэффициент детерминации R^2 , t -статистики, F -статистики, будут указывать на наличие связи там, где на самом деле ее нет. Такой эффект называют **ложной регрессией**.

Показать эффект ложной регрессии можно с помощью метода Монте-Карло. Сгенерируем достаточно много раз два случайных блуждания с независимыми нормально распределенными ошибками ($\varepsilon_t, \xi_t \sim \text{NID}(0,1)$):

$$Y_t = Y_{t-1} + \varepsilon_t, \quad X_t = X_{t-1} + \xi_t.$$

Оценив достаточно много раз регрессию Y_t по константе и X_t вида $Y_t = a + bX_t + u_t$ мы получим экспериментальное распределение различных статистик. Например, эксперименты Монте-Карло показывают, что t-статистика для b при 50 наблюдениях и номинальном уровне значимости 5% в действительности отвергает верную гипотезу об отсутствии связи примерно в 75% случаев. Вместо того, чтобы использовать 5%-ю критическую границу $t_{5\%} \approx 2$ нужно использовать $t_{5\%} = 11,2$.

На рисунке показаны распределения коэффициента детерминации R^2 (в процентах) при длине выборки в 50 наблюдений. Хотя процессы независи-



мы, но регрессия с большой вероятностью дает высокий коэффициент детерминации из-за нестационарности. Два независимых $I(1)$ -процесса примерно в половине случаев дают коэффициент детерминации превышающий 20%. Для $I(2)$ -процессов примерно в половине случаев коэффициент детерминации превышает 80% !

То же самое, хотя и в меньшей степени, можно наблюдать и в случае двух стационарных $AR(1)$ -процессов с коэффициентом автокорреляции ρ близким к 1. Отличие заключается в том, что здесь ложная связь асимптотически (при стремлении размеров выборки к бесконечности) исчезает, а в случае $I(1)$ -процессов — нет. Все же проблема остается серьезной, поскольку на практике экономист имеет дело с конечными и часто довольно малыми выборками.

О процессе типа случайного блуждания без дрейфа говорят как о **стохастическом тренде**, поскольку влияние каждой ошибки не исчезает со временем.

Наличие обычного детерминированного тренда также может приводить к появлению ложной регрессии. Пусть, например Y_t и X_t порождаются процессами $Y_t = a + bt + \varepsilon_t$, $X_t = c + dt + \xi_t$, где ε_t , ξ_t — независимые, одинаково распределенные ошибки. Регрессия Y_t по константе и X_t может иметь высокий коэффициент детерминации и этот эффект только усиливается с ростом размера выборки. К счастью, с “детерминированным” вариантом ложной регрессии достаточно легко бороться. В рассматриваемом случае достаточно добавить в уравнение тренд в качестве регрессора, и эффект ложной регрессии исчезает.

Тестирование стационарности

С осознанием опасности применения ОМНК к нестационарным рядам, появилась необходимость в тестах, которые позволили бы отличить стационарный процесс от нестационарного.

К неформальным методам тестирования стационарности можно отнести визуальный анализ графиков спектральной плотности и автокорреляционной функции.

В настоящее время самым популярным из формальных тестов является тест, разработанный Дики и Фуллером (DF). Базовый порождающий данные процесс (ПДП), который они использовали, — авторегрессионный процесс первого порядка:

$$y_t = \rho y_{t-1} + \varepsilon_t. \quad (A1)$$

При $\rho=1$ это случайное блуждание. Конечно, вряд ли экономическая переменная может быть описана процессом (A1). Более реалистично было бы предположить наличие в этом процессе константы и тренда:

$$y_t = \mu_0 + \rho y_{t-1} + \varepsilon_t. \quad (A2)$$

$$y_t = \mu_0 + \mu_1 t + \rho y_{t-1} + \varepsilon_t. \quad (A3)$$

$$y_t = \mu_0 + \mu_1 t + \mu_2 t^2 + \rho y_{t-1} + \varepsilon_t. \quad (A4)$$

Нулевая гипотеза в тесте Дики-Фуллера состоит в том, что ряд нестационарен и имеет один единичный корень ($\rho=1$) (и при этом $\mu_i=0$), альтернативная — что ряд стационарен ($\rho<1$):

$$H_0: \rho=1, \mu_i=0 \quad H_A: \rho<1.$$

Здесь $i=0$, если оценивается (A2), $i=1$, если оценивается (A3), и $i=2$, если оценивается (A4).

Предполагается, что ошибки ε_t некоррелированы. Это предположение очень важно, без него тест не будет работать!

Для получения статистики, с помощью которой можно было бы проверить нулевую гипотезу, Дики и Фуллер предложили оценить авторегрессию и взять из нее обычную t-статистику для гипотезы о том, что $\rho=1$. При этом тест является *односторонним*, поскольку альтернатива $\rho>1$, соответствующая “взрывному” процессу, не рассматривается.

Необычность DF заключается в том, что с помощью одной t-статистики проверяется гипотеза сразу о двух коэффициентах.⁵ Если мы в регрессии (A3) отвергли нулевую гипотезу, то принимаем альтернативную гипотезу, что процесс описывается уравнением (A3) с $\rho<1$, то есть это стационарный вокруг линейного тренда процесс. В противном случае имеем нестационарный процесс ($\rho=1$), описываемый уравнением (A2), то есть случайное блуждание с дрейфом, но без временного тренда в уравнении авторегрессии.

Часто встречается несколько иная интерпретация этой особенности данного теста: проверяется гипотеза $H_0: \rho=1$ против гипотезы $H_A: \rho<1$, и *оцениваемая регрессия не совпадает с порождающим данные процессом*, каким он предполагается согласно альтернативной гипотезе. Так, чтобы проверить нулевую гипотезу для ПДП типа (A2) нужно построить регрессию (A3) или (A4). Аналогично для тестирования ПДП типа (A3) нужно оценить регрессию (A4). Однако приведенная ранее интерпретация более точная.

⁵ Если применить другую параметризацию, то об одном.

Поскольку полученная статистика имеет нестандартное распределение, для ее использования требуются специальные таблицы. Эти таблицы были получены численно методом Монте-Карло. Все эти статистики получены на основе одного и того же ПДП (A1) с $\rho = 1$, но с асимптотической точки зрения годятся и для других ПДП, несмотря на наличие мешающих параметров, которые приходится оценивать.

Чтобы удобно было использовать стандартные регрессионные пакеты, уравнения регрессии преобразуются так, чтобы зависимой переменной была первая разность. В случае (A1) имеем уравнение ($\phi = \rho - 1$):

$$\Delta y_t = \phi y_{t-1} + \varepsilon_t.$$

Будем обозначать статистику, получаемую в результате оценивания регрессии (A1) τ_{nc} , в результате оценивания регрессии (A2) — τ_c , в результате оценивания регрессии (A3) — τ_{ct} и в результате оценивания регрессии (A4) — τ_{ctt} . Это означает, соответственно, что в регрессии нет константы (nc), есть только константа (c), есть константа и линейный временной тренд (ct), есть константа, линейный тренд и квадратичный тренд (ctt). (Дики и Фуллер использовали другие обозначения, здесь используются обозначения Мак-Киннона).

Следующая таблица показывает, какую статистику можно применять в какой ситуации.

	ПДП с $\rho = 1$, соответствующий нулевой гипотезе		
Регрессия	A1	A2	A2
A1	τ_{nc}		
A2	τ_c	t	t
A3	τ_{ct}	τ_{ct}	t
A4	τ_{ctt}	τ_{ctt}	τ_{ctt}

В таблице t обозначает обычную t-статистику. Дело в том, что когда регрессия совпадает с ПДП и в регрессии есть детерминированные переменные (константа, тренд), то обычная t-статистика асимптотически имеет стандартное нормальное распределение и поэтому для проверки гипотезы годятся обычные критические границы. Правда это свойство существенно асимпто-

тическое, и в малых выборках действительный уровень значимости, как показывают имитации Монте-Карло, может сильно отличаться от номинального. Поэтому предпочтительно добавить в регрессию дополнительную переменную и воспользоваться тестом Дики-Фуллера с нестандартными критическими границами, которые хотя и являются тоже асимптотическими, но связаны с меньшими искажениями размера теста.

Из этой таблицы видно, что если можно предположить, что рассматриваемая переменная нестационарна и имеет тренд, то начать тестирование следует с регрессии (A4) и соответствующего теста τ_{ctt} .

Поскольку неизвестно, присутствуют ли в ПДП константа и тренд, то полезно иметь тесты, которые бы позволили проверить соответствующие гипотезы. Такие тесты были предложены Дики и Фуллером. В случае всех этих тестов (в отличие от DF) действительно проверяемая гипотеза совпадает с номинально проверяемой гипотезой (или, согласно альтернативной интерпретации, оцениваемая регрессия *совпадает* с ПДП, каким он предполагается в соответствии с альтернативной гипотезой). По сути дела используются обычные F- и t-статистики для соответствующих гипотез, только критические границы берут другие. Опять же, при получении этих таблиц методом Монте-Карло используется исключительно ПДП (A1) с $\rho = 1$, поэтому тесты являются асимптотическими.

При оценивании регрессии вида (A2) получаем две статистики: t-статистику для гипотезы $\mu_0 = 0$ и F-статистику для гипотезы $\mu_0 = 0$ и $\rho = 1$. При оценивании регрессии вида (A3) получаем четыре статистики: t-статистику для гипотезы $\mu_0 = 0$, t-статистику для гипотезы $\mu_1 = 0$, F-статистику для гипотезы $\mu_1 = 0$ и $\rho = 1$ и F-статистику для гипотезы $\mu_0 = 0$, $\mu_1 = 0$ и $\rho = 1$.

Было бы естественно предположить, что только что описанные F-статистики было бы предпочтительнее использовать, чем ADF-тесты, поскольку действительная гипотеза для них совпадает с номинальной и является как раз той гипотезой, которая и проверяется в ADF-тестах. Однако эти статистики являются двусторонними и, тем самым, не отбрасывают возможность “взрывного” процесса, что должно приводить к потере мощности теста.

Если гипотеза о наличии единичного корня не была отвергнута, то t-статистики для $\mu = 0$ и $\gamma = 0$ могут быть полезны для определения точного вида нестационарного процесса — имеется ли в нем “дрейф” и тренд.

Предположение о том, что переменная следует авторегрессионному процессу первого порядка и ошибки некоррелированы, является, конечно, слишком ограничительным. Тест Дики-Фуллера был модифицирован для авторегрессионных процессов более высоких порядков и получил название дополненного теста Дики-Фуллера (augmented Dickie-Fuller test, ADF).

Базовые уравнения приобретают следующий вид:

$$\Delta y_t = (\rho - 1) y_{t-1} + \sum_{l=1}^L \Delta y_{t-l} + \varepsilon_t. \quad (B1)$$

$$\Delta y_t = \mu_0 + (\rho - 1) y_{t-1} + \sum_{l=1}^L \Delta y_{t-l} + \varepsilon_t. \quad (B2)$$

$$\Delta y_t = \mu_0 + \mu_1 t + (\rho - 1) y_{t-1} + \sum_{l=1}^L \Delta y_{t-l} + \varepsilon_t. \quad (B3)$$

$$\Delta y_t = \mu_0 + \mu_1 t + \mu_2 t^2 + (\rho - 1) y_{t-1} + \sum_{l=1}^L \Delta y_{t-l} + \varepsilon_t. \quad (B4)$$

Распределения этих тестов асимптотически совпадают с соответствующими обычными тестами Дики-Фуллера, и используют те же таблицы. Грубо говоря, роль дополнительной авторегрессионной компоненты сводится к тому, чтобы убрать автокорреляцию из остатков. Процедура тестирования не отличается от описанной выше.

Как показали эксперименты Монте-Карло, тест Дики-Фуллера чувствителен к наличию процесса типа скользящего среднего в ошибке. Эту проблему частично можно снять, добавляя в регрессию достаточно много лагов первой разности (Said and Dickey, 1984). Чтобы тест был состоятельным, требуется увеличивать L с ростом количества наблюдений по определенному закону.

На практике решающим при использовании ADF является вопрос о том, как выбирать L — порядок AR-процесса в оцениваемой регрессии. Можно предложить следующие подходы.

1) Поскольку важно, чтобы остатки были как можно более похожи на “белый шум”, то следует выбирать такое число L , чтобы тест на автокорреляцию остатков показал отсутствие значимой автокорреляции. Поскольку дополнительные лаги не меняют асимптотические результаты, то лучше взять больше лагов, чем меньше. Однако этот последний аргумент верен только с асимптотической точки зрения.

2) Другой подход состоит в том, чтобы выбирать L на основе обычных t - и F -статистик для соответствующих дополнительных регрессоров.

ADF может давать разные результаты в зависимости от того, каким выбрано количество лагов. Даже добавление лага, который “не нужен” согласно только что приведенным критериям, может резко изменить результат тестирования.

Особую проблему создает наличие сезонной компоненты в переменной. Если сезонность имеет детерминированный характер, то достаточно добавить в регрессию фиктивные сезонные переменные — это не изменяет асимптотического распределения ADF-статистики. Для случая стохастической сезонности также есть специальные модификации теста.

Пока мы рассмотрели тесты $I(1)$ против $I(0)$. Временной ряд может быть интегрированным и более высокого порядка. Как несложно понять, тесты $I(2)$ против $I(1)$ сводятся к рассмотренным, если взять не уровень тестируемого ряда, а первую разность. Аналогично для более высоких порядков интегрирования.

Имитации показали, что следует проверять гипотезы последовательно, начиная с наиболее высокого порядка интегрирования, который можно ожидать априорно. Т. е., сначала следует проверить гипотезу о том, что ряд является $I(2)$, и лишь после этого, если гипотеза была отвергнута, что он является $I(1)$. (См. Dickey and Pantula, 1987.)

Коинтеграция. Регрессии с интегрированными переменными

Как уже говорилось выше, привычные методы регрессионного анализа не подходят, если переменные нестационарны. Однако не всегда при применении МНК имеет место эффект ложной регрессии.

Говорят, что $I(1)$ -процессы Y_{1t} и Y_{2t} являются **коинтегрированными** первого порядка ($CI(1,0)$), если существует их линейная комбинация, которая является $I(0)$, то есть стационарна. То есть $Y_{1t}, Y_{2t} \sim I(1)$, коинтегрированы, если существует коэффициент λ , такой что $Y_{1t} - \lambda Y_{2t} \sim I(0)$. (Granger(1981)).

Понятие коинтеграции тесно связано с моделью исправления ошибки. Коинтегрированные процессы Y_{1t} и Y_{2t} связаны между собой долгосрочным стационарным соотношением, и следует предположить, что существует не-

кий корректирующий механизм, который при отклонениях возвращает Y_{1t} и Y_{2t} к их долгосрочному отношению.

Если $\lambda=1$, то разность Y_{1t} и Y_{2t} будет стационарной и, грубо говоря, Y_{1t} и Y_{2t} будут двигаться “параллельно” во времени. Следующий рисунок (Рис. 7) изображает две таких коинтегрированных переменных, динамика которых задана моделью исправления ошибки:

$$X_{1t} = X_{1t-1} - 0.2(Y_{1t-1} - Y_{2t-1} + 2) + \varepsilon_{1t}$$

$$Y_{2t} = Y_{2t-1} + 0.5(Y_{1t-1} - Y_{2t-1} + 2) + \varepsilon_{2t}$$

$$\varepsilon_{1t}, \varepsilon_{2t} \sim \text{NID}(0,1).$$

Определение коинтеграции естественным образом распространяется на случай нескольких коинтегрированных переменных произвольного порядка интегрирования. Компоненты n -мерного векторного процесса $Y_t = (Y_{1t}, \dots, Y_{nt})$ называют **коинтегрированными** порядка d, b , что обозначается $Y_t \sim \text{CI}(d, b)$, если (1) Y_{it} является $I(d)$ $i=1, \dots, n$ и (2) существует отличный от нуля вектор β , такой что $Y_t \beta \sim I(d-b)$, $d \geq b > 0$. Вектор β называют **коинтегрирующим вектором**.

В рассмотренном ранее примере коинтеграционный вектор имеет вид $\beta = (-1, \lambda)$. Его можно пронормировать также как $(-1/\lambda, 1)$.

Если переменные в регрессии не стационарны, но действительно связаны друг с другом стационарной линейной комбинацией (модель специфицирована верно), то полученные оценки коэффициентов этой линейной комбинации будут на самом деле сверхсостоятельными, то есть сходятся по вероят-

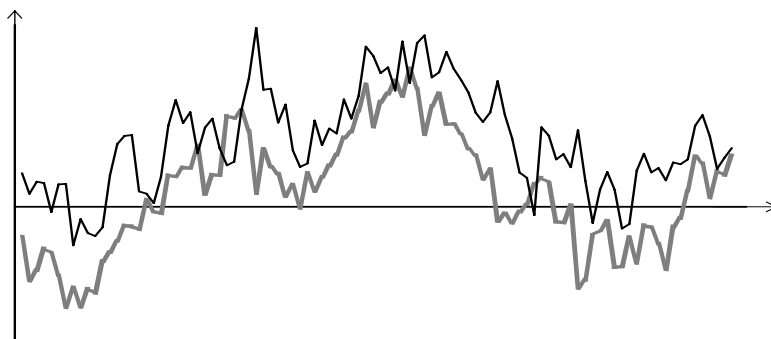


Рис. 7. Два коинтегрированных процесса при $\lambda=1$.

ности к истинным коэффициентам со скоростью, пропорциональной не квад-

ратному корню количества наблюдений, как в регрессии со стационарными переменными, а со скоростью, пропорциональной просто количеству наблюдений. Другими словами в обычной регрессии $\sqrt{N}(\hat{\lambda} - \lambda)$ имеет невырожденное асимптотическое распределение, а в регрессии с I(1)-переменными $N(\hat{\lambda} - \lambda)$ имеет невырожденное асимптотическое распределение.

Обычные асимптотические аргументы сохраняют свою силу, если речь идет об оценках параметров краткосрочной динамики в модели исправления ошибок. Таким образом, можно использовать t-статистики, получаемые обычным методом наименьших квадратов, для проверки гипотез о значимости отдельных переменных. Важно помнить, что это относится к оценкам краткосрочных параметров. Этот подход не годится для проверки гипотез о коэффициентах коинтеграционной комбинации.

Оценивание коинтеграционной регрессии: подход Энгла-Грейнджера

Если бы коэффициент λ был известен, то проверка на коинтегрированность была бы эквивалентна проверке $Y_{1t} - \lambda Y_{2t}$ на стационарность. Но в практических проблемах обычно стационарная линейная комбинация неизвестна. Значит, необходимо оценить коинтегрирующий вектор. Следует также проверить, действительно ли этот вектор дает стационарную линейную комбинацию.

Простейшим методом отыскания стационарной линейной комбинации является **метод Энгла-Грейнджера**. Энгл и Грейнджер предложили использовать оценки, полученные из обычной регрессии с помощью метода наименьших квадратов. Одна из переменных должна стоять в левой части регрессии, другая — в правой:

$$Y_{1t} = \lambda Y_{2t} + u_t.$$

Для тестирования стационарности полученной линейной комбинации предлагается применить метод Дики-Фуллера к остаткам из коинтеграционной регрессии. Пусть — \hat{u}_t остатки из этой регрессии. Тест Энгла-Грейнджера проводится с помощью регрессии

$$\hat{u}_t = \rho \hat{u}_{t-1} + \text{остатки}.$$

Распределение t-статистики для гипотезы $\rho=1$ в этой регрессии будет отличаться (даже асимптотически), от распределения DF-статистики, но имеются соответствующие таблицы. Нулевой гипотезой, таким образом, является

отсутствие коинтеграции. Если мы отвергаем гипотезу об отсутствии коинтеграции, то это дает уверенность в том, что полученные результаты не являются ложной регрессией.

Игнорирование детерминированных компонент ведет к неверным выводам о коинтеграции. Чтобы этого избежать, в коинтеграционную регрессию следует добавить соответствующие переменные — константу, тренд, квадрат тренда, сезонные фиктивные переменные. Добавление константы, тренда, и квадрата тренда, как и в случае DF, меняет асимптотическое распределение теста Энгла-Грейнджера. Следует помнить, что, в отличие от DF, регрессия, из которой берется t-статистика, остается неизменной, то есть в нее не нужно добавлять детерминированные регрессоры.

В МНК регрессии с коинтегрированными переменными оценки должны быть смещенными из-за того, что в правой части стоит эндогенная переменная, коррелированная с ошибкой. Кроме того, ошибка содержит пропущенные переменные. Коинтеграционная регрессия Энгла-Грейнджера является статической по форме, то есть не содержит лагов переменных. С асимптотической точки зрения не приводит к смещенности оценок, поскольку ошибка является величиной меньшего порядка, чем регрессор, дисперсия которого стремится к бесконечности. Как уже говорилось, оценки на самом деле сверхсостоятельны. Однако в малых выборках смещение может быть существенным.

После того, как найдена стационарная линейная комбинация, можно оценить модель исправления ошибок, которая делает переменные коинтегрированными. В этой регрессии нужно использовать первые разности исходных переменных и остатки из коинтеграционной регрессии, которые будут представлять корректирующий член модели исправления ошибок.

Подчеркнем роль корректирующего члена. До появления метода Энгла-Грейнджера исследователи часто оценивали регрессии в первых разностях, что, хотя и приводило к стационарности переменных, но не учитывался стационарный корректирующий член, то есть регрессионная модель была неверно специфицирована (проблема пропущенной переменной).

Несмотря на то, что в модели исправления ошибок используется оценка коинтегрирующего вектора, оценки коэффициентов, полученные из такой модели будут иметь такие же асимптотические свойства, как если бы коинтегрирующий вектор был точно известен. В частности, можно использовать t-статистики из этой регрессии, поскольку оценки стандартных ошибок яв-

ляются состоятельными. Это является следствием сверхсостоятельности оценок коинтегрирующего вектора.

Коинтеграция в динамических системах: подход Йохансена

Другой популярный метод нахождения стационарных комбинаций — **метод Йохансена**. Этот метод служит также для тестирования стационарности найденных линейных комбинаций, и по сути дела распространяет методику Дики-Фуллера на случай **векторной авторегрессии** (то есть такой модели, в которой несколько зависимых переменных и зависят они от собственных лагов и от лагов других переменных). Если в обычной авторегрессии мы рассматривали один коэффициент ρ , то здесь следует рассматривать уже матрицу коэффициентов. Предполагается (как и в ADF), что если добавить достаточное число лагов в авторегрессионную модель, то ошибка не будет сериально коррелированной.

Если векторный процесс состоит более чем из двух процессов ($S > 2$), то может существовать несколько коинтегрирующих векторов. Если существует ровно r линейно независимых коинтегрирующих векторов, то говорят, что **ранг коинтеграции** равен r .

Обозначим β матрицу, составленную из таких векторов. Набор коинтегрирующих векторов не является однозначным, на самом деле речь должна идти о коинтеграционном пространстве. Нормировку следует выбирать исходя из экономической теории рассматриваемых процессов.

Метод Йохансена позволяет не только найти матрицу коинтеграционных векторов при данном ранге коинтеграции, но и проверять гипотезы о ранге коинтеграции (количестве коинтегрирующих векторов). Метод непосредственно работает с **векторной моделью исправления ошибок**. Пусть $Y_t = (Y_{1t}, \dots, Y_{nt})$ — векторный процесс (вектор-строка), каждая из компонент которого является $I(1)$ (или $I(0)$). Порождающий данные процесс задается формулой

$$\Delta Y_t = \mu_0 + \mu_1 t + Y_{t-1} \Pi + \Delta Y_{t-1} \Gamma_1 + \dots + \Delta Y_{t-L+1} \Gamma_{L-1} + \varepsilon_t.$$

Предполагается, что ошибки, относящиеся к разным моментам времени, независимы, и $\varepsilon_t \sim N(0, \Omega)$. В модели оцениваются вектор-строка констант μ_0 и коэффициентов при трендах μ_1 , матрицы коэффициентов $\Gamma_1, \dots, \Gamma_{L-1}$ и Π ($n \times n$), а также ковариационная матрица Ω . Поскольку по предположению $\Delta Y_t \sim I(0)$, то должно быть выполнено $Y_{t-1} \Pi \sim I(0)$. Ограничения на ранг коин-

теграции задаются как ограничения на матрицу Π . При нулевой гипотезе, что ранг коинтеграции равен r , ее можно представить в виде

$$H_0(r): \Pi = \beta\alpha^T,$$

где матрицы α и β имеют размерность $(n \times r)$; β — матрица коинтегрирующих векторов, α — матрица корректирующих коэффициентов. Если $r=0$, то $\Pi=0$ и не существует стационарных линейных комбинаций переменных Y_{1t}, \dots, Y_{nt} . В другом крайнем случае, когда $n=r$ любая линейная комбинация этих переменных стационарна, то есть все они $I(0)$.

Для оценивания модели используется метод максимального правдоподобия. При данной матрице β можно получить оценки максимального правдоподобия для остальных неизвестных параметров обычным методом наименьших квадратов. Йохансен показал также, что максимизация функции правдоподобия по β эквивалентна задаче отыскания собственных чисел для некоторой симметричной положительно определенной матрицы. При ранге коинтеграции r выбираются r минимальных собственных чисел. Если расположить собственные числа в порядке возрастания $(\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n)$, то следует выбрать $\lambda_1, \lambda_2, \dots, \lambda_r$. (Йохансен записал ПДП в несколько ином виде, и поэтому у него собственные числа идут в порядке убывания и выбираются r максимальных собственных чисел.) Столбцами матрицы β (коинтегрирующими векторами) будут соответствующие собственные вектора. Конечно, β определяется только с точностью до некоторой нормировки. После того, как найдена оценка максимального правдоподобия $\hat{\beta}$, вычисляются оценки других параметров.

Для проверки гипотез об r используется статистика отношения правдоподобия. **Статистика следа** используется для проверки гипотезы (H_0) о том, что ранг равен r , против гипотезы (H_A) о том, что ранг равен n . Статистика имеет вид

$$LR^{\text{trace}} = -T \sum_{i=r+1}^n \ln(1 - \lambda_i).$$

Тестирование проводится последовательно для $r = n-1, \dots, 0$ и заканчивается, когда нулевая гипотеза не будет отвергнута в первый раз. Можно проводить тестирование в обратном порядке $r = 0, \dots, n-1$. В этом случае тестирование заканчивается, когда нулевая гипотеза будет отвергнута в первый раз.

Можно также использовать **статистику максимального собственного числа**, которая используется для проверки гипотезы (H_0) о том, что ранг равен r , против гипотезы (H_A) о том, что ранг равен $r+1$. Эта статистика равна

$$LR^{\lambda\text{-max}} = -\ln(1 - \lambda_{r+1}).$$

Обе статистики имеют нестандартные асимптотические распределения. К счастью, их распределения не зависят от мешающих параметров. Распределение этих статистик зависит только от $n-r$ и от того, как входят в модель константа и тренд.

Можно выделить пять основных случаев, касающихся статуса векторов μ_0 и μ_1 в модели. В порядке перехода от частного к более общему:

Случай 0. $\mu_0 = 0, \quad \mu_1 = 0.$

Случай 1*. $\mu_0 = \gamma_0 \alpha^\top, \quad \mu_1 = 0.$

Случай 1. μ_0 произвольный, $\mu_1 = 0.$

Случай 2*. μ_0 произвольный, $\mu_1 = \gamma_1 \alpha^\top.$

Случай 2. μ_0 произвольный, μ_1 произвольный.

Здесь γ_0 и γ_1 — вектора-строки длины r . Случай 0 легко понять — константы и тренды в модели полностью отсутствуют. В Случае 1 константа входит в коинтеграционное пространство и, тем самым, в корректирующие механизмы, но не входит в сам процесс Y_t в виде дрейфа. Это легко увидеть, если переписать модель следующим образом.

$$\Delta Y_t = (\gamma_0 + Y_{t-L} \beta) \alpha^\top + \Delta Y_{t-1} \Gamma_1 + \dots + \Delta Y_{t-L+1} \Gamma_{L-1} + \varepsilon_t.$$

В Случае 1 μ_0 можно записать как $\mu_0 = \gamma_0 \alpha^\top + \mu_0^*$, где γ_0 входит в коинтеграционное пространство, а μ_0^* соответствует дрейфу в векторной модели исправления ошибок. Дрейф в модели исправления ошибок означает, что в Y_t входит линейный тренд. (См. выше рассмотрение простого авторегрессионного процесса с дрейфом.)

Аналогичные рассуждения верны по отношению ко временному тренду в Случаях 2* и 2. В Случае 2* тренд входит в коинтеграционное пространство, но не входит в Y_t в виде квадратичного тренда. В Случае 2 тренд входит и в коинтеграционное пространство, и в Y_t в виде квадратичного тренда.

Методом Монте-Карло получены таблицы LR^{trace} и $LR^{\lambda\text{-max}}$ для всех пяти случаев и нескольких значений $n-r$ (на данный момент имеются таблицы для $n-r=1, \dots, 12$).

Как и в случае ADF очень важным вопросом является выбор длины лага L . Способы по сути дела являются теми же самыми. Для проверки гипотез о длине лага можно использовать тест отношения правдоподобия, который в данном случае имеет обычное распределение χ^2 . Если процесс состоит из n компонент, и проверяется гипотеза о том, что следует увеличить L на единицу то количество степеней свободы соответствующей статистики равно n . Важно также, чтобы отсутствовала автокорреляция остатков.

Метод Йохансена можно использовать также для оценивания моделей с линейными ограничениями на матрицу коинтегрирующих векторов β и на матрицу корректирующих коэффициентов α . Для проверки таких ограничений предлагается использовать все тот же тест отношения правдоподобия, который здесь имеет обычное асимптотическое распределение χ^2 .

Литература по единичным корням и коинтеграции

- Banerjee, A., J.J. Dolado, D.F. Hendry, and G.W. Smith, "Exploring Equilibrium Relationships in Econometrics Through Static Models: Some Monte Carlo Evidence," *Oxford Bulletin of Economics and Statistics*, **48** (1986), 253-277.
- Banerjee, A. J.J. Dolado, J.W. Galbraith and D.F. Hendry, *Co-integration, Error Correction, and the Econometric Analysis of Nonstationary Data*. Oxford: Oxford University Press, 1993.
- Dickey, D.A., W.R. Bell and R.B Miller, "Unit Roots in Time Series Models: Tests and Implications," *American Statistician*, **40** (1986), 12-26.
- Dickey, D.A. and W.A.Fuller, "Distributions of the Estimators for Autoregressive Time Series With a Unit Root," *Journal of American Statistical Association*, **75** (1979), 427-431.
- Dickey, D.A. and S.G. Pantula, "Determining the Order of Differencing in Autoregressive Processes," *Journal of Business and Economic Statistics*, **5** (1987), 455-461.
- Engle, R.F. and C.W.J. Granger, "Co-integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, **55** (1987), 251-276.
- Engle, R.F. and B.S. Yoo, "Forecasting and Testing in Cointegrated Systems," *Journal of Econometrics*, **35** (1987), 143-159.
- Fuller, W.A. *Introduction to Statistical Time Series*. NY: Wiley, 1976.
- Granger C.W.J., "Some Properties of Time Series Data and their Use in Econometric Model Specification," *Journal of Econometrics*, **16** (1981) 121-130.
- Hendry, D.F. "Econometric Modelling with Cointegrated Variables: An Overview," *Oxford Bulletin of Economics and Statistics*, **48** (1986), 201-212.
- Johansen, S., "Statistical Analysis of Cointegration vectors," *Journal of Economic Dynamics and Control*, **12** (1988), 231-254.
- Johansen, S., "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, **59** (1991), 1551-1580.
- Johansen, S., "The Role of the Constant and Linear Terms in Cointegration Analysis of Nonstationary Data," *Econometric Reviews*, **13** (1994), 205-229.
- Johansen, S. and K. Juselius, "Maximum Likelihood Estimation and Inference on Cointegration with Application to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, **52** (1990), 169-210.

- Ouliaris, S., J.Y. Park and P.C.B. Phillips, “Testing for a Unit Root in the Presence of a Maintained Trend,” Ch. 1 in *Advances in Econometrics*, ed. B. Raj, Boston: Klumer Academic Publishers, 1989.
- Perron, P. “Trends and Random Walks in Macroeconomic Time Series: Further Evidence from a New Approach,” *Journal of Economic Dynamics and Control*, **12** (1988), 297-332.
- Phillips, P.C.B., “Time Series Regression with a Unit Root,” *Econometrica*, **55** (1987), 277-301.
- Phillips, P.C.B. and P. Perron, “Testing for a Unit Root in Time Series Regression,” *Biometrika*, **75** (1988) 335-346.
- Said, E.S. and D.A. Dickey, “Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order,” *Biometrika*, **71** (1984), 599-607.
- Sims, C.A., J.H. Stock and M. Watson, “Inference in Linear Time Series Models with some Unit Roots,” *Econometrica*, **58** (1990), 113-144.
- Stock, J.H., “Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors,” *Econometrica*, **55** (1987), 1035-1056.
- Stock, J.H. and M. Watson, “Variable Trends in Economic Time Series,” *Journal of Economic Perspectives*, **2** (1988), 147-174.
- Stock, J.H. and M. Watson, “Testing for Common Trends,” *Journal of the American Statistical Association*, **83** (1988), 1097-1107.
- West, K.D., “Asymptotic Normality When Regressors Have a Unit Root,” *Econometrica*, **56** (1988), 1397-1417.

II. Метод максимального правдоподобия в эконометрии

Базовые понятия

Пусть Y — реализация N -мерной случайной величины, распределенной как:

- а) $P_{\theta}(x)$ (вероятность) — в случае дискретного распределения.
- б) $p_{\theta}(x)$ (плотность) — в случае непрерывного распределения.

Здесь $P_{\theta}(x)$ ($p_{\theta}(x)$) характеризует семейство распределений задаваемое параметром $\theta \in \Theta$, $\Theta \subset \mathbb{R}^m$ — пространство параметров. В эконометрии принято говорить об этом семействе распределений как о **порождающем данные процессе** (ПДП). Будем считать, что рассматриваемый вектор наблюдений (выборка) порожден распределением из этого семейства с параметром $\theta_0 \in \Theta$, которое будем называть истинным распределением, а θ_0 — истинным параметром.

Функция $\mathcal{L}(Y, \theta) = P_{\theta}(Y)$ (соответственно $\mathcal{L}(Y, \theta) = p_{\theta}(Y)$) называется **функцией правдоподобия**.

Оценкой максимального правдоподобия ($\hat{\theta}$), сокращенно оценкой МП, называется решение задачи

$$\mathcal{L}(Y, \theta) \rightarrow \max_{\theta \in \Theta}.$$

Будем считать в дальнейшем, что решение задачи единственно.¹

Такой метод оценивания называют **методом максимального правдоподобия**.

Обычно удобнее пользоваться **логарифмической функцией максимального правдоподобия**

$$\ell(Y, \theta) = \ln(\mathcal{L}(Y, \theta)).$$

Логарифм — (бесконечно) дифференцируемая возрастающая функция: поэтому можно находить оценки МП решая задачу $\ell(Y, \theta) \rightarrow \max_{\theta \in \Theta}$.

В частном случае вектор наблюдений представляет собой выборку независимых одинаково распределенных случайных величин: $Y_i \sim \text{IID}$, $i = 1, \dots, N$. При этом

$$\mathcal{L}(Y, \theta) = \prod_i \mathcal{L}_i(Y_i, \theta), \quad \ell(Y, \theta) = \sum_i \ell_i(Y_i, \theta).$$

Вообще говоря вектор наблюдений Y состоит из зависимых между собой и/или неодинаково распределенных случайных величин, поэтому не является

выборкой в обычном смысле слова. В общем случае это равенство тоже будет верным если обозначить

$$\mathcal{L}_i(Y_i, \theta) = p_{\theta}(Y_i | Y_{i-1}, \dots, Y_1) \quad \text{и} \quad \ell_i(Y_i, \theta) = \ln(\mathcal{L}_i(Y_i, \theta)).$$

Тем самым задается разбиение функции правдоподобия на **вклады отдельных наблюдений**.

Поскольку Y — случайная величина, то функция правдоподобия — случайная величина при данном значении параметров. Оценка максимального правдоподобия является функцией вектора наблюдений: $\hat{\theta} = \hat{\theta}(Y)$, поэтому это тоже случайная величина. Соответственно, точно так же случайными величинами является значение функции правдоподобия в максимуме $\hat{\mathcal{L}}(Y) = \mathcal{L}(Y, \hat{\theta})$ и многие другие рассматриваемые далее величины (градиент, гессиан и т. п.).

Пусть функция правдоподобия дифференцируема по θ и достигает максимума во внутренней точке ($\hat{\theta} \in \text{int}(\Theta)$), тогда оценка МП $\hat{\theta}$ должна удовлетворять следующему условию первого порядка:

$$\frac{\partial \mathcal{L}}{\partial \theta}(Y, \hat{\theta}) = 0 \quad \text{или} \quad \frac{\partial \ell}{\partial \theta}(Y, \hat{\theta}) = 0.$$

Таким образом, **градиент** логарифмической функции правдоподобия $g(\theta)$ при $\theta = \hat{\theta}$ должен быть равен нулю.

Для того, чтобы оценки, удовлетворяющие этим **уравнениям правдоподобия** действительно давали максимум правдоподобия, необходимо и достаточно, чтобы были выполнены условия второго порядка (предполагаем, что функция правдоподобия дважды дифференцируема). А именно, **матрица Гессе (гессиан)** логарифмической функции правдоподобия должна быть всюду отрицательно определена. Далее мы встретим случаи, когда это свойство действительно выполнено (логит и пробит), и когда может быть несколько локальных максимумов (“полная” функция правдоподобия для регрессии с AR(1)-ошибкой). Матрица Гессе \mathcal{H} по определению есть матрица вторых производных:

$$\mathcal{H}_{j,l}(Y, \theta) = \frac{\partial^2 \ell}{\partial \theta_j \partial \theta_l}(Y, \theta) \quad j, l = 1, \dots, m.$$

¹ Пример неединственности представляет оценивание процесса скользящего среднего в ошибке.

С помощью матричного дифференцирования можно записать гессиан в виде

$$\mathcal{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

В некоторых моделях функция правдоподобия неограничена сверху и не существует оценок максимального правдоподобия в смысле приведенного выше определения. Согласно альтернативному определению оценками максимального правдоподобия называют корни уравнения правдоподобия, являющиеся локальными максимумами функции правдоподобия, корнями уравнения правдоподобия. Существуют модели, для которых такие оценки состоятельны.

Информационной матрицей для вектора наблюдений размерностью N будем называть матрицу

$$\mathcal{I}^N(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\mathbf{g}(Y, \boldsymbol{\theta}) \mathbf{g}^T(Y, \boldsymbol{\theta})).$$

Заметим, что по этому определению информационная матрица — функция некоторого вектора параметров $\boldsymbol{\theta} \in \Theta$. В дальнейшем для сокращения записи, если это не вносит путаницы, индекс количества наблюдений N будем опускать: $\mathcal{I}(\boldsymbol{\theta})$. Индекс $\boldsymbol{\theta}$ у символа математического ожидания E означает, что ожидание вычисляется в предположении, что $\boldsymbol{\theta}$ — точка истинных параметров. Заметьте, что оператор E без нижнего индекса везде означает ожидание для распределения с параметрами $\boldsymbol{\theta}_0$!

В дальнейшем будет использоваться следующее очевидное свойство функции правдоподобия. Пусть $\varphi(Y)$ есть некоторая функция вектора наблюдений Y . Тогда ее математическое ожидание равно

$$E(\varphi(Y)) = \int_{\mathcal{Y}} \varphi(Y) \mathcal{L}(\boldsymbol{\theta}_0, Y) dY,$$

где \mathcal{Y} обозначает пространство элементарных событий (пространство переменной Y).

Таким образом, можно переписать определение информационной матрицы в виде

$$\mathcal{I}(\boldsymbol{\theta}) = \int_{\mathcal{Y}} \mathbf{g}(Y, \boldsymbol{\theta}) \mathbf{g}^T(Y, \boldsymbol{\theta}) \mathcal{L}(\boldsymbol{\theta}, Y) dY.$$

Асимптотическая информационная матрица есть предел

$$\mathcal{I}^{\infty}(\boldsymbol{\theta}) = \lim_{N \rightarrow \infty} \frac{1}{N} \mathcal{I}^N(\boldsymbol{\theta}).$$

Множитель $1/N$ добавлен в определения для того, чтобы существовал конечный предел (информационная матрица является величиной порядка $O(N)$).

Если мы рассматриваем выборку, то применяя определение информационной матрицы к отдельным наблюдениям (\mathcal{I}_i), имеем

$$\mathcal{I}^N = N \mathcal{I}_i.$$

Таким образом, если наблюдения независимы и одинаково распределены, то информация растет пропорционально количеству наблюдений.

Пример. Линейная регрессия с нормально распределенными ошибками.

Пусть ошибки $\varepsilon_i \sim \text{NID}(0, \sigma^2)$. Эта аббревиатура означает, что случайные величины ε_i независимы и имеют нормальное распределение с параметрами $(0, \sigma^2)$ (normally and independently distributed). Ковариационная матрица вектора ошибок — это единичная матрица с точностью до множителя: $E(\varepsilon\varepsilon^\top) = \sigma^2 \mathbf{I}_N$.

Зависимая переменная связана с ошибками следующим образом:

$$Y = X\beta + \varepsilon,$$

где X — матрица регрессоров ($N \times m$), β — вектор-столбец неизвестных коэффициентов длины m . Таким образом, Y_i имеет нормальное распределение с параметрами $(X_i\beta, \sigma^2)$, где X_i — i -я строка матрицы X :

$$Y_i \sim N(X_i\beta, \sigma^2).$$

Плотность распределения $N(\alpha, \sigma^2)$ равна

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \alpha)^2}{2\sigma^2}\right).$$

Функция правдоподобия для этого набора наблюдений имеет вид

$$\mathcal{L} = (2\pi\sigma^2)^{-N/2} \prod_{i=1}^N \exp\left(-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}\right).$$

Логарифмическая функция правдоподобия:

$$\begin{aligned} \ell &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (Y_i - X_i\beta)^2 = \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y - X\beta)^\top (Y - X\beta) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} e^\top e. \end{aligned}$$

Здесь мы обозначили вектор остатков $e = Y - X\beta$.

В данном случае вектор неизвестных параметров состоит из двух компонент:

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix}.$$

Градиент логарифмической функции правдоподобия тоже состоит из двух частей:

$$\mathbf{g}_{\boldsymbol{\beta}} = \frac{\partial \ell}{\partial \boldsymbol{\beta}^\top} = \frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{e}.$$

$$g_{\sigma^2} = \frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{\mathbf{e}^\top \mathbf{e}}{2\sigma^4} = \frac{1}{2\sigma^4} (\text{RSS}(\boldsymbol{\beta}) - N\sigma^2),$$

где $\text{RSS}(\boldsymbol{\beta}) = \mathbf{e}^\top \mathbf{e}$ — сумма квадратов остатков.

Оценка максимального правдоподобия $\hat{\boldsymbol{\theta}}$ должна удовлетворять равенству $\mathbf{g}(\hat{\boldsymbol{\theta}}) = 0$, откуда получим

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \quad \text{и} \quad \hat{\sigma}^2 = \frac{\text{RSS}(\hat{\boldsymbol{\beta}})}{N} = \frac{1}{N} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

ММП дает ту же оценку вектора коэффициентов регрессии $\boldsymbol{\beta}$, что и МНК. Как известно, оценка дисперсии $\hat{\sigma}^2$ является смещенной:

$$E(\hat{\sigma}^2) = \frac{N-m}{N} \sigma^2.$$

Покажем, каким образом связаны ММП и МНК.

Выразим, используя равенство $g_{\sigma^2} = 0$, дисперсию через $\boldsymbol{\beta}$:

$$\sigma^2(\boldsymbol{\beta}) = \frac{\text{RSS}(\boldsymbol{\beta})}{N}.$$

Если подставить ее в функцию правдоподобия, то получится **концентрированная функция правдоподобия**:

$$\ell^c = -\frac{N}{2} \ln(2\pi\sigma^2(\boldsymbol{\beta})) - \frac{1}{2\sigma^2(\boldsymbol{\beta})} \text{RSS}(\boldsymbol{\beta}) = -\frac{N}{2} \ln\left(2\pi \frac{\text{RSS}(\boldsymbol{\beta})}{N}\right) - \frac{N}{2}.$$

Максимизация ее эквивалентна минимизации суммы квадратов остатков $\text{RSS}(\boldsymbol{\beta})$ по $\boldsymbol{\beta}$.

Гессиан логарифмической функции правдоподобия состоит из следующих компонент:

$$\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X},$$

$$\mathcal{H}_{\boldsymbol{\beta}\sigma^2} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^2} = -\frac{1}{\sigma^4} \mathbf{e}^\top \mathbf{X},$$

$$\mathcal{H}_{\sigma^2\boldsymbol{\beta}} = \frac{\partial^2 \ell}{\partial \sigma^2 \partial \boldsymbol{\beta}^\top} = -\frac{1}{\sigma^4} \mathbf{X}^\top \mathbf{e},$$

$$\mathcal{H}_{\sigma^2\sigma^2} = \frac{\partial^2 \ell}{(\partial \sigma^2)^2} = \frac{N}{2\sigma^4} - \frac{\mathbf{e}^\top \mathbf{e}}{\sigma^6}.$$

В точке истинных параметров $\boldsymbol{e} = \boldsymbol{\varepsilon}$. Используя это, получим, что компоненты информационной матрицы, вычисленной в точке истинных параметров равны:

$$\begin{aligned}\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0) &= \mathbb{E}(\mathbf{g}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)\mathbf{g}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^\top) = \mathbb{E}\left(\frac{1}{\sigma^4} \mathbf{X}^\top \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{X}\right) = \\ &= \frac{1}{\sigma^4} \mathbf{X}^\top \mathbb{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top) \mathbf{X} = \frac{1}{\sigma^4} \sigma^2 \mathbf{X}^\top \mathbf{I} \mathbf{X} = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X},\end{aligned}$$

$$\mathcal{I}_{\boldsymbol{\beta}\sigma^2}(\boldsymbol{\theta}_0) = \mathbb{E}(\mathbf{g}_{\sigma^2}(\boldsymbol{\theta}_0)\mathbf{g}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)^\top) = \mathbb{E}\left(\frac{1}{\sigma^6} (\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - N\sigma^2) \boldsymbol{\varepsilon}^\top \mathbf{X}\right) = \mathbf{0}^\top,$$

$$\mathcal{I}_{\sigma^2\boldsymbol{\beta}}(\boldsymbol{\theta}_0) = \mathbf{0} \text{ (аналогично),}$$

$$\begin{aligned}\mathcal{I}_{\sigma^2\sigma^2}(\boldsymbol{\theta}_0) &= \mathbb{E}(\mathbf{g}_{\sigma^2}(\boldsymbol{\theta}_0)^2) = \mathbb{E}\left(\frac{1}{4\sigma^8} (\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - N\sigma^2)^2\right) =, \\ &= \frac{1}{4\sigma^8} \mathbb{E}((\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon})^2) - 2N \frac{1}{4\sigma^6} \mathbb{E}(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}) + N^2 \frac{1}{4\sigma^4} = \\ &= \frac{1}{4\sigma^8} (\sum_i \mathbb{E}(\varepsilon_i^4) + \sum_{i \neq s} \mathbb{E}(\varepsilon_i^2 \varepsilon_s^2)) - 2N \frac{1}{4\sigma^6} \sum_i \mathbb{E}(\varepsilon_i^2) + N^2 \frac{1}{4\sigma^4} = \\ &= \frac{1}{4\sigma^8} 3N\sigma^4 + \frac{1}{4\sigma^8} (N^2 - N)\sigma^4 - 2N \frac{1}{4\sigma^6} N\sigma^2 + N^2 \frac{1}{4\sigma^4} = N \frac{1}{2\sigma^4}.\end{aligned}$$

В данном случае \mathcal{I} — блочно-диагональная матрица по параметрам $\boldsymbol{\beta}$ и σ^2 . В дальнейшем мы рассмотрим, какие полезные свойства из этого вытекают.

Характеристика ММП.

В статистике применяются три основных метода оценивания:

- Метод наименьших квадратов.
- (Обобщенный) метод моментов.
- Метод максимального правдоподобия.

Интересно сравнить ММП с двумя другими методами.

Условия, при которых можно использовать ММП более ограничительны. Метод требует явного задания вида распределения.

С другой стороны, ММП более универсален. Его можно использовать для любых моделей, задающих вид распределения наблюдаемых переменных. Два другие метода можно использовать лишь тогда, когда распределение переменных можно представить в определенном виде. Если есть гипотеза о точном виде распределения, то всегда понятно, как получать оценки пара-

метров, распределений параметров и различных статистик, как проверять гипотезы, хотя сами расчеты могут быть сложными.

Еще одно свойство — **инвариантность** по отношению к переобозначению параметров. Пусть $\varphi(\cdot): \mathbb{R}^k \rightarrow \mathbb{R}^k$ однозначная обратимая функция. Можно подставить в функцию правдоподобия вместо θ величину $\varphi(\tau)$, где τ — новый вектор параметров, $\tau \in \varphi^{-1}(\theta)$. При этом, если $\hat{\tau}$ — оценка МП в новой задаче, то $\hat{\theta}$ — оценка МП в старой задаче.

Из инвариантности следует, что оценка МП как правило не может быть несмещенной. Пусть, например, $E(\hat{\theta}) = \theta_0$, где θ_0 — истинное значение параметра. Тогда оценка $\hat{\tau}$, полученная нелинейным преобразованием $\theta = \varphi(\tau)$ будет смещенной: $E(\hat{\tau}) \neq \tau_0$, где $\tau_0 = E(\varphi^{-1}(\hat{\theta}))$.

Если правильно выбрать параметризацию, то распределение оценок в малых выборках может быть близко к асимптотическому, если неправильно, то асимптотическое распределение будет очень плохой аппроксимацией.

ММП получил широкое распространение благодаря своим хорошим асимптотическим свойствам:

- состоятельность,
- асимптотическая нормальность,
- асимптотическая эффективность.

С точки зрения эффективности сильные предположения о виде распределения, которые приходится делать, применяя ММП, окупаются (в большей или меньшей степени). Поскольку мы делаем очень ограничительные предположения, то можем доказать более сильные утверждения.

Связь ММП с МНК. Квази-МП методы.

Хотя оценки МП являются специфическими по отношению к определенному виду распределения, значение метода может быть шире.

Идея состоит в том, чтобы процедуру получения оценок для одного распределения распространить на “близкие” распределения. Также методы получили название квази- или псевдо-ММП.

Метод максимального правдоподобия используют для нахождения способа расчетов, а затем уже доказывают, какими свойствами обладает этот метод по отношению к некоторому более широкому классу распределений.

Как мы видели, например, ММП в случае регрессии с нормально распределенными ошибками дает МНК, который на самом деле обладает “хороши-

ми” свойствами и при ошибках, которые уже не имеют нормального распределения (хотя эффективность теряется).

Есть и обратная связь между этими двумя методами. МНК можно использовать как вычислительную процедуру, которая помогает находить оценки МП и строить тесты. Такое техническое использование МНК называют **вспомогательной регрессией**. Кроме того, вслед за Дэвидсоном и МакКинноном будем использовать термин **искусственная регрессия**, если вспомогательную регрессию можно применять как для нахождения оценок, так и для проверки гипотез относительно полученных оценок и проверки правильности спецификации модели.

Связь гессиана и матрицы вкладов в градиент с информационной матрицей

Гессиан и информационная матрица

Покажем, какая связь существует между информационной матрицей и гессианом. Сделаем это только в случае непрерывного распределения. Тот же метод доказательства очевидным образом распространяется на дискретные распределения. Применяя правило дифференцирования логарифма к логарифмической функции правдоподобия, получим следующее тождество:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}.$$

Продифференцируем по $\boldsymbol{\theta}^\top$:

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \frac{1}{\mathcal{L}} \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} - \frac{1}{\mathcal{L}^2} \frac{\mathcal{L}}{\partial \boldsymbol{\theta}^\top} \frac{\mathcal{L}}{\partial \boldsymbol{\theta}}.$$

Отсюда, опять воспользовавшись правилом дифференцирования логарифма, получим

$$\mathcal{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \frac{1}{\mathcal{L}} \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} - \frac{\ell}{\partial \boldsymbol{\theta}^\top} \frac{\ell}{\partial \boldsymbol{\theta}}.$$

Найдем теперь ожидание обеих частей в точке $\boldsymbol{\theta}_0$ (при истинных параметрах распределения):

$$\begin{aligned} \mathbb{E}(\mathcal{H}(Y, \boldsymbol{\theta}_0)) &= \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}(\boldsymbol{\theta}_0)\right) = \\ &= \int_{\mathcal{Y}} \mathcal{L}(\boldsymbol{\theta}_0, Y) \frac{1}{\mathcal{L}(\boldsymbol{\theta}_0; Y)} \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}_0; Y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} dY - \mathbb{E}\left(\frac{\ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} \frac{\ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right). \end{aligned}$$

Второй член разности есть по определению информационная матрица $\mathcal{I}(\boldsymbol{\theta}_0)$. Интеграл равен нулю при условии, что операции интегрирования и

дифференцирования перестановочны (для этого достаточно, в частности, чтобы пространство зависимой переменной \mathcal{Y} не зависело от θ или плотность распределения по краям \mathcal{Y} была стремилась к нулю):

$$\frac{\partial^2}{\partial \theta \partial \theta^\top} \int \mathcal{L}(\theta_0, \mathbf{Y}) d\mathbf{Y} = \frac{\partial^2}{\partial \theta \partial \theta^\top} 1 = 0.$$

Таким образом, используя для краткости обозначения $\mathcal{H}(\mathbf{Y}, \theta_0) = \mathcal{H}_0$ и $\mathcal{I}(\theta_0) = \mathcal{I}_0$,

$$-E(\mathcal{H}_0) = \mathcal{I}_0$$

— информационная матрица равна математическому ожиданию гессиана функции правдоподобия со знаком минус. То же самое свойство верно асимптотически (опять обозначаем $\mathcal{I}^\infty(\theta_0) = \mathcal{I}_0^\infty$):

$$-\lim_{N \rightarrow \infty} \frac{1}{N} E(\mathcal{H}_0) = \mathcal{I}_0^\infty.$$

Матрица вкладов в градиент и информационная матрица

Прежде всего докажем, что математическое ожидание градиента в точке θ_0 равно нулю ($E \mathbf{g}(\mathbf{Y}, \theta_0) = 0$):²

$$\begin{aligned} E \mathbf{g}(\mathbf{Y}, \theta_0) &= \int \mathbf{g}(\mathbf{Y}, \theta_0) \mathcal{L}(\mathbf{Y}, \theta_0) d\mathbf{Y} = \int \frac{\partial \ell}{\partial \theta}(\mathbf{Y}, \theta_0) \mathcal{L}(\mathbf{Y}, \theta_0) d\mathbf{Y} = \\ &= \int \frac{1}{\mathcal{L}(\mathbf{Y}, \theta_0)} \frac{\partial \mathcal{L}}{\partial \theta}(\mathbf{Y}, \theta_0) \mathcal{L}(\mathbf{Y}, \theta_0) d\mathbf{Y} = \int \frac{\partial \mathcal{L}}{\partial \theta}(\mathbf{Y}, \theta_0) d\mathbf{Y} = \\ &= \frac{\partial}{\partial \theta} \int \mathcal{L}(\mathbf{Y}, \theta_0) d\mathbf{Y} = \frac{\partial}{\partial \theta} 1 = 0. \end{aligned}$$

Как уже говорилось, функцию правдоподобия можно разбить по вкладам отдельных наблюдений: $\ell(\mathbf{Y}, \theta) = \sum_i \ell_i(Y_i, \theta)$. То же самое можно проделать с градиентом. Определим **матрицу вкладов в градиент** отдельных наблюдений \mathbf{G} как

$$G_{ij}(\theta) = \frac{\partial \ell_i}{\partial \theta_j}(\theta).$$

$$\text{При этом } \sum_i G_{ij} = \sum_i \frac{\partial \ell_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_i \ell_i = \frac{\partial \ell}{\partial \theta_j} = g_j.$$

² Мы опять предполагаем здесь, что операции интегрирования и дифференцирования перестановочны.)

Используя рассуждения, аналогичные приведенным выше, можно показать, что $E G_{ij}(Y, \theta_0) = 0$.

Мы так разделили функцию правдоподобия на вклады отдельных наблюдений, что $E(\mathbf{G}_i(Y, \theta_0) \mathbf{G}_s(Y, \theta_0)^\top) = 0$, где $\mathbf{G}_i(Y, \theta_0)$ и $\mathbf{G}_s(Y, \theta_0)$ — строки матрицы $\mathbf{G}_0 = \mathbf{G}(Y, \theta_0)$, относящиеся к разным наблюдениям i и s . (Поскольку элементы матрицы \mathbf{G}_0 имеют нулевое математическое ожидание, то это означает что строки матрицы \mathbf{G}_0 , относящиеся к разным наблюдениям, некоррелированы.) Докажем это свойство.

Функция правдоподобия i -го наблюдения по определению есть плотность распределения Y_i (в случае непрерывного распределения) условная по информации, содержащейся в наблюдениях $1, \dots, i-1$ (условная по Y_1, \dots, Y_{i-1}). Обозначим это информационное множество Ω_i . Будем вычислять математическое ожидание по частям — сначала условное, а потом от него безусловное (правило полного мат. ожидания). Предположим, что $i < s$. Тогда

$$\begin{aligned} E(\mathbf{G}_i(Y, \theta_0) \mathbf{G}_s(Y, \theta_0)^\top) &= E(E(\mathbf{G}_i(Y, \theta_0) \mathbf{G}_s(Y, \theta_0)^\top | \Omega_i)) = \\ &= E(\mathbf{G}_i(Y, \theta_0) E(\mathbf{G}_s(Y, \theta_0)^\top | \Omega_i)) = 0. \end{aligned}$$

Равенство $E(\mathbf{G}_s(Y, \theta_0)^\top | \Omega_i) = 0$ доказывается в точности по той же схеме, что и доказанное выше $E \mathbf{g}(Y, \theta_0) = 0$.

Используя это свойство, получим

$$E(\mathbf{G}_0^\top \mathbf{G}_0) = E\left(\sum_i \mathbf{G}_{0i}^\top \mathbf{G}_{0i}\right) = E\left(\left(\sum_i \mathbf{G}_{0i}\right)^\top \left(\sum_i \mathbf{G}_{0i}\right)\right) = E(\mathbf{g}_0 \mathbf{g}_0^\top).$$

Последнее выражение есть по определению информационная матрица. Таким образом,

$E(\mathbf{G}_0^\top \mathbf{G}_0) = \mathcal{I}_0.$
--

Вычисление информационной матрицы

Рассмотрим теперь, как вычислить для конкретной модели информационную матрицу $\mathcal{I}(\theta)$. Здесь существуют три способа. Понятно, что все три способа должны для “хороших” моделей давать один и тот же результат. Во-первых, можно воспользоваться определением информационной матрицы: $\mathcal{I} = E(\mathbf{g}\mathbf{g}^\top)$. Во-вторых, можно воспользоваться равенством $\mathcal{I}_0 = -E(\mathcal{H}_0)$.

Самым простым часто (а именно тогда, когда функцию правдоподобия можно простым образом разбить на вклады наблюдений) оказывается третий способ, который использует только что рассмотренное свойство

$$\mathcal{I}_0 = E(\mathbf{G}_0^\top \mathbf{G}_0) = \sum_i E(\mathbf{G}_{i0}^\top \mathbf{G}_{i0}).$$

Выше была получено выражение для информационной матрицы в случае линейной регрессии с нормально распределенными ошибками прямо по определению. Вычислим теперь ее двумя другими способами.

Гессиан уже был вычислен выше. Математическое ожидание от него со знаком минус равно.

$$\mathcal{I}_0 = -E(\mathcal{H}_0) = -E \left(\begin{bmatrix} -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} & -\frac{1}{\sigma^4} \mathbf{X}^\top \boldsymbol{\varepsilon} \\ -\frac{1}{\sigma^4} \boldsymbol{\varepsilon}^\top \mathbf{X} & \frac{N}{2\sigma^4} - \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\sigma^6} \end{bmatrix} \right) = \begin{bmatrix} \frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0}^\top & \frac{N}{2\sigma^4} \end{bmatrix}.$$

Вклад в логарифмическую функцию правдоподобия i -го наблюдения равен

$$\ell_i = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2.$$

Продифференцировав его, получим вклад в градиент i -го наблюдения в точке истинных параметров:

$$\mathbf{G}_{i0} = \left(\frac{1}{\sigma^2} \mathbf{X}_i^\top \boldsymbol{\varepsilon}_i, \frac{\boldsymbol{\varepsilon}_i^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right).$$

Вклад в информационную матрицу i -го наблюдения в точке истинных параметров равен

$$\mathcal{I}_{i0} = E(\mathbf{G}_{i0}^\top \mathbf{G}_{i0}) = \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{X}_i^\top \mathbf{X}_i & \mathbf{0} \\ \mathbf{0}^\top & \frac{1}{2\sigma^4} \end{bmatrix}.$$

Таким образом,

$$\mathcal{I}_0 = \sum_i \mathcal{I}_{i0} = \begin{bmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{0} \\ \mathbf{0}^\top & \frac{N}{2\sigma^4} \end{bmatrix}.$$

Все три способа, как и следовало ожидать, привели к одному и тому же результату.

Заметим попутно, что \mathcal{I}_{i0} — положительно определенная матрица, \mathcal{I}_0 при любом количестве наблюдений — положительно определенная матрица (в предположении, что матрица регрессоров имеет полный ранг). Из этого можно сделать вывод, что информация в некотором смысле увеличивается с ростом количества наблюдений. Это одно из объяснений названия "информаци-

онная матрица". В частности, определитель информационной матрицы увеличивается с ростом количества наблюдений:

$$|\mathcal{I}_0^{N+1}| > |\mathcal{I}_0^N|.$$

Распределение градиента и оценок максимального правдоподобия

Асимптотическое распределение градиента и оценок максимального правдоподобия

Оценки максимального правдоподобия имеют нормальное асимптотическое распределение. Для доказательства этого мы воспользуемся предположением, что градиент функции правдоподобия в точке истинных значений параметров θ_0 имеет асимптотическое нормальное распределение.

Градиент $\mathbf{g}(\mathbf{Y}, \theta_0)$ будет иметь нормальное распределение (асимптотически), если к нему применима *центральная предельная теорема*. Надо представить \mathbf{g}_0 как сумму некоторой последовательности случайных величин. Для этого подходит разложение градиента на вклады отдельных наблюдений

$$g_i(\mathbf{Y}, \theta_0) = \sum_j G_{ij}(\mathbf{Y}, \theta_0).$$

Как сказано выше, каждое из слагаемых здесь имеет нулевое математическое ожидание. Если выполнены некоторые условия регулярности (см. литературу, посвященную центральной предельной теореме), то $\sum G_{ij}(\mathbf{Y}, \theta_0)$ стремится к нормальному распределению с ростом количества наблюдений. Ковариационная матрица градиента в точке θ_0 есть информационная матрица, поскольку его математическое ожидание равно нулю: $V(\mathbf{g}_0) = E(\mathbf{g}_0 \mathbf{g}_0^\top) = \mathcal{I}$. Последнее равенство выполнено по определению.

Окончательно получаем

$$\frac{1}{\sqrt{N}} \mathbf{g}_0 \stackrel{a}{\sim} N(\mathbf{0}, \mathcal{I}_0^\infty).$$

Используя это свойство градиента мы докажем асимптотическую нормальность оценок ММП. Для этого используем разложение в ряд Тейлора в точке θ_0 до членов первого порядка:

$$\mathbf{0} = \mathbf{g}(\hat{\theta}) = \mathbf{g}(\theta_0) + \mathcal{H}(\bar{\theta})(\hat{\theta} - \theta_0),$$

где \mathcal{H} — гессиан (матрица вторых производных от логарифмической функции правдоподобия), $\bar{\theta}_j$ — выпуклая комбинация $\hat{\theta}_j$ и θ_{0j} . Поскольку $\hat{\theta}_j$ — состоятельная оценка параметра θ_{0j} , то $\bar{\theta}_j$ тоже должна быть состоятельной

оценкой θ_{0j} . Поскольку $-1/N \mathcal{H}_0 \stackrel{a}{=} \mathcal{I}_0^\infty$, то имеем асимптотическое равенство:
 $-1/N \mathcal{H}(\bar{\theta}) \stackrel{a}{=} \mathcal{I}_0^\infty$.

Таким образом, $\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{=} (\mathcal{I}_0^\infty)^{-1} \frac{1}{\sqrt{N}} \mathbf{g}_0 \stackrel{a}{\sim} N(\mathbf{0}, (\mathcal{I}_0^\infty)^{-1} \mathcal{I}_0^\infty (\mathcal{I}_0^\infty)^{-1})$.

Окончательно получим

$$\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} N(\mathbf{0}, (\mathcal{I}_0^\infty)^{-1}).$$

Это соотношение позволяет оценить ковариационную матрицу оценок $\hat{\theta}$. С этой точки зрения оценка обратной информационной матрицы является оценкой ковариационной матрицы МП-оценок \hat{V}_θ (с точностью до множителя $1/N$), и эти термины можно использовать как синонимы. Понятно, что для этого должны быть выполнены соответствующие условия, гарантирующие, что операции интегрирования и дифференцирования коммутируют и что справедлива центральная предельная теорема, что мы всегда в дальнейшем будем предполагать.

Выборочная оценка распределения градиента и оценок максимального правдоподобия

Для получения выборочной оценки распределений МП-оценок $\hat{\theta}$ и градиента, можно воспользоваться формулами для их асимптотических распределений. Все эти величины асимптотически нормально распределены и их асимптотические матрицы ковариаций являются функциями асимптотической информационной матрицы в точке истинных параметров (\mathcal{I}_0^∞) . Таким образом, требуется получить состоятельную оценку \mathcal{I}_0^∞ , чтобы подставить ее в соответствующие формулы. Будем обозначать символом $\hat{\mathcal{I}}$ такую матрицу, что $1/N \hat{\mathcal{I}}$ — состоятельная оценка \mathcal{I}_0^∞ :

$$\text{Plim}_{N \rightarrow \infty} \frac{1}{N} \hat{\mathcal{I}} = \mathcal{I}_0^\infty.$$

Поскольку $\hat{\theta}$ — состоятельная оценка истинных параметров θ_0 , то $1/N \mathcal{I}(\hat{\theta})$ — состоятельная оценка \mathcal{I}_0^∞ . Это дает первый способ оценивания. Он состоит в том, чтобы сначала для данной модели найти функцию $\mathcal{I}(\theta)$, а затем подставить в нее оценки максимального правдоподобия $\hat{\theta}$ (конечно, по-

дойдут и любые другие состоятельные оценки). Методы нахождения $\mathcal{I}(\theta)$ описаны ниже.

Другой способ основывается на равенстве для информационной матрицы $\mathcal{I}_0^\infty = -\lim_{N \rightarrow \infty} 1/N E(\mathcal{H}_0)$ и на том, что ожидаемый гессиан $E(\mathcal{H}_0)$ асимптотически равен **эмпирическому гессиану** $\hat{\mathcal{H}} = \mathcal{H}(Y, \hat{\theta}) = \frac{\partial^2 \ell}{\partial \theta \partial \theta^\top}(Y, \hat{\theta})$. Этот способ обычно проще предыдущего, поскольку не требует вычисления математических ожиданий. Получить матрицу вторых производных данной функции правдоподобия можно и с помощью компьютерной программы.

Особой простотой, и потому притягательностью (требуется найти только первые производные), отличается третий способ оценивания информационной матрицы, использующий матрицу вкладов в градиент G . Этот способ предложен в статье Berndt, Hall, Hall, and Hausman (1974) и поэтому называется ВННН. Другое название — метод внешнего произведения градиента (outer product of the gradient, сокращенно OPG). Этот способ основан на том, что $E(G_0^\top G_0) = \mathcal{I}_0$. Предлагается использовать матрицу $G(Y, \hat{\theta})^\top G(Y, \hat{\theta})$ в качестве $\hat{\mathcal{I}}$.

Таким образом, имеем три варианта матрицы $\hat{\mathcal{I}}$:

$$\text{I. } \mathcal{I}(\hat{\theta}); \quad \text{II. } \mathcal{H}(Y, \hat{\theta}); \quad \text{III. } G(Y, \hat{\theta})^\top G(Y, \hat{\theta}).$$

Как показывают эксперименты методом Монте-Карло, тесты, использующие $G(Y, \hat{\theta})^\top G(Y, \hat{\theta})$ самые неточные в конечных выборках, а тесты, основанные на $\mathcal{I}(\hat{\theta})$ обычно не уступают тестам, основанным на $\mathcal{H}(Y, \hat{\theta})$.

Три рассмотренных способа нахождения $\hat{\mathcal{I}}$ подходят для любых распределений. Есть также более специфические методы, которые можно использовать только в случае моделей определенного вида. Например, метод Гаусса-Ньютона используется в нелинейных регрессиях, метод удвоенной регрессии — в квазирегрессионных моделях с неизвестными параметрами в правой части.

Особого рассмотрения требует нахождение оценки ковариационной матрицы оценок в случае квази-МП методов (их называют также псевдо-МП методами). Если предполагается, что ошибки в модели имеют нормальное распределение и гомоскедастичны, а на самом деле это не так, то часто только что рассмотренные методы дают несостоятельные оценки. Оказывается, что во многих случаях следующие оценки состоятельны (конечно, при вычисле-

нии этих величин используется не настоящая, а псевдо функция правдоподобия):

$$\mathcal{H}(Y, \hat{\theta})^{-1} \mathcal{I}(\hat{\theta}) \mathcal{H}(Y, \hat{\theta})^{-1}.$$

$$\mathcal{H}(Y, \hat{\theta})^{-1} \mathbf{G}(Y, \hat{\theta})^\top \mathbf{G}(Y, \hat{\theta}) \mathcal{H}(Y, \hat{\theta})^{-1}.$$

Поясним интуитивно, откуда берутся эти формулы. При выводе асимптотического распределения оценок максимального правдоподобия, мы пользовались тем, что “усредненный” гессиан $-1/N \mathcal{H}_0$ равен асимптотически \mathcal{I}_0^∞ . В общем случае нужно воспользоваться пределом $1/N E(\mathcal{H})$ — “асимптотическим” ожидаемым гессианом в точке истинных оценок (\mathcal{H}_0^∞) . Формула приобретет следующий вид:

$$\sqrt{N} (\hat{\theta} - \theta_0) \stackrel{a}{\sim} (\mathcal{H}_0^\infty)^{-1} \frac{1}{\sqrt{N}} \mathbf{g}_0 \stackrel{a}{\sim} N(\mathbf{0}, (\mathcal{H}_0^\infty)^{-1} \mathcal{I}_0^\infty (\mathcal{H}_0^\infty)^{-1}).$$

Численные методы нахождения оценок максимального правдоподобия

Рассмотрим семейство универсальных алгоритмов вычисления оценок максимального правдоподобия, тесно связанных с только что рассмотренными способами получения матрицы $\hat{\mathcal{I}}$. Эти алгоритмы являются итеративными градиентными методами и t -й шаг алгоритма задается формулой

$$\theta^{t+1} = \theta^t + (\hat{\mathcal{I}}^t)^{-1} \mathbf{g}(\theta^t).$$

Стационарная точка этого процесса $\theta^{t+1} = \theta^t$ будет удовлетворять уравнениям правдоподобия $\mathbf{g}=0$ и (с соответствующими оговорками) будет оценкой максимального правдоподобия.

Если в качестве $\hat{\mathcal{I}}^t$ взять информационную матрицу в точке оценок $\mathcal{I}(\theta^t)$, то мы получаем метод, называемый по-английски **method of scoring**:

$$\theta^{t+1} = \theta^t - \mathcal{I}(\theta^t)^{-1} \mathbf{g}(\theta^t).$$

Если в качестве $\hat{\mathcal{I}}^t$ взять минус гессиан $-\mathcal{H}(\theta^t)$, то мы получаем классический **метод Ньютона**:

$$\theta^{t+1} = \theta^t - \mathcal{H}(\theta^t)^{-1} \mathbf{g}(\theta^t).$$

Метод Ньютона, как правило, быстрее сходится в ближайшей окрестности оценок МП, зато метод, использующий информационную матрицу обычно менее чувствителен к выбору начальных приближений.

Шаг метода ВННН (OPG) можно получить с помощью вспомогательной (искусственной) регрессии, зависимой переменной в которой будет вектор, составленный из единиц (обозначим его $\mathbf{1}$), а матрицей регрессоров — матрица $\mathbf{G}(\boldsymbol{\theta}^t)$. Если $\Delta\boldsymbol{\theta}^t$ — оценки коэффициентов в этой вспомогательной регрессии на t -м шаге, то итерация имеет вид

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta\boldsymbol{\theta}^t, \quad \text{где } \Delta\boldsymbol{\theta}^t = (\mathbf{G}(\boldsymbol{\theta}^t)^\top \mathbf{G}(\boldsymbol{\theta}^t))^{-1} \mathbf{G}(\boldsymbol{\theta}^t)^\top \mathbf{1}.$$

Хотя этот последний алгоритм является самым простым, но, как правило, сходится очень медленно. Если учесть, что обычно при использовании этого метода $(\mathbf{G}(\boldsymbol{\theta}^t)^\top \mathbf{G}(\boldsymbol{\theta}^t))^{-1}$ берут в качестве оценки ковариационной матрицы оценок, то использовать его нежелательно.

Возможны различные модификации этой основной идеи.

Шаг алгоритма можно вычислять, домножая исходный шаг на параметр λ :

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \lambda (\hat{\mathcal{I}}^t)^{-1} \mathbf{g}(\boldsymbol{\theta}^t).$$

Разумно выбирать параметр λ , максимизируя по нему функцию правдоподобия в точке $\boldsymbol{\theta}^{t+1}$:

$$\lambda = \operatorname{argmin} \ell(\boldsymbol{\theta}^t + \lambda (\hat{\mathcal{I}}^t)^{-1} \mathbf{g}(\boldsymbol{\theta}^t)).$$

В частном случае матрица \mathcal{I}_0 является блочно-диагональной. Тогда шаг алгоритма можно разбить на несколько “подшагов”, один для каждого блока. Изменяются при этом только параметры, соответствующие данному блоку.

Если из условий первого порядка выразить одни оцениваемые параметры через другие и подставить их в функцию правдоподобия, то получится **концентрированная функция правдоподобия**. Действуя таким образом, задачу поиска оценок МП можно упростить, сведя к задаче максимизации концентрированной функции правдоподобия по меньшему числу параметров. Задача может упроститься до одномерного поиска.

Существует много других алгоритмов. Есть алгоритмы специально сконструированные для конкретной модели; с примерами их мы встретимся в дальнейшем. Есть универсальные методы, которые можно применять к широкому классу моделей, такие как метод удвоенной регрессии и итеративный обобщенный МНК. Можно, конечно, использовать универсальные оптимизационные алгоритмы, которые подходят не только для максимизации функции правдоподобия.

ММП и проверка гипотез

Асимптотическое распределение и асимптотическая эквивалентность трех классических статистик

Предположим, что мы хотим проверить гипотезу о том, что вектор истинных параметров θ_0 удовлетворяет набору ограничений, который в векторном виде можно записать как

$$r(\theta_0) = 0.$$

Тогда с учетом этой информации задача получения оценки максимального правдоподобия эквивалентна задаче нахождения седловой точки лагранжиана:

$$L(\theta, \lambda) = \ell(\theta) - r^\top(\theta) \lambda.$$

Ограниченная оценка $\tilde{\theta}$ должна вместе с вектором множителей Лагранжа $\tilde{\lambda}$ удовлетворять следующей системе условий первого порядка:

$$g(\tilde{\theta}) = R^\top(\tilde{\theta}) \tilde{\lambda},$$

$$r(\tilde{\theta}) = 0,$$

где $R(\theta)$ — матрица первых производных ограничений: $R = \frac{\partial r}{\partial \theta}$.

Для вывода распределений интересующих нас статистик используем тот же прием, с помощью которого выше получено распределение оценок. Поскольку мы предполагаем, что оценки МП состоятельны и нас интересуют асимптотические распределение, то для разложений в ряд Тейлора будем писать приближенные равенства. Более строгие рассуждения должны быть аналогичны использованным выше.

Разложим градиент и ограничения в ряд Тейлора до членов первого порядка в точке истинных параметров θ_0 :

$$g(\tilde{\theta}) \approx g_0 + \mathcal{H}_0(\tilde{\theta} - \theta_0),$$

$$r(\tilde{\theta}) \approx R_0(\tilde{\theta} - \theta_0).$$

При получении второго соотношения мы использовали, что в точке истинных параметров ограничения выполняются: $r(\theta_0) = 0$.

Подставив эти приближения в условия первого порядка, получим следующие асимптотические равенства:

$$g_0 + \mathcal{H}_0(\tilde{\theta} - \theta_0) \stackrel{a}{=} R_0^\top \tilde{\lambda},$$

$$R_0(\tilde{\theta} - \theta_0) \stackrel{a}{=} 0.$$

Перепишем систему в блочной форме:

$$\begin{bmatrix} -\mathcal{H}_0 & \mathbf{R}_0^\top \\ \mathbf{R}_0 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \tilde{\boldsymbol{\lambda}} \end{bmatrix} \stackrel{\text{a}}{=} \begin{bmatrix} \mathbf{g}_0 \\ \mathbf{0} \end{bmatrix}.$$

Отсюда, домножая на \sqrt{N} и $\frac{1}{\sqrt{N}}$, чтобы получились величины порядка $O(1)$, получим асимптотическое равенство:

$$\begin{bmatrix} \sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \frac{1}{\sqrt{N}}\tilde{\boldsymbol{\lambda}} \end{bmatrix} \stackrel{\text{a}}{=} \begin{bmatrix} \mathcal{I}_0^\infty & \mathbf{R}_0^\top \\ \mathbf{R}_0 & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{N}}\mathbf{g}_0 \\ \mathbf{0} \end{bmatrix}.$$

Используем следующее правило блочного обращения матрицы:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1} \\ (\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1}\mathbf{C}\mathbf{A}^{-1} & -(\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1} \end{bmatrix}.$$

В данном случае

$$\mathbf{A} = \mathcal{I}_0^\infty, \quad \mathbf{B} = \mathbf{R}_0^\top, \quad \mathbf{C} = \mathbf{R}_0, \quad \mathbf{D} = \mathbf{0}, \quad (\mathbf{C}\mathbf{A}^{-1}\mathbf{B} - \mathbf{D})^{-1} = (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}.$$

Таким образом,

$$\begin{aligned} & \begin{bmatrix} \mathcal{I}_0^\infty & \mathbf{R}_0^\top \\ \mathbf{R}_0 & \mathbf{0} \end{bmatrix}^{-1} = \\ & = \begin{bmatrix} (\mathcal{I}_0^\infty)^{-1}(\mathbf{I} - \mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}) & (\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1} \\ (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1} & -(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1} \end{bmatrix}. \end{aligned}$$

Получим выражения, асимптотически эквивалентные оценкам $\tilde{\boldsymbol{\theta}}$ и множителям Лагранжа $\tilde{\boldsymbol{\lambda}}$:

$$\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{\text{a}}{=} (\mathcal{I}_0^\infty)^{-1}(\mathbf{I} - \mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}) \left(\frac{1}{\sqrt{N}}\mathbf{g}_0\right), \quad (1)$$

$$\frac{1}{\sqrt{N}}\tilde{\boldsymbol{\lambda}} \stackrel{\text{a}}{=} (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1} \left(\frac{1}{\sqrt{N}}\mathbf{g}_0\right). \quad (2)$$

Вспомним, что $\frac{1}{\sqrt{N}}\mathbf{g}_0 \stackrel{\text{a}}{\sim} \mathbf{N}(\mathbf{0}, \mathcal{I}_0^\infty)$.

Отсюда получим асимптотическое распределение вектора множителей Лагранжа:

$$\frac{1}{\sqrt{N}}\tilde{\boldsymbol{\lambda}} \stackrel{\text{a}}{\sim} \mathbf{N}(\mathbf{0}, (\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathcal{I}_0^\infty(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top(\mathbf{R}_0(\mathcal{I}_0^\infty)^{-1}\mathbf{R}_0^\top)^{-1}),$$

$$\frac{1}{\sqrt{N}} \tilde{\lambda} \stackrel{a}{\sim} N(\mathbf{0}, (\mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top)^{-1}).$$

Статистикой множителя Лагранжа называют следующую величину:

$$\text{LM} \equiv \tilde{\lambda}^\top \tilde{\mathbf{R}} \tilde{\mathcal{I}}^{-1} \tilde{\mathbf{R}}^\top \tilde{\lambda}.$$

Здесь $\tilde{\mathcal{I}}$ — матрица, полученная на основании выборочной информации в точке $\tilde{\theta}$, такая что $1/N \tilde{\mathcal{I}}$ — состоятельная оценка \mathcal{I}_0^∞ . Величина LM имеет распределение χ^2 с p степенями свободы, где p — размерность вектора ограничений \mathbf{r} :

$$\text{LM} \stackrel{a}{\sim} \chi^2(p).$$

Это следует из формулы для распределения $\tilde{\lambda}$, состоятельности оценки $\tilde{\theta}$ и невырожденности матрицы $\mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top$.

Вспомним, что одно из условий первого порядка максимума функции правдоподобия имеет вид $\tilde{\mathbf{g}} = \tilde{\mathbf{R}}^\top \tilde{\lambda}$. Это позволяет выразить статистику множителя Лагранжа через градиент логарифмической функции правдоподобия:

$$\text{LM} = \tilde{\mathbf{g}}^\top \tilde{\mathcal{I}}^{-1} \tilde{\mathbf{g}} \stackrel{a}{\sim} \chi^2(p).$$

Хотя статистика множителя Лагранжа получила свое название благодаря тому, что ее можно выразить через множители Лагранжа, на практике гораздо чаще используют градиентную форму (score form of LM test).

Если вспомнить асимптотическое выражение (2) для $\tilde{\lambda}$, то можно выразить (асимптотически) LM-тест через \mathbf{g}_0 :

$$\text{LM} \stackrel{a}{=} 1/N \mathbf{g}_0^\top (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 (\mathcal{I}_0^\infty)^{-1} \mathbf{g}_0.$$

Статистика отношения правдоподобия по определению есть

$$\text{LR} \equiv 2(\tilde{\ell} - \hat{\ell}).$$

Найдем ее асимптотическое распределение. Используем для этого разложение в ряд Тейлора:

$$\tilde{\ell} = \hat{\ell} + \frac{1}{2} (\tilde{\theta} - \hat{\theta}) \mathcal{H}(\bar{\theta}) (\tilde{\theta} - \hat{\theta}),$$

где $\bar{\theta}_j$ — выпуклая линейная комбинация $\tilde{\theta}_j$ и $\hat{\theta}_j$. Поскольку $\tilde{\theta}$ и $\hat{\theta}$ — состоятельные оценки θ_0 , то $-\mathcal{H}(\bar{\theta}) \stackrel{a}{=} N \mathcal{I}_0^\infty$.

$$LR = 2(\tilde{\ell} - \hat{\ell}) \stackrel{a}{=} N(\tilde{\theta} - \hat{\theta})\mathcal{I}_0^\infty(\tilde{\theta} - \hat{\theta}).$$

Асимптотический эквивалент этой статистики также можно записать в терминах θ_0 , \mathcal{I}_0^∞ , R_0 и g_0 .

Отняв от $\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{=} (\mathcal{I}_0^\infty)^{-1} \frac{1}{\sqrt{N}} g_0$ доказанное ранее равенство (1) получаем, что $\sqrt{N}(\hat{\theta} - \tilde{\theta}) \stackrel{a}{=} (\mathcal{I}_0^\infty)^{-1} R_0^\top (R_0 (\mathcal{I}_0^\infty)^{-1} R_0^\top)^{-1} R_0 (\mathcal{I}_0^\infty)^{-1} g_0$.

Отсюда следует, что статистика отношения правдоподобия асимптотически равна той же самой случайной величине, что и статистика множителя Лагранжа:

$$LR \stackrel{a}{=} 1/N g_0^\top (\mathcal{I}_0^\infty)^{-1} R_0^\top (R_0 (\mathcal{I}_0^\infty)^{-1} R_0^\top)^{-1} R_0 (\mathcal{I}_0^\infty)^{-1} g_0$$

Эта это означает, что статистика отношения правдоподобия также имеет асимптотическое распределение χ^2 с p степенями свободы.

Третья классическая статистика основана на распределении $r(\hat{\theta})$. Поскольку $\hat{\theta}$ — оценка, полученная без учета ограничений, то в общем случае $r(\hat{\theta}) \neq 0$, однако, если верна нулевая гипотеза, то $r(\hat{\theta}) \stackrel{a}{=} 0$. Разложим $\hat{r} = r(\hat{\theta})$ в ряд Тейлора в точке θ_0 , учитывая, что $r(\theta_0) = 0$:

$$\hat{r} \approx r(\theta_0) + R_0(\hat{\theta} - \theta_0) = R_0(\hat{\theta} - \theta_0).$$

Ранее было выведено, что $\sqrt{N}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} N(0, \mathcal{I}_0)$. Отсюда получим **статистику Вальда**:

$$W \equiv \hat{r}^\top (\hat{R} \hat{\mathcal{I}}^{-1} \hat{R}^\top)^{-1} \hat{r}$$

Здесь $\hat{\mathcal{I}}$ — матрица, полученная на основании выборочной информации в точке $\hat{\theta}$, такая что $1/N \hat{\mathcal{I}}$ — состоятельная оценка \mathcal{I}_0^∞ .

Как и в случае двух других тестов $W \stackrel{a}{\sim} \chi^2(p)$.

В пределе $\hat{r} \stackrel{a}{=} R_0(\hat{\theta} - \theta_0) \stackrel{a}{=} 1/N R_0 (\mathcal{I}_0^\infty)^{-1} g_0$. Значит,

$$LR \stackrel{a}{=} 1/N g_0^\top (\mathcal{I}_0^\infty)^{-1} R_0^\top (R_0 (\mathcal{I}_0^\infty)^{-1} R_0^\top)^{-1} R_0 (\mathcal{I}_0^\infty)^{-1} g_0.$$

Тем самым мы показали, что с ростом количества наблюдений все три статистики стремятся к одной и той же случайной переменной, которая имеет распределение $\chi^2(p)$. Другими словами, три классических теста **асимптотически эквивалентны**.

Все три статистики совпадают, если логарифмическая функция правдоподобия является квадратичной. Это верно, в частности, для линейной регрессии с известной дисперсией, например,

$$Y = X\beta + \varepsilon, \text{ где } \varepsilon_i \sim \text{NID}(0,1).$$

Рассмотрим, к примеру, логарифмическую функцию правдоподобия с единственным (скалярным) параметром θ :

$$\ell = -(a - b\theta)^2 + \text{const.}$$

Гессиан $\mathcal{H} = \frac{\partial^2 \ell}{\partial \theta^2} = -2b^2$ является постоянной величиной, информационная матрица, таким образом, равна $\mathcal{I} = 2b^2$ при всех θ .

Возьмем ограничение вида $r(\theta) = 3\theta - 1$

Получим $\hat{\theta} = \frac{a}{b}$, $\tilde{\theta} = \frac{1}{3}$. Откуда $\hat{\ell} = 0 + \text{const}$, $\tilde{\ell} = -(a - \frac{b}{3})^2 + \text{const}$.

$$\text{LR} = 2(\hat{\ell} - \tilde{\ell}) = 2(a - \frac{b}{3})^2.$$

Градиент равен $g = 2b(a - b\theta)$. Таким образом, $\tilde{g} = 2b(a - \frac{1}{3}b)$.

$$\text{LM} = \tilde{g}^T \tilde{\mathcal{I}}^{-1} \tilde{g} = 2b(a - \frac{1}{3}b) \frac{1}{2b^2} 2b(a - \frac{1}{3}b) = 2(a - \frac{b}{3})^2.$$

Найдем ту же статистику через множитель Лагранжа.

$$L = -(a - b\theta)^2 - \lambda(3\theta - 1) \rightarrow \max_{\theta}$$

$$\frac{\partial L}{\partial \theta} = 2(a - b\theta) - 3\lambda = 0 \Rightarrow \tilde{\lambda} = \frac{2}{3}(a - b\tilde{\theta}) = \frac{2}{3}(a - \frac{b}{3}).$$

$$R = \frac{\partial r}{\partial \theta} = 3, \quad \tilde{R} = 3.$$

$$\text{LM} = \tilde{\lambda}^T \tilde{R} \tilde{\mathcal{I}}^{-1} \tilde{R}^T \tilde{\lambda} = \frac{2}{3}(a - \frac{b}{3}) 3 \frac{1}{2b^2} 3 \frac{2}{3}(a - \frac{b}{3}) = 2(a - \frac{b}{3})^2.$$

Теперь найдем статистику Вальда.

$$\hat{r} = 3\hat{\theta} - 1 = 3\frac{a}{b} - 1. \quad \hat{R} = 3.$$

$$W = \hat{r}^T (\hat{R} \hat{\mathcal{I}}^{-1} \hat{R}^T)^{-1} \hat{r} = (3\frac{a}{b} - 1) (3 \frac{1}{2b^2} 3)^{-1} (3\frac{a}{b} - 1) = 2(a - \frac{b}{3})^2.$$

Соотношения между статистиками

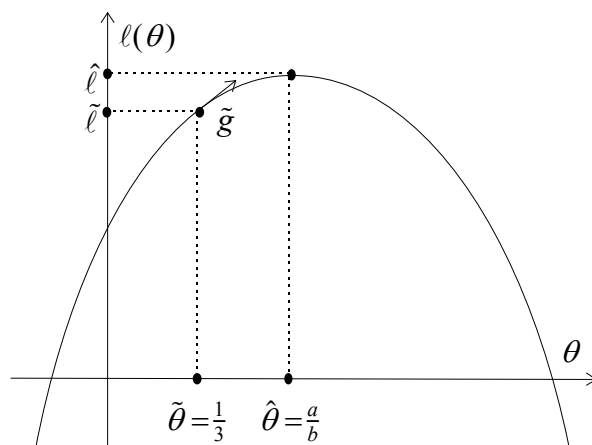


Рис. 8

Все три классические статистики совпадают в случае “бесконечно большой выборки”. В выборках конечных размеров их поведение может существенно отличаться от асимптотического. Поэтому не всегда на эти классические статистики можно полагаться. В этом их отличие от F-статистик, которые имеют *точное* распределение в конечной выборке в случае классической линейной регрессии с нормальными ошибками и линейной проверяемой гипотезой. Рассмотрим этот случай более подробно.

Предположим, что проверяется гипотеза

$$\mathbf{Q}\boldsymbol{\beta} = \mathbf{q},$$

где \mathbf{Q} — известная матрица ($p \times m$), \mathbf{q} — известный вектор ($p \times 1$). В использованных выше обозначениях $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}(\boldsymbol{\beta}, \sigma^2) = \mathbf{Q}\boldsymbol{\beta} - \mathbf{q}$, матрица $\mathbf{R}(\boldsymbol{\theta})$ равна $[\mathbf{Q} \ \mathbf{0}]$ при всех значениях $\boldsymbol{\theta}$ (нулевой вектор относится к параметру σ^2). Проверяемой гипотезе соответствует следующая статистика Вальда (вспомним, что $\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$):

$$\begin{aligned} W &= \hat{\mathbf{r}}^\top (\hat{\mathbf{R}} \hat{\mathcal{I}}^{-1} \hat{\mathbf{R}}^\top)^{-1} \hat{\mathbf{r}} = (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q})^\top (\mathbf{Q} \hat{\mathcal{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \mathbf{Q}^\top)^{-1} (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q}) = \\ &= \frac{1}{\hat{\sigma}^2} (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q})^\top (\mathbf{Q} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{Q}^\top)^{-1} (\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{q}). \end{aligned}$$

Нам нужно максимизировать функцию правдоподобия при ограничении $\mathbf{Q}\boldsymbol{\beta} = \mathbf{q}$. Лагранжиан рассматриваемой задачи условной максимизации имеет вид

$$L = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{Q}\boldsymbol{\beta} - \mathbf{q})^\top \boldsymbol{\lambda}.$$

В максимуме должно выполняться

$$\frac{\partial L}{\partial \beta^T}(\tilde{\theta}, \tilde{\lambda}) = \frac{1}{\tilde{\sigma}^2} X^T(Y - X\tilde{\beta}) - Q^T \tilde{\lambda} = 0.$$

Отсюда

$$\tilde{\beta} = (X^T X)^{-1} X^T Y - \tilde{\sigma}^2 (X^T X)^{-1} Q^T \tilde{\lambda} = \hat{\beta} - \tilde{\sigma}^2 (X^T X)^{-1} Q^T \tilde{\lambda},$$

где $\hat{\beta} = (X^T X)^{-1} X^T Y$ — оценки ОМНК (без ограничений). Домножая это равенство слева на Q и учитывая, что $Q\tilde{\beta} = q$, получим

$$\tilde{\sigma}^2 \tilde{\lambda} = (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q).$$

Таким образом, оценки с учетом ограничений равны

$$\tilde{\beta} = \hat{\beta} - (X^T X)^{-1} Q^T (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q).$$

Из условия $\frac{\partial L}{\partial \beta^T}(\tilde{\theta}, \tilde{\lambda}) = 0$ несложно получить, что $\tilde{\sigma}^2 = \frac{R\tilde{S}S}{N}$, где $R\tilde{S}S$ —

сумма квадратов остатков в регрессии с ограничениями (так же как $\hat{\sigma}^2 = \frac{R\hat{S}S}{N}$ в регрессии без ограничений).

Статистика множителя Лагранжа равна:

$$\begin{aligned} LM &= \tilde{\lambda}^T \tilde{R} \tilde{I}^{-1} \tilde{R}^T \tilde{\lambda} = \tilde{\sigma}^2 \tilde{\lambda}^T Q(X^T X)^{-1} Q^T \tilde{\lambda} = \\ &= \frac{1}{\tilde{\sigma}^2} (Q\hat{\beta} - q)^T (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q). \end{aligned}$$

Можно также показать (пропускаем эти преобразования), что

$$(Q\hat{\beta} - \hat{\beta})^T (Q(X^T X)^{-1} Q^T)^{-1} (Q\hat{\beta} - q) = R\tilde{S}S - R\hat{S}S.$$

Это позволяет выразить LM и W через суммы квадратов остатков:

$$LM = N \frac{R\tilde{S}S - R\hat{S}S}{R\tilde{S}S}, \quad W = N \frac{R\tilde{S}S - R\hat{S}S}{R\hat{S}S},$$

Логарифмическая функция правдоподобия в максимуме равна

$$\tilde{\ell} = -\frac{N}{2} \ln(2\pi \frac{R\tilde{S}S}{N}) - \frac{N}{2}.$$

В регрессии без ограничений $\hat{\ell} = -\frac{N}{2} \ln(2\pi \frac{R\hat{S}S}{N}) - \frac{N}{2}$.

Отсюда найдем статистику отношения правдоподобия:

$$LR = 2(\tilde{\ell} - \hat{\ell}) = N (\ln(R\tilde{S}S) - \ln(R\hat{S}S)).$$

Так как логарифм — строго вогнутая функция, то выполнено следующее точное неравенство:

$$W > LR > LM.$$

Таким образом, тест Вальда будет чаще отвергать гипотезу, тест множителя Лагранжа — реже. Это же неравенство верно и для нелинейных регрессий.

F-статистика для проверки той же гипотезы равна

$$F = \frac{R\tilde{S}S - R\hat{S}S}{R\hat{S}S} \frac{N - m}{p}.$$

Она распределена как $F(p, N - m)$.

В линейной регрессии лучше, конечно использовать t- и F-статистики. Кроме того, распределение этих статистик лучше аппроксимируется их номинальным распределением и в других моделях: нелинейных регрессиях, некоторых моделях, являющихся развитием регрессионных, некоторых искусственных регрессиях и т. п. Хотя здесь t- и F-статистики не будут иметь точного распределения, но, как показали эксперименты, они, как правило, лучше в конечных выборках, чем их асимптотические аналоги (N и χ^2 соответственно). Такие t- и F-статистики называют **асимптотическими t- и F-статистиками**. Три классические статистики можно преобразовать в асимптотические F-статистики по следующим формулам:

$$F_W = \frac{W}{N} \frac{N - m}{p}, \quad F_{LM} = \frac{LM}{N - LM} \frac{N - m}{p}, \quad F_{LR} = \left(\exp\left(\frac{LR}{N}\right) - 1 \right) \frac{N - m}{p}.$$

Все эти статистики распределены приближенно как $F(p, N - m)$.

Понятно, что тесты на основе W , LR , LM и асимптотические F-тесты дают противоречивые результаты в конечных выборках. Одни из них могут отвергать гипотезу при выбранном уровне значимости, другие же говорить в пользу принятия гипотезы.

Для того, чтобы исследовать поведение асимптотических статистик в конечных выборках, используют метод Монте-Карло. С помощью этого метода можно, в частности, выяснить, какой из тестов более подходит для данного типа моделей, какую оценку ковариационной матрицы оценок лучше всего использовать.

Модели с дискретной зависимой переменной

Модели с бинарной зависимой переменной (логит и пробит)

Бинарная зависимая переменная Y_i называется так, потому, что принимает два значения, обычно 0 и 1. Обозначим через P_i вероятность появления единицы, или, что в данном случае то же самое, математическое ожидание Y_i :

$$P_i = \text{Prob}(Y_i = 1) = E(Y_i).$$

Вероятность P_i в линейной модели с бинарной зависимой переменной зависит от $X_i \beta$, где X_i — строка матрицы регрессоров, β — вектор коэффициентов регрессии:

$$P_i = F(X_i \beta).$$

Здесь $F(\cdot)$ — (кумулятивная) функция распределения некоторого непрерывного распределения.

В **логите** используется (стандартное) логистическое распределение с функцией распределения

$$F(z) = \frac{1}{1 + e^{-z}}$$

и плотностью распределения

$$f(z) = \frac{e^z}{(1 + e^z)^2}.$$

В **пробите** используется стандартное нормальное распределение с функцией распределения

$$F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Логарифмическая функция правдоподобия равна

$$\begin{aligned} \ell &= \sum_{i \in I_0} \ln P_i(\beta) + \sum_{i \in I_1} \ln(1 - P_i(\beta)) = \\ &= \sum_{i=1}^N [Y_i \ln P_i(\beta) + (1 - Y_i) \ln(1 - P_i(\beta))]. \end{aligned}$$

где I_0 и I_1 — множества наблюдений, для которых $Y_i = 0$ и $Y_i = 1$ соответственно.

Градиент функции правдоподобия равен:

$$\mathbf{g}^\top = \frac{\partial \ell}{\partial \beta} = \sum_i \left[\frac{Y_i}{P_i} - \frac{1 - Y_i}{1 - P_i} \right] \frac{\partial P_i(\beta)}{\partial \beta} = \sum_i \frac{Y_i - P_i}{P_i(1 - P_i)} f(X_i \beta) X_i.$$

Логит

Для логита верно, что $f(z) = F(z)(1-F(z))$, поэтому $f(\mathbf{X}_i\boldsymbol{\beta}) = P_i(1-P_i)$. Это позволяет упростить формулу градиента:

$$\mathbf{g}^\top = \sum_i (Y_i - P_i)\mathbf{X}_i$$

где $P_i = \frac{1}{1 + e^{-X_i\boldsymbol{\beta}}}$.

Гессиан в случае логита равен:

$$\begin{aligned}\mathcal{H} &= \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = - \sum_i \frac{\partial P_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \mathbf{X}_i^\top = - \sum_i f(\mathbf{X}_i\boldsymbol{\beta}) \mathbf{X}_i^\top \mathbf{X}_i = \\ &= - \sum_i P_i(1-P_i) \mathbf{X}_i^\top \mathbf{X}_i.\end{aligned}$$

Видно, что гессиан всюду отрицательно определенный (кроме вырожденных случаев). Таким образом, логарифмическая функция правдоподобия всюду вогнута.

Гессиан не зависит от случайного вектора \mathbf{Y} , поэтому ожидаемый гессиан равен просто гессиану, то есть информационная матрица равна минус гессиану:

$$\mathcal{I} = -\mathcal{H} = \sum_i P_i(1-P_i) \mathbf{X}_i^\top \mathbf{X}_i.$$

Для поиска максимума можно использовать метод Ньютона (он же в данном случае и method of scoring):

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\mathcal{H}(\boldsymbol{\beta}^t))^{-1} \mathbf{g}^t(\boldsymbol{\beta}^t) = \boldsymbol{\beta}^t - \Delta \boldsymbol{\beta}^t.$$

Поскольку максимизируемая функция вогнута, то метод Ньютона всегда сходится. Шаг алгоритма удобно находить как оценки коэффициентов во вспомогательной регрессии \mathbf{Y}^* по \mathbf{X}^* , где

$$\mathbf{Y}_i^* = \frac{Y_i - P_i}{\sqrt{P_i(1-P_i)}}, \quad \mathbf{X}_i^* = \sqrt{P_i(1-P_i)} \mathbf{X}_i.$$

Ίδιαιότητα

В случае пробита выражение для гессиана несколько более громоздкое:

$$\begin{aligned}\mathcal{H} &= \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \sum_i \left[\frac{Y_i}{P_i} - \frac{1-Y_i}{1-P_i} \right] f(\mathbf{X}_i\boldsymbol{\beta}) \mathbf{X}_i = \\ &= - \sum_i \left[\frac{Y_i}{(P_i)^2} + \frac{1-Y_i}{(1-P_i)^2} \right] f(\mathbf{X}_i\boldsymbol{\beta})^2 \mathbf{X}_i^\top \mathbf{X}_i \\ &\quad + \sum_i \left[\frac{Y_i}{P_i} - \frac{1-Y_i}{1-P_i} \right] f(\mathbf{X}_i\boldsymbol{\beta}) \frac{\partial f(\mathbf{X}_i\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \mathbf{X}_i = \\ &= - \sum_i \left[\left(\frac{Y_i - P_i}{P_i(1-P_i)} \right)^2 f(\mathbf{X}_i\boldsymbol{\beta}) \mathbf{X}_i^\top - \frac{Y_i - P_i}{P_i(1-P_i)} \frac{\partial f(\mathbf{X}_i\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \right] f(\mathbf{X}_i\boldsymbol{\beta}) \mathbf{X}_i.\end{aligned}$$

Для нормального распределения верно, что $\frac{\partial f(z)}{\partial z} = -zf(z)$. Это позволяет несколько упростить выражение для гессiana, так как

$$\frac{\partial f(\mathbf{X}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = -\mathbf{X}_i \boldsymbol{\beta} f(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i^\top.$$

Обозначим

$$v_i = f(\mathbf{X}_i \boldsymbol{\beta}) \frac{Y_i - P_i}{P_i(1 - P_i)}.$$

Тогда

$$\mathcal{H} = -\sum_i v_i (v_i + \mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i^\top \mathbf{X}_i.$$

В тех же обозначениях градиент равен

$$\mathbf{g}^\top = \sum_i v_i \mathbf{X}_i.$$

Как и в случае логита, можно показать, что гессиан является отрицательно определенным.

Чтобы найти информационную матрицу для пробита, воспользуемся тем, что $E(Y_i) = P_i$, $E(Y_i - P_i)^2 = P_i(1 - P_i)$.

$$\mathcal{I} = -E(\mathcal{H}) = -\sum_i E(v_i^2) \mathbf{X}_i^\top \mathbf{X}_i = -\sum_i \frac{f^2(\mathbf{X}_i \boldsymbol{\beta})}{P_i(1 - P_i)} \mathbf{X}_i^\top \mathbf{X}_i.$$

Для поиска максимума, как и в случае логита, можно использовать градиентный алгоритм:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\mathcal{I}^t)^{-1} \mathbf{g}^t = \boldsymbol{\beta}^t - \Delta \boldsymbol{\beta}^t.$$

В методе Ньютона с $\mathcal{I}^t = -\mathcal{H}(\boldsymbol{\beta}^t)$ используется вспомогательная регрессия с переменными

$$Y_i^* = \frac{\sqrt{v_i}}{\sqrt{v_i + \mathbf{X}_i \boldsymbol{\beta}}}, \quad X_i^* = \sqrt{v_i(v_i + \mathbf{X}_i \boldsymbol{\beta})} \mathbf{X}_i.$$

Если использовать информационную матрицу в точке оценок $\mathcal{I}^t = \mathcal{I}(\boldsymbol{\beta}^t)$ (method of scoring), то надо взять

$$Y_i^* = \frac{Y_i - P_i}{\sqrt{P_i(1 - P_i)}}, \quad X_i^* = \frac{f(\mathbf{X}_i \boldsymbol{\beta})}{\sqrt{P_i(1 - P_i)}} \mathbf{X}_i.$$

Вспомогательные регрессии для пробита и логита являются искусственными регрессиями, то есть, с помощью них можно проверять все те гипотезы, которые можно проверять в случае обычной регрессии, в частности, использовать t-статистики.

Метод максимального правдоподобия для моделей с дискретной зависимой переменной по сути является нелинейным методом наименьших квадратов (НМНК). Математическое ожидание Y_i равно P_i . Разность Y_i и P_i должна иметь нулевое математическое ожидание, то есть подходит в качестве ошибки в нелинейной регрессии Y_i по P_i . Однако эта ошибка будет гетероскедастична. Действительно,

$$V(Y_i) = E(Y_i - P_i)^2 = P_i(1 - P_i)^2 + (1 - P_i)P_i^2 = P_i(1 - P_i).$$

Таким образом, следует воспользоваться взвешенным НМНК, где веса рассматриваются как фиксированные:

$$\sum_{i=1}^N \frac{(Y_i - P_i)^2}{P_i(1 - P_i)} \rightarrow \min.$$

Поскольку веса неизвестны, то приходится использовать итерационные процедуры, которые совпадают с описанными выше. Оба метода дают одни и те же оценки, поскольку достигают экстремума одновременно.

Пуассонова регрессия

Распределение Пуассона — дискретное распределение, задаваемая формулой

$$\text{Prob}(Y=r) = \frac{\mu^r}{r!} e^{-\mu},$$

где μ — параметр распределения.

Распределение Пуассона имеет случайная величина Y , равная количеству событий, произошедших за некоторый промежуток времени, если эти события независимы и происходят с постоянной скоростью (равномерно по времени). Это, например, может быть количество покупателей, посетивших магазин в течении часа.

Моменты распределения:

$$E(Y) = \mu, \quad \text{Var}(Y) = \mu.$$

В регрессионной модели с распределением Пуассона параметр μ зависит от набора факторов и неизвестных параметров.

В линейной модели:

$$\mu_i = \exp(X_i \beta).$$

Тогда логарифмическая функция правдоподобия равна

$$\ell = \sum_i [Y_i X_i \beta - \exp(X_i \beta) - \ln Y_i!] \rightarrow \max_{\beta}.$$

Градиент равен:

$$\mathbf{g}^\top = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_i [Y_i \mathbf{X}_i - \exp(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i].$$

Условие первого порядка максимума:

$$\sum_i [Y_i - \exp(\mathbf{X}_i \boldsymbol{\beta})] \mathbf{X}_i.$$

Гессиан не содержит случайных компонент, и поэтому информационная матрица равна минус гессиану.

$$\mathcal{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = - \sum_i \exp(\mathbf{X}_i \boldsymbol{\beta}) \mathbf{X}_i \mathbf{X}_i^\top.$$

Для поиска максимума можно использовать метод Ньютона:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\mathcal{H}^t)^{-1} \mathbf{g}^t.$$

Метод Ньютона легко реализовать с помощью вспомогательной регрессии.

Обозначим

$$v_i = \exp\left(\frac{1}{2} \mathbf{X}_i \boldsymbol{\beta}\right), \quad Y_i^* = Y_i / v_i - v_i, \quad X_i^* = X_i v_i.$$

Тогда если $\Delta \boldsymbol{\beta}$ — оценки коэффициентов в регрессии Y^* по X^* , то шаг метода Ньютона задается формулой:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \Delta \boldsymbol{\beta}^t.$$

Оценка ковариационной матрицы оценок есть $-(\mathcal{H})^{-1} = (\mathbf{X}^{*\top} \mathbf{X}^*)^{-1}$, поэтому тесты для коэффициентов матрицы и т. п. можно получить из регрессии Y^{**} по X^* , где $Y^{**} = Y^*/s$, $s = \sqrt{Y^{*\top} Y^*/N}$, и Y^* берется в точке оценок МП (на последней итерации метода Ньютона). Проверять ограничения на коэффициенты можно как с помощью χ^2 статистики Вальда, так и с помощью соответствующих t- и F-статистик из вспомогательной регрессии. В качестве аналога стандартного F-теста на равенство нулю коэффициентов при всех переменных кроме константы можно использовать статистику отношения правдоподобия. Пусть $\tilde{\ell}$ — значение логарифмической функции правдоподобия когда $\mu_i = \mu \forall i$.

$$\text{LR} = 2 (\hat{\ell} - \tilde{\ell}) \stackrel{a}{\sim} \chi^2(m-1),$$

где m — количество регрессоров (столбцов \mathbf{X}).

Найдем $\tilde{\ell}$:

$$\ell = \sum_i [Y_i \ln \mu - \mu - \ln Y_i!].$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\mu} \sum_i Y_i - N = 0 \quad \Rightarrow \quad \tilde{\mu} = \frac{1}{N} \sum_i Y_i = \bar{Y}.$$

Таким образом, $\tilde{\ell} = N \bar{Y} \ln \bar{Y} - N \bar{Y} - \sum_i \ln Y_i!$.

Обобщенный метод наименьших квадратов

Отбросим предположение, что в регрессионной модели ошибки независимы. Пусть ошибки коррелированы и структура матрицы ковариаций ошибок V известна. Найдем оценки в этой регрессии методом МП при нормально распределенных ошибках. Заметим, что метод можно использовать и для более широкого класса распределений.

Модель имеет вид

$$Y = X\beta + \varepsilon, \text{ где } \varepsilon \sim N(\mathbf{0}, V).$$

Плотность распределения ε равна

$$p_{\varepsilon}(z) = (2\pi)^{-N/2} |V|^{-1/2} \exp\left(-\frac{1}{2} z^T V z\right).$$

Отсюда получим логарифмическую функцию правдоподобия

$$\ell = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln|V| - \frac{1}{2} (Y - X\beta)^T V^{-1} (Y - X\beta).$$

Далее мы рассмотрим случай, когда V известна с точностью до множителя:

$$V = \sigma^2 W.$$

Подставим это выражение в функцию правдоподобия:

$$\ell = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \ln|W| - \frac{1}{2\sigma^2} (Y - X\beta)^T W^{-1} (Y - X\beta).$$

Воспользовавшись условием первого порядка

$$\frac{\partial \ell}{\partial \sigma^2} = 0,$$

выразим σ^2 через β : $\sigma^2(\beta) = \text{RSS}(\beta)/N$ где

$$\text{RSS}(\beta) = (Y - X\beta)^T W^{-1} (Y - X\beta)$$

— обобщенная сумма квадратов остатков.

Подставив $\sigma^2(\beta)$ в логарифмическую функцию правдоподобия, получим концентрированную функцию правдоподобия:

$$\ell^c = -\frac{N}{2} \ln\left(2\pi \frac{\text{RSS}(\beta)}{N}\right) - \frac{1}{2} \ln|W| - \frac{N}{2}.$$

Поскольку W — константа, то максимизация ℓ^c эквивалентна минимизации обобщенной суммы квадратов: $\text{RSS}(\beta) \rightarrow \min_{\beta}$.

$$\frac{\partial \ell^c}{\partial \beta} = 0 \Leftrightarrow Y^T W^{-1} X = \hat{\beta}^T X^T W^{-1} X.$$

Получим оценку МП для β :

$$\hat{\beta} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y.$$

На практике удобно использовать не эту формулу, а преобразовать X и Y так, чтобы можно было делать расчеты с помощью ОМНК. Поскольку W — симметричная положительно определенная матрица, то к ней можно применить разложение Холецкого (или любое другое аналогичное представление):

$$W = T T^T,$$

где T — нижняя или верхняя треугольная матрица. Отсюда

$$W^{-1} = (T^{-1})^T T^{-1},$$

$$\hat{\beta} = (X^T (T^{-1})^T T^{-1} X)^{-1} X^T (T^{-1})^T T^{-1} Y.$$

Обозначим $T^{-1} X = X^*$, $T^{-1} Y = Y^*$. Тогда выражение для $\hat{\beta}$ примет вид:

$$\hat{\beta} = (X^{*T} X^*)^{-1} X^{*T} Y^*.$$

Таким образом, оценки коэффициентов β можно найти, применив обычный метод наименьших квадратов к регрессии Y^* по X^* . Вообще говоря, предположения о нормальности не требуется для состоятельности оценок, и на метод максимального правдоподобия можно было не ссылаться, поскольку предложенное преобразование сразу приводит к классической регрессионной модели.

Несложно проверить, что информационная матрица равна

$$\mathcal{I} = \frac{1}{\sigma^2} X^T W^{-1} X,$$

поэтому и ковариационная матрица оценок $\hat{\beta}$ полученная из той же регрессии будет оценкой ковариационной матрицы оценок максимального правдоподобия:

$$V(\hat{\beta}) = \hat{\sigma}^2 (X^T W^{-1} X)^{-1} = \hat{\sigma}^2 (X^{*T} X^*)^{-1},$$

где $\hat{\sigma}^2$ — оценка дисперсии в регрессии Y^* по X^* , которая является оценкой параметра σ^2 в исходной модели. Это позволяет использовать t- и F-

статистики в преобразованной модели для проверки гипотез о коэффициентах β .

В более общем случае матрица ковариаций ошибок V зависит от вектора неизвестных параметров α . Эту модель в дальнейшем будем называть моделью обобщенного метода наименьших квадратов, хотя ковариационная матрица ошибок в обобщенном методе наименьших квадратов в собственном смысле слова зависит от единственного неизвестного множителя. Предполагается, что два вектора параметров не связаны между собой, т.е. математическое ожидание Y не зависит от α , а матрица ковариаций V не зависит от β .

$$\ell = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln |V(\alpha)| - \frac{1}{2} (Y - X\beta)^\top V(\alpha)^{-1} (Y - X\beta).$$

Информационная матрица будет блочно-диагональной: ее часть, соответствующая “взаимодействию” α и β будет равна нулю:

$$\frac{\partial \ell}{\partial \beta} = (Y - X\beta)^\top V(\alpha)^{-1} X.$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \alpha^\top} = (Y - X\beta)^\top \frac{\partial V(\alpha)^{-1}}{\partial \alpha^\top} X.$$

$$\mathcal{I}_{\beta\alpha} = E\left(\frac{\partial^2 \ell}{\partial \beta \partial \alpha^\top}(\beta_0, \alpha_0)\right) = E(\varepsilon^\top \frac{\partial V(\alpha)^{-1}}{\partial \alpha^\top} X) = E(\varepsilon^\top) \frac{\partial V(\alpha)^{-1}}{\partial \alpha^\top} X = \mathbf{O}.$$

Отсюда следует, что для проведения тестов относительно α можно использовать просто диагональный блок информационной матрицы, относящийся к α , не учитывая, что $\hat{\beta}$ — оценки, и наоборот, для проведения тестов относительно β можно использовать просто диагональный блок информационной матрицы, относящийся к β :

$$\mathcal{I}_{\beta\beta} = E\left(\frac{\partial \ell}{\partial \beta} \frac{\partial \ell}{\partial \beta}^\top\right) = X^\top V(\alpha)^{-1} X.$$

Если имеется способ получить состоятельную оценку параметров α ($\check{\alpha}$); то эффективную оценку параметров β благодаря блочно-диагональности информационной матрицы можно получить за один шаг:

$$\check{\beta} = (X^\top V(\check{\alpha})^{-1} X)^{-1} X^\top V(\check{\alpha})^{-1} Y.$$

Такую оценку принято называть **одношаговой эффективной оценкой**, а метод называют **возможный обобщенный метод наименьших квадратов** (feasible generalized least squares).

Если на основании оценок β можно из условий первого порядка максимума правдоподобия вычислить оценки α , то можно использовать **итеративный обобщенный метод наименьших квадратов** (iterated generalized least squares). Этот метод сходится к оценкам максимального правдоподобия.

Регрессии с неодинаковой дисперсией и тестирование гетероскедастичности

Взвешенный метод наименьших квадратов

Обобщенный метод наименьших квадратов имеет много применений. Его частным случаем является взвешенный метод наименьших квадратов, позволяющий оценивать регрессии с гетероскедастичной ошибкой. Гетероскедастичность означает, что хотя матрица ковариаций ошибок диагональная, но дисперсии (стоящие по диагонали) разные.

Пусть ошибки независимы и i -я ошибка имеет дисперсию $\sigma_i^2 = \sigma^2 w_i$. В данном случае матрица W — диагональная с типичным диагональным элементом w_i . Матрица T — тоже диагональная с типичным элементом $\sqrt{w_i}$, а T^{-1} — диагональная с типичным элементом $\frac{1}{\sqrt{w_i}}$. Переменные во вспомогательной регрессии будут иметь вид:

$$Y_i^* = \frac{Y_i}{\sqrt{w_i}}, \quad X_i^* = \frac{X_i}{\sqrt{w_i}}.$$

Такую регрессию называют **взвешенной регрессией**.

Если веса зависят от неизвестных параметров $w_i = w_i(\gamma)$, то следует воспользоваться методом максимального правдоподобия. Логарифмическая функция правдоподобия равна

$$\ell = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \sum_i \ln w_i(\gamma) - \frac{1}{2\sigma^2} \sum_i \frac{1}{w_i(\gamma)} (Y_i - X_i \beta)^2.$$

Концентрируем функцию правдоподобия по σ^2 :

$$\ell^c = -\frac{1}{2} \sum_i \ln \left(\frac{1}{w_i(\gamma)} (Y_i - X_i \beta)^2 \right) - \frac{1}{2} \sum_i \ln w_i(\gamma) + \text{const}.$$

Максимизация функции правдоподобия эквивалентна минимизации суммы квадратов остатков взвешенной регрессии по β и γ , если взять нормированные веса:

$$Y_i^n = \frac{Y_i}{\sqrt{w_i^n(\gamma)}}, \quad X_i^n = \frac{X_i}{\sqrt{w_i^n(\gamma)}}.$$

Здесь $w_i^n(\boldsymbol{\gamma}) = \frac{w_i(\boldsymbol{\gamma})}{\bar{w}(\boldsymbol{\gamma})}$, $\bar{w}(\boldsymbol{\gamma})$ — среднее геометрическое весов ($\bar{w} = \prod_i w_i^{1/N}$).

Важно, что используются *нормированные* веса, в противном случае минимизация суммы квадратов привела бы к неправильному результату.

Такой метод малоприменим для вычислений. Ниже рассматривается более удобный метод, который годится в частном случае линейной мультипликативной гетероскедастичности.

Проверка гипотезы о наличии гетероскедастичности известного вида

Выдвинем явную гипотезу о виде гетероскедастичности в регрессии:

$$w_i(\boldsymbol{\gamma}) = h(\mathbf{Z}_i \boldsymbol{\gamma}),$$

где $h(\cdot)$ — дифференцируемая строго монотонная функция, такая что $h(0) = 1$, $\mathbf{Z}_i \boldsymbol{\gamma}$ — линейная комбинация известных переменных \mathbf{Z} с неизвестными коэффициентами $\boldsymbol{\gamma}$.

Дисперсия ошибки i -го наблюдения равна $\sigma_i^2 = \sigma^2 h(\mathbf{Z}_i \boldsymbol{\gamma})$. Функция правдоподобия i -го наблюдения будет иметь вид:

$$\ell_i = -\frac{1}{2} \ln(2\pi\sigma^2 h(\mathbf{Z}_i \boldsymbol{\gamma})) - \frac{1}{2\sigma^2 h(\mathbf{Z}_i \boldsymbol{\gamma})} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2.$$

Как мы уже видели, информационная матрица в модели обобщенного МНК имеет блочно-диагональную форму, поэтому гипотезы о $\boldsymbol{\gamma}$ можно проверять независимо от $\boldsymbol{\beta}$. Поэтому в дальнейшем будем рассматривать градиент функции правдоподобия и информационную матрицу только в той части, которая относится к $\boldsymbol{\gamma}$ и σ^2 , которые вместе составляют вектор $\boldsymbol{\alpha} = (\sigma^2, \boldsymbol{\gamma})^T$.

Для проверки гипотезы об отсутствии гетероскедастичности удобнее всего использовать LM-тест (нулевая гипотеза $H_0: \boldsymbol{\gamma} = \mathbf{0}$), поскольку для него не требуется оценивать модель при $\boldsymbol{\gamma} \neq \mathbf{0}$. Достаточно оценить регрессию обычным методом наименьших квадратов.

Найдем вклад в градиент i -го наблюдения:

$$\begin{aligned} \frac{\partial \ell_i}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \frac{\varepsilon_i^2}{h(\mathbf{Z}_i \boldsymbol{\gamma})}, \\ \frac{\partial \ell_i}{\partial \sigma^2} \Big|_{H_0} &= \frac{1}{2\sigma^2} \left(\frac{\varepsilon_i^2}{\sigma^2} - 1 \right) = \frac{1}{2\sigma^2} \mu_i, \\ \frac{\partial \ell_i}{\partial \boldsymbol{\gamma}} &= -\frac{1}{2} \frac{h'(\mathbf{Z}_i \boldsymbol{\gamma})}{h(\mathbf{Z}_i \boldsymbol{\gamma})} \mathbf{Z}_i + \frac{1}{2\sigma^2} \frac{h'(\mathbf{Z}_i \boldsymbol{\gamma})}{h^2(\mathbf{Z}_i \boldsymbol{\gamma})} \varepsilon_i^2 \mathbf{Z}_i. \end{aligned}$$

$$\left. \frac{\partial \ell_i}{\partial \gamma} \right|_{H_0} = \frac{h'(0)}{2} \left(\frac{\varepsilon_i^2}{\sigma^2} - 1 \right) \mathbf{Z}_i = \frac{h'(0)}{2} \mu_i \mathbf{Z}_i.$$

Здесь мы обозначили $\mu_i = \frac{\varepsilon_i^2}{\sigma^2} - 1$ и воспользовались тем, что $h(0) = 1$. Информационную матрицу удобно находить через матрицу вкладов в градиент. Воспользуемся тем, что $E(\mu_i^2) = 2$, поскольку для нормального распределения

$$E\left(\frac{\varepsilon_i^2}{\sigma^2}\right) = 1 \quad \text{и} \quad E\left(\frac{\varepsilon_i^4}{\sigma^4}\right) = 3.$$

Отсюда получим при выполнении нулевой гипотезы

$$\begin{aligned} E\left(\left(\frac{\partial \ell_i}{\partial \sigma^2}\right)^2\right) &= \frac{1}{4\sigma_i^4} E(\mu_i^2) = \frac{1}{2\sigma_i^4}, \\ E\left(\frac{\partial \ell_i}{\partial \sigma^2} \frac{\partial \ell_i}{\partial \gamma}\right) &= \frac{h'(0)}{4\sigma^2} E(\mu_i^2) \mathbf{Z}_i = \frac{h'(0)}{2\sigma^2} \mathbf{Z}_i, \\ E\left(\frac{\partial \ell_i}{\partial \gamma} \frac{\partial \ell_i}{\partial \gamma}^\top\right) &= \frac{(h'(0))^2}{4} E(\mu_i^2) \mathbf{Z}_i \mathbf{Z}_i^\top = \frac{(h'(0))^2}{2} \mathbf{Z}_i \mathbf{Z}_i^\top. \end{aligned}$$

Таким образом, информационная матрица равна

$$\begin{aligned} \mathcal{I}_{\alpha\alpha}^N &= E(\mathbf{G}_{\alpha 0}^\top \mathbf{G}_{\alpha 0}) = \begin{bmatrix} \frac{N}{2\sigma_i^4} & \frac{h'(0)}{2\sigma^2} \sum_i \mathbf{Z}_i \\ \frac{h'(0)}{2\sigma^2} \sum_i \mathbf{Z}_i^\top & \frac{(h'(0))^2}{2} \sum_i \mathbf{Z}_i \mathbf{Z}_i^\top \end{bmatrix} = \\ &= \begin{bmatrix} \frac{1}{2\sigma_i^4} \mathbf{1}^\top \mathbf{1} & \frac{h'(0)}{2\sigma^2} \mathbf{1}^\top \mathbf{Z} \\ \frac{h'(0)}{2\sigma^2} \mathbf{Z}^\top \mathbf{1} & \frac{(h'(0))^2}{2} \mathbf{Z}^\top \mathbf{Z} \end{bmatrix}. \end{aligned}$$

где $\mathbf{1}$ — вектор-столбец, составленный из N единиц. Если обозначить

$$\mathbf{Z}^* = \left(\frac{1}{\sqrt{2}\sigma^2} \mathbf{1}, \frac{h'(0)}{\sqrt{2}} \mathbf{Z} \right),$$

то

$$\mathcal{I}_{\alpha\alpha}^N = \mathbf{Z}^{*\top} \mathbf{Z}^*.$$

Статистика множителя Лагранжа для проверяемой гипотезы равна

$$LM = \tilde{\mathbf{g}}_\alpha^\top (\tilde{\mathcal{I}}_{\alpha\alpha}^N)^{-1} \tilde{\mathbf{g}}_\alpha,$$

где градиент и информационная матрица берутся в точке $(\tilde{\beta}, \tilde{\sigma}^2; 0)$ оценок ОМНК.

Градиент равен $\tilde{\mathbf{g}}_\alpha = \left(\frac{1}{2\tilde{\sigma}^2} \mathbf{1}^\top \tilde{\boldsymbol{\mu}}, \frac{h'(0)}{2} \mathbf{Z}^\top \tilde{\boldsymbol{\mu}} \right)$, где $\tilde{\mu}_i = \frac{e_i^2}{\tilde{\sigma}^2} - 1$, e_i — остатки из регрессии. (Оценка дисперсии $\tilde{\sigma}^2$, полученная методом максимального правдоподобия такова, что $\mathbf{1}^\top \tilde{\boldsymbol{\mu}} = 0$, ò. ê. производная функции правдоподобия равна нулю.) В терминах матрицы \mathbf{Z}^*

$$\tilde{\mathbf{g}}_\alpha = \frac{1}{\sqrt{2}} \mathbf{Z}^{*\top} \tilde{\boldsymbol{\mu}}.$$

В таком случае можно заметить, что LM-статистика равна *объясненной сумме квадратов* из регрессии $\frac{1}{\sqrt{2}} \tilde{\boldsymbol{\mu}} \hat{\mathbf{I}} \mathbf{Z}^*$ или, что то же самое, *половине объясненной суммы квадратов* из регрессии $\tilde{\boldsymbol{\mu}} \hat{\mathbf{I}} \mathbf{Z}^*$:

$$\text{LM} = \frac{1}{\sqrt{2}} \tilde{\boldsymbol{\mu}}^\top \mathbf{Z}^* (\mathbf{Z}^{*\top} \mathbf{Z}^*)^{-1} \frac{1}{\sqrt{2}} \mathbf{Z}^{*\top} \tilde{\boldsymbol{\mu}} = \frac{1}{2} \tilde{\boldsymbol{\mu}}^\top \mathbf{Z}^* (\mathbf{Z}^{*\top} \mathbf{Z}^*)^{-1} \mathbf{Z}^{*\top} \tilde{\boldsymbol{\mu}}.$$

Если домножить регрессоры на отличные от нуля константы, то подпространство, которое на них натянуто, не изменится. Поэтому регрессия $\tilde{\boldsymbol{\mu}} \hat{\mathbf{I}} \mathbf{Z}^*$ дает ту же самую объясненную сумму квадратов, что и регрессия $\tilde{\boldsymbol{\mu}} \hat{\mathbf{I}} \mathbf{1} \in \mathbf{Z}$. Таким образом, окончательно получаем, что LM-статистика для тестирования гетероскедастичности равна половине объясненной суммы квадратов из регрессии $\tilde{\boldsymbol{\mu}} \hat{\mathbf{I}}$ константе $\in \mathbf{Z}$. Статистика распределена асимптотически как $\chi^2(r)$, где r — размерность вектора $\boldsymbol{\gamma}$.

Примечательно, что в этой статистике не фигурируют производные функции $h(\cdot)$, формула будет одна и та же независимо от выбора $h(\cdot)$. Когда статистика множителя Лагранжа одна и та же для широкого класса альтернативных гипотез, тогда эти альтернативные модели принято называть **локально эквивалентными альтернативами**.

Регрессия с мультипликативной гетероскедастичностью

В регрессии с (линейной) **мультипликативной гетероскедастичностью** дисперсия ошибки равна

$$\sigma_i^2(\boldsymbol{\alpha}) = \exp(\mathbf{Z}_i \boldsymbol{\alpha}).$$

Здесь \mathbf{Z} — матрица, состоящая из переменных, от которых зависит дисперсия (как правило, в ней должен быть столбец, состоящий из единиц), $\boldsymbol{\alpha}$ — вектор параметров.

Регрессия задана формулой:

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim \text{NID}(0, \sigma_i^2(\boldsymbol{\alpha})).$$

Предполагается, что неизвестные параметры в “среднем” и в дисперсии не связаны между собой.

Логарифмическая функция правдоподобия i -го наблюдения для этой модели имеет вид:

$$\begin{aligned} \ell_i &= -\frac{1}{2} \ln(2\pi\sigma_i^2(\boldsymbol{\alpha})) - \frac{1}{2\sigma_i^2(\boldsymbol{\alpha})} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 = \\ &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \mathbf{Z}_i\boldsymbol{\alpha} - \frac{e_i^2}{2 \exp(\mathbf{Z}_i\boldsymbol{\alpha})}. \end{aligned}$$

Найдем вклад в градиент i -го наблюдения:

$$\begin{aligned} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma_i^2(\boldsymbol{\alpha})} e_i \mathbf{X}_i, \\ \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} &= -\frac{1}{2} \mathbf{Z}_i + \frac{e_i^2}{2 \exp(\mathbf{Z}_i\boldsymbol{\alpha})} \mathbf{Z}_i = \frac{1}{2} \left(\frac{e_i^2}{\sigma_i^2(\boldsymbol{\alpha})} - 1 \right) \mathbf{Z}_i. \end{aligned}$$

Вклад в информационную матрицу i -го наблюдения равен

$$\begin{aligned} \mathbb{E} \left(\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}}^\top \right) &= \frac{\mathbb{E}(\varepsilon_i^2)}{\sigma_i^4} \mathbf{X}_i \mathbf{X}_i^\top = \frac{1}{\sigma_i^2} \mathbf{X}_i \mathbf{X}_i^\top, \\ \mathbb{E} \left(\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}}^\top \right) &= \frac{1}{2} \mathbb{E} \left(\frac{\varepsilon_i}{\sigma_i^2(\boldsymbol{\alpha})} \left(\frac{\varepsilon_i^2}{\sigma_i^2(\boldsymbol{\alpha})} - 1 \right) \right) \mathbf{X}_i \mathbf{Z}_i^\top = 0, \\ \mathbb{E} \left(\frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}}^\top \right) &= \frac{1}{4} \mathbb{E} \left(\frac{\varepsilon_i^4}{\sigma_i^4(\boldsymbol{\alpha})} - 2 \frac{\varepsilon_i^2}{\sigma_i^2(\boldsymbol{\alpha})} + 1 \right) \mathbf{Z}_i \mathbf{Z}_i^\top = \\ &= \frac{1}{4} (3 - 2 + 1) \mathbf{Z}_i \mathbf{Z}_i^\top = \frac{1}{2} \mathbf{Z}_i \mathbf{Z}_i^\top. \end{aligned}$$

Таким образом, информационная матрица (как и следовало ожидать) блочно-диагональная и блоки ее равны:

$$\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^N = \mathbf{X}^\top \text{diag} \left(\frac{1}{\sigma_1^2}, \dots, \frac{1}{\sigma_N^2} \right) \mathbf{X}, \quad \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N = \frac{1}{2} \mathbf{Z}^\top \mathbf{Z}.$$

При данном векторе $\boldsymbol{\alpha}$, коэффициенты регрессии $\boldsymbol{\beta}$ можно найти из взвешенной регрессии:

$$\boldsymbol{\beta} = (\mathbf{X}^{*\top} \mathbf{X}^*)^{-1} (\mathbf{X}^{*\top} \mathbf{Y}^*),$$

где $\mathbf{X}^* = \frac{1}{\sigma_i} \mathbf{X}$, $\mathbf{Y}^* = \frac{1}{\sigma_i} \mathbf{Y}$. Обозначим остатки из этой регрессии $\mathbf{e}^*(\boldsymbol{\alpha})$.

Используем итерации по $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{t+1} = \boldsymbol{\alpha}^t + \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N(\boldsymbol{\alpha}^t)^{-1} \mathbf{g}^{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^t + \Delta \boldsymbol{\alpha}^t.$$

$\Delta\alpha^t$ можно находить с помощью вспомогательной регрессии $\frac{1}{\sqrt{2}} \tilde{\mu}(\alpha^t)$ по $\frac{1}{\sqrt{2}} \mathbf{Z}$, где $\tilde{\mu}_i = \frac{e_i^2}{\sigma_i^2} - 1 = (e_i^*)^2 - 1$.

Обе используемые в этом алгоритме вспомогательные регрессии дают состоятельные оценки ковариационных матриц соответствующих оценок параметров и могут использоваться для проверки гипотез.

Нелинейная регрессия. Метод Гаусса-Ньютона

Пусть нелинейная регрессия задана уравнением

$$Y_i = f_i(\theta) + \varepsilon_i,$$

Имея на t -м шаге приближение θ^t , следующее приближение θ^{t+1} получаем с помощью регрессии $Y - f(\theta^t)$ по $F(\theta^t)$, где $F(\cdot)$ — матрица производных f по θ :

$$F_{ij} = \frac{\partial f_i}{\partial \theta_j}.$$

По сути дела мы здесь линеаризируем функцию f в окрестности точки θ^t . Пусть $\Delta\theta^t$ — оценки ОМНК из этой вспомогательной регрессии:

$$\Delta\theta^t = (F(\theta^t)^T F(\theta^t))^{-1} F(\theta^t)^T (Y - f(\theta^t)).$$

Тогда следующее приближение метода Гаусса-Ньютона будет:

$$\theta^{t+1} = \theta^t + \Delta\theta^t.$$

Повторяем эти итерации пока метод не сойдется.

Последняя из регрессий Гаусса-Ньютона даст состоятельную оценку матрицы ковариаций оценок $\hat{\theta}$ ($\hat{\sigma}^2 (\hat{F}^T \hat{F})^{-1}$), при условии, что верны обычные предположения: что ε_i независимо нормально распределены с нулевым мат. ожиданием и одинаковой дисперсией. Ясно, что можно, используя t - и F -статистики из этой вспомогательной регрессии, проверять различные гипотезы по принципу теста Вальда. Таким образом, регрессия Гаусса-Ньютона является искусственной регрессией.

Оценивание регрессии с AR-ошибкой

AR(1) имеет вид:

$$\varepsilon_i = \rho \varepsilon_{i-1} + \xi_i.$$

“Инновации” ξ_i независимы и имеют одинаковую дисперсию ω^2 . Из этого следует, что ε_{i-1} и ξ_i независимы. Дисперсия ε_i находится из соотношения

$$\begin{aligned} V(\varepsilon_i) &= V(\rho \varepsilon_{i-1}) + \omega^2, \\ \sigma^2 &= \rho^2 \sigma^2 + \omega^2, \end{aligned}$$

где $\sigma^2 = V(\varepsilon_i)$. Отсюда

$$\sigma^2 = \frac{\omega^2}{1 - \rho^2}.$$

Чтобы дисперсия не увеличивалась до бесконечности с ростом количества наблюдений, должно выполняться условие стационарности $|\rho| < 1$.

Найдем матрицу ковариаций ошибок ε_i . Рекуррентную формулу для ε_i можно развернуть следующим образом:

$$\varepsilon_i = \xi_i + \rho \xi_{i-1} + \rho^2 \xi_{i-2} + \dots + \rho^\tau \xi_{i-\tau}.$$

$$\text{Отсюда } \text{cov}(\varepsilon_i, \varepsilon_{i-\tau}) = \text{cov}(\rho^\tau \varepsilon_{i-\tau}, \varepsilon_{i-\tau}) = \rho^\tau \sigma^2.$$

Получаем следующую ковариационную матрицу

$$V(\sigma^2, \rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{N-1} \\ \rho & \ddots & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{N-1} & \dots & \rho & 1 \end{bmatrix} = \sigma^2 W(\rho).$$

Предположим, что в линейной регрессионной модели $Y_i = X_i \beta + \varepsilon_i$ ошибка порождается авторегрессионным процессом первого порядка. Рассмотрим два способа оценивания такой регрессионной модели.

Нелинейная регрессия с пропущенным первым наблюдением

Подставим $\varepsilon_i = Y_i - X_i \beta$ в уравнение авторегрессионного процесса:

$$Y_i - X_i \beta = \rho(Y_{i-1} - X_{i-1} \beta) + \xi_i.$$

Получим следующую нелинейную регрессионную модель:

$$Y_i = X_i \beta + \rho(Y_{i-1} - X_{i-1} \beta) + \xi_i.$$

Ошибки ξ_i независимы с одинаковой дисперсией и их ковариационная матрица равна $\omega^2 I$.

Для оценивания нелинейной регрессии можно использовать метод Гаусса-Ньютона. Поскольку Y_0 , X_0 и ε_0 неизвестны, то берут $\varepsilon_0 = 0$. С помощью вспомогательной регрессии метода Гаусса-Ньютона можно не только оценить модель, но и проверить гипотезу об отсутствии автокорреляции ($\rho = 0$).

Если неизвестно, есть ли автокорреляция ошибок, лучше сначала получить оценки ОМНК и проверить гипотезу в этой точке (принцип LM-теста).

Такой тест можно реализовать воспользовавшись той же регрессией Гаусса-Ньютона. В точке оценок ОМНК (когда $\rho = 0$) матрица производных равна $[X, e_{-1}]$, т.е. к X добавляется столбец лагов остатков (где первое наблюдение равно нулю). Вспомогательная регрессия имеет вид

$$e = Xb + r e_{-1} + \text{ошибка.}$$

Проверяем гипотезу, что $r = 0$. Для этого можно использовать обычную t -статистику. Поскольку модель линейна, то это все равно, что тест на добавление переменной e_{-1} в исходную регрессию, так как можно заменить в левой части e на Y .

Этот тест и этот метод годятся даже тогда, когда в правой части стоят лаги зависимой переменной (DW-статистика в этом случае непригодна). Идею теста предложил Дарбин.

Описанный метод дает оценки МП при предположении, что ошибки распределены нормально, но не является точным методом МП, поскольку не учитывает распределение первого наблюдения.

Оценивание регрессии с AR(1)-ошибкой полным методом максимального правдоподобия

Ковариационная матрица ошибок ε_i имеет вид:

$$V(\sigma^2, \rho) = \sigma^2 W(\rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{N-1} \\ \rho & \ddots & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{N-1} & \dots & \rho & 1 \end{bmatrix}.$$

Дисперсии ε_i и ξ_i связаны между собой соотношением $\sigma^2 = \frac{\omega^2}{1-\rho^2}$.

Можно проверить, что

$$W^{-1} = (1-\rho^2) \begin{bmatrix} 1 & -\rho & & & \mathbf{O} \\ -\rho & 1+\rho^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ \mathbf{O} & & \ddots & 1+\rho^2 & -\rho \\ & & & -\rho & 1 \end{bmatrix}.$$

Применим следующее разложение Холецкого: $TT^T = W^{-1}(1-\rho^2)$, где

$$T = \begin{bmatrix} \sqrt{1-\rho^2} & -\rho & & & \mathbf{O} \\ & 1 & \ddots & & \\ & \mathbf{O} & & \ddots & -\rho \\ & & & & 1 \end{bmatrix}.$$

Определим переменные вспомогательной регрессии следующим образом:

$$Y^* = T^T Y, \quad X^* = T^T X.$$

При фиксированном ρ регрессия Y^* по X^* дает оценки максимума правдоподобия для β (это будет оценка МП только в том случае, если X не содержит лагов зависимой переменной !!!). Этот прием называется преобразованием Прэйса-Винстена (Prais-Winsten transformation). Распишем его более подробно:

$$\begin{aligned} Y_1^* &= \sqrt{1-\rho^2} Y_1, & X_{1j}^* &= \sqrt{1-\rho^2} X_{1j}, \\ Y_i^* &= Y_i - \rho Y_{i-1}, & X_{ij}^* &= X_i - \rho X_{i-1j} \quad \forall i > 1. \end{aligned}$$

Как и следовало ожидать, при $i > 1$ преобразование совпадает с рассмотренным выше преобразованием при пропущенном первом наблюдении.

В данном случае формула для первого наблюдения отличается от формулы для прочих наблюдений. Поскольку в ней отсутствуют лаги, то ошибка будет равна ε_i , а не ξ_i и дисперсия для первого наблюдения будет равна $\frac{\omega^2}{1-\rho^2}$, а не σ^2 . Поэтому первое наблюдение домножается на $\sqrt{1-\rho^2}$, чтобы избавиться от гетероскедастичности.

Пусть $\beta(\rho)$ — оценки коэффициентов из вспомогательной регрессии, $e^*(\rho)$ — остатки из вспомогательной регрессии. Мы можем максимизировать по ρ концентрированную функцию правдоподобия:

$$\begin{aligned} \ell^c(\rho) &= -\frac{1}{2} \ln |W(\rho)| - \frac{N}{2} \ln ((Y - X\beta(\rho))^T W(\rho)^{-1} (Y - X\beta(\rho))) + \\ &+ \text{const} = \frac{1}{2} \ln(1-\rho^2) - \frac{N}{2} \ln(e^*(\rho)^T e^*(\rho)) + \text{const} \rightarrow \max_{\rho}. \end{aligned}$$

Здесь мы воспользовались тем, что $TT^T = W^{-1}(1-\rho^2)$ и

$$\begin{aligned} \ln |W| &= -\ln |W^{-1}| = -N \ln(1-\rho^2) - \ln |TT^T| = \\ &= -N \ln(1-\rho^2) - 2 \ln |T| = -N \ln(1-\rho^2) - 2 \ln(\sqrt{1-\rho^2}) = \\ &= -(N+1) \ln(1-\rho^2). \end{aligned}$$

Можно показать, что условие первого порядка максимума концентрированной функции правдоподобия представимо в виде кубического уравнения. Максимум находится как средний корень этого уравнения.

Удобно, что при этом оценка ρ всегда лежит в интервале стационарности $(-1; 1)$.

Информационная матрица, как и всегда в модели обобщенного метода наименьших квадратов, является блочно-диагональной. Приводим ее выборочный аналог без доказательства

$$\mathcal{I}^N(\hat{\theta})^{-1} = \begin{bmatrix} \hat{\omega}^2 (\hat{X}^{*\top} \hat{X}^*)^{-1} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \frac{N}{1-\hat{\rho}^2} + \frac{3\hat{\rho}^2-1}{(1-\hat{\rho}^2)^2} & \frac{2\hat{\rho}}{\hat{\omega}(1-\hat{\rho})^2} \\ \mathbf{O} & \frac{2\hat{\rho}}{\hat{\omega}(1-\hat{\rho})^2} & \frac{2N}{\hat{\omega}^2} \end{bmatrix}.$$

Иногда первое наблюдение очень важно и добавляет много новой информации.

Регрессия с МА-ошибкой

Оценивание регрессия с МА(1)-процессом в ошибке полным методом максимального правдоподобия

Будем рассматривать регрессию $Y = X\beta + \varepsilon$ с МА(1)-процессом в ошибке:

$$\begin{aligned} \varepsilon_i &= \xi_i + \mu \xi_{i-1} & \xi_i &\sim N(0, \omega^2) \\ \sigma^2 &= \text{var}(\varepsilon_i) = (1 + \mu^2)\omega^2 \end{aligned}$$

Ковариационная матрица ошибок ε_i имеет вид $V(\sigma^2, \mu) = \sigma^2 W(\mu)$.

$$W = \begin{bmatrix} 1 + \mu^2 & \mu & & \mathbf{O} \\ \mu & 1 + \mu^2 & \ddots & \\ & \ddots & \ddots & \ddots \\ \mathbf{O} & & \ddots & 1 + \mu^2 & \mu \\ & & & \mu & 1 + \mu^2 \end{bmatrix} = (1 + \mu^2)\mathbf{I} + \mu\Phi,$$

$$\text{где } \Phi = \begin{bmatrix} 0 & 1 & & \mathbf{O} \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & \ddots \\ \mathbf{O} & & \ddots & 0 & 1 \\ & & & 1 & 0 \end{bmatrix}.$$

Симметричную положительно определенную матрицу можно представить в виде $W = H^T \Lambda H$, где H — ортогональная матрица собственных векторов ($H^T = H^{-1}$), а Λ — диагональная матрица, диагональ которой состоит из соответствующих собственных чисел. В данном случае, собственные вектора матрицы W совпадают с собственными векторами матрицы Φ , и поэтому не зависят от μ . Типичный элемент матрицы H равен

$$H_{kl} = \sin\left(\frac{kl\pi}{N+1}\right) \sqrt{\frac{2}{N+1}}.$$

Типичный диагональный элемент матрицы Λ (собственное число) равен

$$\lambda_k = \mu^2 + 2\mu \cos\left(\frac{2\pi k}{N+1}\right) + 1.$$

Матрица W такова, что

$$W^{-1} = H^T \Lambda^{-1} H.$$

Несложно также найти определитель матрицы W :

$$|W| = \frac{1 - \mu^{2N+2}}{1 - \mu^2}.$$

Обозначим $Y^* = DHY$, $X^* = DHX$, где $D = \Lambda^{-1/2}$ — диагональная матрица.

Пусть $e^*(\mu)$ — остатки из вспомогательной регрессии, $\beta(\mu)$ — оценки коэффициентов из этой регрессии. Тогда

$$(Y - X\beta(\mu))^T W(\mu)^{-1} (Y - X\beta(\mu)) = e^*(\mu)^T e^*(\mu).$$

Концентрированная функция правдоподобия после исключения параметров σ^2 и β приобретает вид

$$\ell^c(\mu) = -\frac{1}{2} (\ln(1 - \mu^{2N+2}) - \ln(1 - \mu^2)) - \frac{N}{2} \ln(e^*(\mu)^T e^*(\mu)) + \text{const}.$$

Остается с помощью одномерного поиска максимизировать концентрированную функцию правдоподобия по μ на отрезке $[-1, 1]$. В максимуме функции правдоподобия можно с ненулевой вероятностью получить $\mu = 1$ или $\mu = -1$.

Можно предложить и другую вспомогательную регрессию, которая применима и в общем случае $MA(l)$ -процесса. Обозначим $\eta = \xi_0$. Тогда модель можно преобразовать к виду

$$\begin{aligned} 0 &= -\eta + \xi_0, \\ Y_1 &= X_1\beta + \mu\eta + \xi_1, \\ Y_2 - \mu Y_1 &= (X_2 - \mu X_1)\beta - \mu^2\eta + \xi_2, \end{aligned}$$

и так далее для $i = 3, \dots, N$.

Более компактно это можно записать как уравнение регрессии:

$$\dot{Y} = \dot{X}\beta + \dot{Z}\eta + \xi.$$

Здесь \dot{Y} , \dot{X} и \dot{Z} имеют по $N+1$ наблюдению и вычисляются по рекуррентным формулам:

$$\dot{Y}_i = Y_i - \mu \dot{Y}_{i-1}, \quad \dot{Y}_0 = 0,$$

$$\begin{aligned}\dot{X}_i &= X_i - \mu \dot{X}_{i-1}, & \dot{X}_0 &= \mathbf{0}, \\ \dot{Z}_i &= -\mu \dot{Z}_{i-1}, & \dot{Z}_0 &= -1.\end{aligned}$$

Пусть $\xi(\mu)$ — остатки из вспомогательной регрессии, $\beta(\mu)$ — оценки коэффициентов β из этой регрессии. Тогда можно показать, что $(Y - X\beta(\mu))^T W(\mu)^{-1} (Y - X\beta(\mu)) = \xi(\mu)^T \xi(\mu)$. Соответственно, концентрированная функция правдоподобия равна

$$\ell^c(\mu) = -\frac{1}{2} (\ln(1 - \mu^{2N+2}) - \ln(1 - \mu^2)) - \frac{N}{2} \ln(\xi(\mu)^T \xi(\mu)) + \text{const.}$$

Оценивание регрессии с МА-ошибкой нелинейным МНК

Как и в случае регрессии с AR-процессом, можно получить оценку, которая асимптотически эквивалентна точной оценке максимального правдоподобия, если пренебречь первыми наблюдениями. В данном случае удобно считать, что довыборочные ошибки ξ_i ($i < 1$) равны нулю. При этом из функции правдоподобия исчезает мешающий член $-1/2 (\ln(1 - \mu^{2N+2}) - \ln(1 - \mu^2))$, и модель сводится к нелинейной регрессии, которую можно оценить с помощью метода Гаусса-Ньютона. Требуется минимизировать сумму квадратов остатков

$$\sum_{i=1}^N \xi_i^2(\beta, \mu) \rightarrow \min.$$

Остатки вычисляются рекуррентно по формуле

$$\xi_i(\beta, \mu) = Y_i - X_i \beta - \mu \xi_{i-1}(\beta, \mu) \quad (\xi_0(\beta, \mu) = 0).$$

Производные функции $\xi_i(\beta, \mu)$, необходимые для использования метода Гаусса-Ньютона также находятся рекуррентно:

$$\begin{aligned}\frac{\partial \xi_i}{\partial \beta} &= -X_i - \mu \frac{\partial \xi_{i-1}}{\partial \beta} \quad \left(\frac{\partial \xi_0}{\partial \beta} = 0 \right), \\ \frac{\partial \xi_i}{\partial \mu} &= -\mu \frac{\partial \xi_{i-1}}{\partial \mu} - \xi_{i-1} \quad \left(\frac{\partial \xi_0}{\partial \mu} = 0 \right).\end{aligned}$$

Регрессия с ARCH-процессом в ошибке

Часто в эконометрических моделях остатки становятся то большими на какой-то период, то не очень большими, и так далее и в этом нет определенной закономерности. Особенно это относится к моделям финансовых рынков. Даже если безусловная дисперсия ошибок постоянна, условная дисперсия может быть подвержена случайным колебаниям. Условные прогнозы диспер-

сии могут иметь практическое значение. Владельцев активов интересует, как получить прогноз риска на ряд следующих периодов, если известна информация за текущий и предшествующие периоды.

Для моделирования таких процессов используется понятие условной авторегрессионной гетероскедастичности (ARCH, autoregressive conditional heteroskedasticity). Ввел это понятие Энгл (Engle, 1982). Основная идея состоит в том, что дисперсия ошибки в момент i зависит от величины квадрата ошибки в предшествующие периоды времени.

Процесс ARCH(p) имеет следующий вид:

$$\sigma_i^2 \equiv E(\varepsilon_i^2 | \Omega_i) = \mu + \gamma_1 \varepsilon_{i-1}^2 + \gamma_2 \varepsilon_{i-2}^2 + \dots + \gamma_p \varepsilon_{i-p}^2.$$

Ω_i обозначает информацию, которая имеется к моменту i , т. е. все наблюдаемые переменные в момент $i-1$ и ранее. В частности, для ARCH(1)

$$\sigma_i^2 = \mu + \gamma_1 \varepsilon_{i-1}^2.$$

Предполагают, что $\mu > 0$, $\gamma_k \geq 0$, чтобы дисперсия не могла оказаться отрицательной.

В случае, когда ε_i имеют нормальное распределение, очередное значение задается формулой

$$\varepsilon_i = \sigma_i \xi_i, \quad \text{где } \xi_i \sim \text{NID}(0,1).$$

или
$$\varepsilon_i = \xi_i \sqrt{\mu + \gamma_1 \varepsilon_{i-1}^2 + \gamma_2 \varepsilon_{i-2}^2 + \dots + \gamma_p \varepsilon_{i-p}^2}.$$

Такой процесс имеет большое сходство с авторегрессионным.

Обозначим безусловную дисперсию через σ^2 . Тогда, если взять ожидания от обеих частей в формуле для дисперсии, получим

$$\sigma^2 = \mu + \gamma_1 \sigma^2 + \gamma_2 \sigma^2 + \dots + \gamma_p \sigma^2.$$

Здесь мы воспользовались формулой полного математического ожидания:

$$E(\sigma_i^2) = E(E(\varepsilon_i^2 | \Omega_i)) = E(\varepsilon_i^2) = \sigma^2.$$

Получаем выражение для безусловной дисперсии

$$\sigma^2 = \frac{\mu}{1 - \sum_{k=1}^p \gamma_k}.$$

Отсюда видно, что для того, чтобы ARCH-процесс был стационарным необходимо, чтобы $\sum_{k=1}^p \gamma_k < 1$.

На практике структура лага $\gamma_1, \dots, \gamma_p$ неизвестна и ее приходится оценивать. Если p большое, то приходится оценивать слишком много коэффициентов. В таком случае можно наложить априорные ограничения. Энгл, например, взял

$$\gamma_1 = 0.4\gamma^*, \quad \gamma_2 = 0.3\gamma^*, \quad \gamma_3 = 0.2\gamma^*, \quad \gamma_4 = 0.1\gamma^*,$$

поэтому вместо четырех коэффициентов, ему надо было оценить только один γ^* .

Если воспользоваться бесконечным геометрическим лагом, то можно преобразовать модель к виду

$$\sigma^2 = \mu + \gamma_1 \varepsilon_{i-1}^2 + \delta_1 \sigma_{i-1}^2.$$

Эта модель называется GARCH(1,1). Как ARCH похожа AR на, так GARCH похожа на ARMA. Модель GARCH предложена Боллерселевом (Bollerslev, 1986).

GARCH(p, q) имеет вид:

$$\begin{aligned} \sigma_i^2 &= \mu + \gamma_1 \varepsilon_{i-1}^2 + \dots + \gamma_p \varepsilon_{i-p}^2 + \delta_1 \sigma_{i-1}^2 + \dots + \delta_q \sigma_{i-q}^2 = \\ &= \mu + \sum_{k=1}^p \gamma_k \varepsilon_{i-k}^2 + \sum_{k=1}^q \delta_k \sigma_{i-k}^2. \end{aligned}$$

Безусловная дисперсия равна

$$\sigma^2 = \frac{\mu}{1 - \sum_{k=1}^p \gamma_k - \sum_{k=1}^q \delta_k}.$$

Регрессионная модель с нормально распределенной GARCH-ошибкой имеет вид:

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i^2),$$

где σ_i^2 вычисляется по данной выше формуле. К сожалению, не существует простого способа оценивания регрессии с GARCH-процессом в ошибке.

Логарифмическая функция правдоподобия для одного наблюдения равна

$$\ell_i = \frac{1}{2} \ln \sigma_i^2 - \frac{1}{2} \frac{e_i^2}{\sigma_i^2} = \frac{1}{2} \ln \sigma_i^2 - \frac{1}{2} \frac{1}{\sigma_i^2} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2,$$

для всего вектора наблюдений

$$\ell = \sum_i \ell_i = \frac{1}{2} \sum_i \ln \sigma_i^2 - \frac{1}{2} \sum_i \frac{e_i^2}{\sigma_i^2} =$$

$$= \frac{1}{2} \sum_i \ln \sigma_i^2 - \frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2,$$

где σ_i^2 вычисляется рекуррентно:

$$\sigma_i^2(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \mu + \sum_{k=1}^p \gamma_k e_{i-k}^2(\boldsymbol{\beta}) + \sum_{k=1}^q \delta_k \sigma_{i-k}^2(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}).$$

Обозначим $\boldsymbol{\alpha} = (\mu, \gamma_1, \dots, \gamma_p, \delta_1, \dots, \delta_q)^\top$, $\mathbf{z}_i = (1, \varepsilon_{i-1}^2, \dots, \varepsilon_{i-p}^2, \sigma_{i-1}^2, \dots, \sigma_{i-p}^2)^\top$.

Тогда $\sigma_i^2 = \boldsymbol{\alpha}^\top \mathbf{z}_i$.

Вклад в градиент отдельного наблюдения:

$$\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} = \frac{1}{2} \frac{1}{\sigma_i^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} \left(\frac{e_i^2}{\sigma_i^2} - 1 \right) + \frac{1}{\sigma_i^2} e_i \mathbf{X}_i,$$

$$\frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} = \frac{1}{2} \frac{1}{\sigma_i^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \left(\frac{e_i^2}{\sigma_i^2} - 1 \right).$$

Производные условной дисперсии вычисляются рекуррентно:

$$\begin{aligned} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} &= -2 \sum_{k=1}^p \gamma_k \mathbf{X}_{i-k} e_{i-k} + \sum_{k=1}^q \delta_k \frac{\partial \sigma_{i-k}^2}{\partial \boldsymbol{\beta}}, \\ \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} &= \mathbf{z}_i + \sum_{k=1}^q \delta_k \frac{\partial \sigma_{i-k}^2}{\partial \boldsymbol{\alpha}}. \end{aligned}$$

Приведенные формулы уже позволяют применить метод ВННН (OPG), как предложил Боллерслев. Однако, метод, как обычно, работает плохо в смысле сходимости и точности оценивания ковариационной матрицы оценок МП. Другие методы, — Ньютона, удвоенной регрессии, — работают гораздо лучше. Рассмотрим метод, использующий оценку информационной матрицы, в которой ожидания берутся только частично — только условные ожидания.

Чтобы найти оценку информационной матрицы, возьмем в точке истинных параметров математическое ожидание условное по информации, имеющейся на момент i (Ω_i), от матрицы внешнего произведения градиента i -го наблюдения. Поскольку ожидание условное, то единственной случайной компонентой будет ошибка (в точке истинных параметров остатки равны ошибкам). При этом пользуемся тем, что

$$E\left(\left(\frac{\varepsilon_i^2}{\sigma_i^2} - 1\right) \varepsilon_i \mid \Omega_i\right) = 0, \quad E\left(\left(\frac{\varepsilon_i^2}{\sigma_i^2} - 1\right)^2 \mid \Omega_i\right) = 2 \quad \text{и} \quad E(\varepsilon_i^2 \mid \Omega_i) = \sigma_i^2.$$

$$E\left(\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}^\top} \mid \Omega_i\right) = -\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top} - \frac{1}{\sigma_i^2} \mathbf{X}_i \mathbf{X}_i^\top,$$

$$E\left(\frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}^\top} \mid \Omega_i\right) = -\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}^\top},$$

$$E\left(\frac{\partial \ell_i}{\partial \boldsymbol{\alpha}} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}^\top} \mid \Omega_i\right) = -\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top}.$$

Информационная матрица равна безусловному ожиданию суммы условных мат. ожиданий гессианов со знаком минус:

$$\mathcal{I}^N(\boldsymbol{\theta}) = -E(\sum_i E\left(\frac{\partial \ell_i}{\partial \boldsymbol{\theta}} \frac{\partial \ell_i}{\partial \boldsymbol{\theta}^\top} \mid \Omega_i\right))$$

$$\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^N = \sum_i E\left(\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top} + \frac{1}{\sigma_i^2} \mathbf{X}_i \mathbf{X}_i^\top\right),$$

$$\mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N = \sum_i E\left(\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}^\top}\right), \quad \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^N = \sum_i E\left(-\frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top}\right).$$

Можно показать, хотя доказательство достаточно громоздкое, что $\mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^N = \mathbf{O}$. Таким образом, информационная матрица является блочно-диагональной между коэффициентами регрессии и параметрами GARCH-процесса.

Информационную матрицу можно состоятельно оценить матрицей:

$$\hat{\mathcal{I}}^N = \begin{bmatrix} \hat{\mathcal{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^N & \mathbf{O} \\ \mathbf{O} & \hat{\mathcal{I}}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N \end{bmatrix},$$

где

$$\hat{\mathcal{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^N = \sum_i \frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\beta}^\top} + \frac{1}{\sigma_i^2} \mathbf{X}_i \mathbf{X}_i^\top \text{ и } \hat{\mathcal{I}}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N = \sum_i \frac{1}{2} \frac{1}{(\sigma_i^2)^2} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}} \frac{\partial \sigma_i^2}{\partial \boldsymbol{\alpha}^\top}.$$

Для нахождения оценок максимального правдоподобия можно применить обычный алгоритм:

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + \lambda_{\boldsymbol{\beta}} (\hat{\mathcal{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^N)^{-1} \mathbf{g}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^t),$$

$$\boldsymbol{\alpha}^{t+1} = \boldsymbol{\alpha}^t + \lambda_{\boldsymbol{\alpha}} (\hat{\mathcal{I}}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^N)^{-1} \mathbf{g}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}^t).$$

Прежде чем оценивать GARCH-регрессию, имеет смысл проверить наличие GARCH-процесса в ошибке с помощью теста множителя Лагранжа в точ-

ке оценок ОМНК. Используем блочно-диагональность информационной матрицы:

$$LM = \tilde{\mathbf{g}}_{\alpha}^{\top} (\tilde{\mathbf{I}}_{\alpha\alpha}^N)^{-1} \tilde{\mathbf{g}}_{\alpha}$$

При нулевой гипотезе об отсутствии GARCH-процесса $\sigma_i^2 = \mu = \sigma^2$, и статистику можно упростить. Найдем эту статистику только в частном случае ARCH модели:

$$\frac{\partial \sigma_i^2}{\partial \alpha} = \mathbf{z}_i^{\top} \Rightarrow \frac{\partial \ell_i}{\partial \alpha} = \frac{1}{2} \frac{1}{\sigma^2} \mathbf{z}_i^{\top} \left(\frac{e_i^2}{\sigma^2} - 1 \right), \quad \tilde{\mathbf{I}}_{\alpha\alpha}^N = \sum_i \frac{1}{2} \frac{1}{(\sigma^2)^2} \mathbf{z}_i^{\top} \mathbf{z}_i.$$

Пусть $\boldsymbol{\mu}$ — вектор-столбец, состоящий из $\frac{e_i^2}{\sigma^2} - 1$, \mathbf{Z} — матрица, строками которой являются $\mathbf{z}_i^{\top} = (1, e_{i-1}^2, \dots, e_{i-p}^2)$. Тогда

$$\tilde{\mathbf{g}}_{\alpha}^{\top} = \frac{\partial \ell}{\partial \alpha} = \frac{1}{2} \frac{1}{\sigma^2} \boldsymbol{\mu}^{\top} \mathbf{Z}, \quad \tilde{\mathbf{I}}_{\alpha\alpha}^N = \frac{1}{2} \frac{1}{(\sigma^2)^2} \mathbf{Z}^{\top} \mathbf{Z}.$$

$$LM = \frac{1}{2} \boldsymbol{\mu}^{\top} \mathbf{Z} (\mathbf{Z}^{\top} \mathbf{Z})^{-1} \mathbf{Z}^{\top} \boldsymbol{\mu} \stackrel{a}{\sim} \chi^2(p).$$

Энгл предложил несколько изменить форму статистики. Для этого используется то, что поскольку ошибки распределены нормально, то

$$Plim \boldsymbol{\mu}^{\top} \boldsymbol{\mu} / N = 2.$$

Асимптотически эквивалентной статистикой будет

$$LM^* = \frac{N}{\boldsymbol{\mu}^{\top} \boldsymbol{\mu}} \boldsymbol{\mu}^{\top} \mathbf{Z} (\mathbf{Z}^{\top} \mathbf{Z})^{-1} \mathbf{Z}^{\top} \boldsymbol{\mu} \stackrel{a}{\sim} \chi^2(p).$$

Эта статистика равняется NR^2 , где R^2 — коэффициент детерминации в регрессии квадратов остатков по \mathbf{Z} , то есть по константе и p лагам квадратов остатков.

Якобиан преобразования плотности распределения в функции правдоподобия

Функция правдоподобия модели типа $\boldsymbol{\varepsilon} = f(Y, \boldsymbol{\theta}_1)$

Рассмотрим модель по отношению к которой регрессия является частным случаем:

$$\boldsymbol{\varepsilon} = f(Y, \boldsymbol{\theta}_1).$$

Здесь Y — зависимая переменная, $\boldsymbol{\varepsilon}$ — ошибка, причем Y и $\boldsymbol{\varepsilon}$ — вектора-столбцы одинаковой размерности. “Независимые переменные” (регрессоры)

X неявно содержится в функции $f(\cdot)$. θ_1 — неизвестные параметры. Обозначаем их θ_1 , а не θ , потому что распределение ошибок ε само может зависеть от вектора неизвестных параметров (θ_2), так что

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}.$$

В частном случае линейной регрессии

$$f(Y, \theta_1) = Y - X\beta, \quad \theta_1 = \beta, \quad \theta_2 = \sigma^2.$$

Как правило, при построении эконометрической модели делают предположение о распределении ошибки, а уже из распределения ошибки выводят распределение зависимой переменной. Таким образом, задача состоит в том, чтобы из плотности распределения ε получить плотность распределения Y (если мы имеем дело с непрерывным распределением).

Плотности распределения связаны между собой соотношением:

$$p_Y(Y, \theta) = p_\varepsilon(f(Y, \theta_1), \theta_2) \text{ abs } |\mathbf{J}(\theta_1)|,$$

где $\mathbf{J}(\theta_1)$ — матрица Якоби (якобиан), соответствующий преобразованию Y в ε :

$$\mathbf{J}(\theta_1) = \frac{\partial f}{\partial Y} = \left\{ \frac{\partial f_i}{\partial Y_l} \right\}_{il}$$

— матрица первых производных f по Y . В выражении для плотности здесь стоит модуль определителя якобиана.

Функция правдоподобия — это по определению плотность распределения Y . Таким образом, логарифмическая функция правдоподобия равна

$$\ell = \ln p_\varepsilon(f(Y, \theta_1), \theta_2) + \ln \text{abs } |\mathbf{J}(\theta_1)|.$$

Будем второе слагаемое здесь называть якобианным членом. Якобианный член уже присутствовал в логарифмических функциях правдоподобия, которые мы рассматривали выше (см. напр. регрессии с автокоррелированными ошибками). В модели с AR(1)-ошибкой $\varepsilon_i = \rho \varepsilon_{i-1} + \xi_i$, где $\varepsilon_i = Y_i - X_i \beta$. Выразим ξ через Y :

$$f_i(Y, \theta_1) = \xi_i = (Y_i - X_i \beta) - \rho(Y_{i-1} - X_{i-1} \beta), \quad i = 1, \dots, N.$$

$$f_1(Y, \theta_1) = \sqrt{1 - \rho^2} (Y_1 - X_1 \beta).$$

Здесь $\theta_1 = \begin{pmatrix} \beta \\ \rho \end{pmatrix}$. $f_1(Y, \theta_1)$ определена таким образом, чтобы все элементы f

имели одинаковую дисперсию. Для этой модели

$$\mathbf{J}(\boldsymbol{\theta}_1) = \frac{\partial \mathbf{f}}{\partial \mathbf{Y}} = \begin{bmatrix} \sqrt{1-\rho^2} & \mathbf{O} \\ -\rho & 1 \\ \vdots & \vdots \\ \mathbf{O} & -\rho \mathbf{1} \end{bmatrix}, \quad \text{abs } |\mathbf{J}(\boldsymbol{\theta}_1)| = \sqrt{1-\rho^2}.$$

Преобразование зависимой переменной. Модель Бокса-Кокса

В частном случае рассмотренной модели $\boldsymbol{\theta}_1$ состоит из $\boldsymbol{\beta}$ и δ и

$$f_i(Y_i, \boldsymbol{\theta}_1) = h_i(Y_i, \delta) - \mathbf{X}_i(\delta)\boldsymbol{\beta}.$$

Модель является квазирегрессионной. Здесь $\mathbf{X}(\delta)$ — матрица “регрессоров”, $\boldsymbol{\beta}$ — вектор регрессионных коэффициентов. Якобиан $\mathbf{J} = \frac{\partial \mathbf{f}}{\partial \mathbf{Y}} = \frac{\partial \mathbf{h}}{\partial \mathbf{Y}}$ зависит только от параметров δ и является диагональной матрицей.

Такая модель возникает из регрессии, если применить к зависимой переменной преобразование, зависящее от оцениваемых параметров.

Пусть ошибки нормально распределены $\varepsilon_i \sim N(0, \sigma^2)$ и некоррелированы.

$$p_{\varepsilon}(\mathbf{z}) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \mathbf{z}^T \mathbf{z}\right).$$

Логарифмическая функция правдоподобия для такой модели:

$$\begin{aligned} \ell &= \ln p_{\varepsilon}(\mathbf{f}(\mathbf{Y}, \boldsymbol{\theta}_1), \boldsymbol{\theta}_2) + \ln \text{abs } |\mathbf{J}(\boldsymbol{\theta}_1)| = \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \mathbf{f}^T \mathbf{f} + \ln \text{abs } \left| \frac{\partial \mathbf{f}}{\partial \mathbf{Y}} \right| = \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (h_i(Y_i, \delta) - \mathbf{X}_i(\delta)\boldsymbol{\beta})^2 + \sum_i \ln \text{abs} \left(\frac{\partial h_i}{\partial Y_i} \right). \end{aligned}$$

Самое популярное преобразование зависимой переменной — это **преобразование Бокса-Кокса**:

$$h_i(Y_i, \delta) = \frac{Y_i^{\delta} - 1}{\delta}.$$

В общем случае его можно применять при положительных Y_i .

Если $\delta \rightarrow 0$, то $\frac{Y_i^{\delta} - 1}{\delta} \rightarrow \ln Y_i$, поэтому берут $h(Y, 0) = \ln Y$.

Таким образом, имеем следующую простейшую модель Бокса-Кокса³:

$$\frac{Y_i^{\delta} - 1}{\delta} = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i.$$

³ В моделях этого типа переменные в правой части также могут подвергаться преобразованию Бокса-Кокса.

Здесь регрессоры X детерминированы и не зависят от неизвестных параметров.

Якобиан равен

$$\mathbf{J} = \begin{bmatrix} Y_1^{\delta-1} & & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & & Y_N^{\delta-1} \end{bmatrix}.$$

Функция правдоподобия для модели Бокса-Кокса равна

$$\ell = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i \left(\frac{Y_i^\delta - 1}{\delta} - X_i \boldsymbol{\beta} \right)^2 - (1 - \delta) \sum_i \ln Y_i.$$

Концентрируя функцию правдоподобия по σ^2 , получим

$$\ell^c = -\frac{N}{2} \ln(2\pi \frac{\text{RSS}}{N}) - \frac{N}{2} - (1 - \delta) \sum_i \ln Y_i.$$

Обозначим \check{Y} среднее геометрическое Y_i :

$$\check{Y} = (\prod_i Y_i)^{1/N}.$$

Тогда $\ell^c = -\frac{N}{2} \ln(\text{RSS}) - N(1 - \delta) \ln \check{Y} + \text{const}$.

Максимизация ℓ^c эквивалентна минимизации следующей суммы квадратов:

$$\sum_i (\check{Y}^{1-\delta} \left(\frac{Y_i^\delta - 1}{\delta} - X_i \boldsymbol{\beta} \right))^2.$$

Можно предложить два метода оценивания.

Первый метод заключается в одномерной минимизации по δ , поскольку при фиксированном δ задача сводится к ОМНК. Строим регрессию $\check{Y}^{1-\delta} (Y_i^\delta - 1)/\delta$ по $\check{Y}^{1-\delta} X_i$.

Второй метод заключается в использовании нелинейного МНК, в котором зависимой переменной является вектор, состоящий из нулей, а в правой части стоит $\check{Y}^{1-\delta} ((Y_i^\delta - 1)/\delta - X_i \boldsymbol{\beta})$.

В обоих случаях мы найдем МП-оценки, но не найдем состоятельную оценку матрицы ковариаций оценок (ковариационные матрицы из этих вспомогательных регрессий не годятся).

Тест на нормальность

Задача этого параграфа — получить статистику множителя Лагранжа, которая позволила бы проверять гипотезу о том, что ошибки в регрессии распределены нормально. Идея состоит в том, чтобы рассмотреть модель с ошибкой из некоторого семейства непрерывных распределений, так чтобы нормальное распределение было частным случаем. Удобно взять, например, пирсоновское семейство распределений.

Плотность распределения (с нулевым мат. ожиданием) из пирсоновского семейства задается экспонентой функции

$$\psi(\varepsilon, c) = \int \frac{\varepsilon}{c_2 t^2 - c_1 t + c_0} dt.$$

Поскольку интеграл плотности распределения должен быть равен 1, то эту функцию следует пронормировать:

$$p_\varepsilon(u) = \frac{\exp \psi(u)}{\int_{-\infty}^{+\infty} \exp \psi(t) dt}.$$

Нулевая гипотеза (“нормальность”) заключается в том, что ошибки в линейной регрессии Y по X распределены нормально. Нормальное распределение является пирсоновским распределением с параметрами $c_1 = 0$, $c_2 = 0$:

$$H_0: c_1 = 0, c_2 = 0 \Rightarrow \varepsilon \sim N(0, c_0) \text{ (при } c_0 = \sigma^2 \text{)}.$$

Логарифмическая функция правдоподобия есть логарифм плотности распределения. Для i -го наблюдения:

$$l_i = \psi(Y_i - X_i \beta) - \ln \int_{-\infty}^{+\infty} \exp \psi(t) dt.$$

Найдем вклад в градиент i -го наблюдения при выполнении нулевой гипотезы.

$$\frac{\partial l_i}{\partial \beta} = - \frac{\partial \psi}{\partial \varepsilon} X_i = - \frac{c_1 - e_i}{c_2 e_i^2 - c_1 e_i + c_0} X_i.$$

$$\left. \frac{\partial l_i}{\partial \beta} \right|_{H_0} = \frac{\varepsilon_i}{c_0} X_i = \frac{1}{\sigma^2} \varepsilon_i X_i.$$

Производные по параметрам c_k пирсоновского распределения равны

$$\begin{aligned} \frac{\partial \ell_i}{\partial c_k} &= \frac{\partial \Psi}{\partial c_k} - \frac{\int_{-\infty}^{+\infty} \frac{\partial \Psi(t)}{\partial c_k} \exp \psi(t) dt}{\int_{-\infty}^{+\infty} \exp \psi(t) dt} = \frac{\partial \Psi}{\partial c_k} - \int_{-\infty}^{+\infty} \frac{\partial \Psi(t)}{\partial c_k} p_{\varepsilon}(t) dt = \\ &= \frac{\partial \Psi}{\partial c_k} - E\left(\frac{\partial \Psi}{\partial c_k}\right) \quad (k = 0, 1, 2). \end{aligned}$$

Чтобы их вычислить, нужно вычислить производные функции $\psi(\cdot)$ по c_k ($k = 0, 1, 2$). Достаточно найти их при нулевой гипотезе:

$$\left. \frac{\partial \psi(u)}{\partial c_0} \right|_{H_0} = \frac{1}{c_0^2} \int^u t dt = \frac{u^2}{2\sigma^4}.$$

$$\left. \frac{\partial \psi(u)}{\partial c_1} \right|_{H_0} = \frac{1}{c_0} \int^u dt - \frac{1}{c_0^2} \int^u t^2 dt = \frac{u}{\sigma^2} - \frac{u^3}{3\sigma^4}.$$

$$\left. \frac{\partial \psi(u)}{\partial c_2} \right|_{H_0} = \frac{1}{c_0^2} \int^u t^3 dt = \frac{u^4}{4\sigma^4}.$$

Математические ожидания этих производных как функций от ε_i равны

$$E\left(\left. \frac{\partial \psi(\varepsilon_i)}{\partial c_0} \right|_{H_0}\right) = E\left(\frac{\varepsilon_i^2}{2\sigma^4}\right) = \frac{1}{2\sigma^2},$$

$$E\left(\left. \frac{\partial \psi(\varepsilon_i)}{\partial c_1} \right|_{H_0}\right) = E\left(\frac{\varepsilon_i}{\sigma^2} - \frac{\varepsilon_i^3}{3\sigma^4}\right) = 0,$$

$$E\left(\left. \frac{\partial \psi(\varepsilon_i)}{\partial c_2} \right|_{H_0}\right) = E\left(\frac{\varepsilon_i^4}{4\sigma^4}\right) = \frac{3}{4}.$$

Подставим найденные выражения в градиент логарифмической функции правдоподобия, введя обозначение $\tilde{\varepsilon}_i = \varepsilon_i/\sigma$.

$$G_{0i}^0 = \left. \frac{\partial \ell_i}{\partial c_0} \right|_{H_0} = \frac{1}{2\sigma^4} (\varepsilon_i^2 - \sigma^2) = \frac{1}{2\sigma^2} (\tilde{\varepsilon}_i^2 - 1),$$

$$G_{0i}^1 = \left. \frac{\partial \ell_i}{\partial c_1} \right|_{H_0} = \frac{\varepsilon_i}{\sigma^2} - \frac{\varepsilon_i^3}{3\sigma^4} = \frac{1}{3\sigma} (3\tilde{\varepsilon}_i - \tilde{\varepsilon}_i^3),$$

$$G_{0i}^2 = \left. \frac{\partial \ell_i}{\partial c_2} \right|_{H_0} = \frac{\varepsilon_i^4}{4\sigma^4} - \frac{3}{4} = \frac{1}{4} (\tilde{\varepsilon}_i^4 - 3).$$

В тех же обозначениях

$$G_{0i}^{\beta} = \left. \frac{\partial \ell_i}{\partial \beta} \right|_{H_0} = \frac{1}{\sigma} \tilde{\varepsilon}_i X_i.$$

Найдем информационную матрицу, учитывая, что моменты стандартного нормального распределения ($\eta \sim N(0,1)$) равны

$$E(\eta^k) = \begin{cases} 0 & k \text{ — нечетное} \\ 1 \cdot 3 \cdot \dots \cdot (k-1) & k \text{ — четное} \end{cases},$$

$$E(\eta^4) = 3, E(\eta^6) = 15, E(\eta^8) = 105.$$

Информационная матрица для i -го наблюдения:

$$E(G_{0i}^{\beta \top} G_{0i}^{\beta}) = \frac{1}{\sigma^2} \mathbf{X}_i^{\top} \mathbf{X}_i E \tilde{\varepsilon}_i^2 = \frac{1}{\sigma^2} \mathbf{X}_i^{\top} \mathbf{X}_i.$$

$$E(G_{0i}^0 G_{0i}^{\beta}) = \mathbf{0}^{\top}, \quad E(G_{0i}^2 G_{0i}^{\beta}) = \mathbf{0}^{\top},$$

$$E(G_{0i}^1 G_{0i}^{\beta}) = \frac{1}{3\sigma^2} (3E \tilde{\varepsilon}_i^2 - E \tilde{\varepsilon}_i^4) \mathbf{X}_i = \frac{1}{3\sigma^2} (3 - 3) \mathbf{X}_i = \mathbf{0}^{\top},$$

$$E((G_{0i}^0)^2) = \frac{1}{4\sigma^4} (E \tilde{\varepsilon}_i^4 - 2E \tilde{\varepsilon}_i^2 + 1) = \frac{1}{4\sigma^4} (3 - 2 + 1) = \frac{1}{2\sigma^4},$$

$$E(G_{0i}^0 G_{0i}^1) = 0, \quad E(G_{0i}^1 G_{0i}^2) = 0,$$

$$E((G_{0i}^1)^2) = \frac{1}{9\sigma^2} (E \tilde{\varepsilon}_i^6 - 6E \tilde{\varepsilon}_i^4 + 9E \tilde{\varepsilon}_i^2) = \frac{1}{9\sigma^2} (15 - 6 \cdot 3 + 9) = \frac{2}{3\sigma^2},$$

$$\begin{aligned} E(G_{0i}^0 G_{0i}^2) &= \frac{1}{8\sigma^2} (E \tilde{\varepsilon}_i^6 - E \tilde{\varepsilon}_i^4 - 3E \tilde{\varepsilon}_i^2 + 3) = \\ &= \frac{1}{8\sigma^2} (15 - 3 - 3 + 3) = \frac{3}{2\sigma^2}. \end{aligned}$$

$$E((G_{0i}^2)^2) = \frac{1}{16} (E \tilde{\varepsilon}_i^8 - 6E \tilde{\varepsilon}_i^4 + 9) = \frac{1}{16} (105 - 6 \cdot 3 + 9) = 6.$$

Просуммируем по всем наблюдениям и составим блок информационной матрицы, относящийся к \mathbf{c} . Поскольку информационная матрица блочно-диагональная между \mathbf{c} и $\boldsymbol{\beta}$, то для нахождения интересующей нас статистики достаточно этого блока:

$$\mathcal{I}_{cc} = N \begin{bmatrix} \frac{1}{2\sigma^4} & 0 & \frac{3}{2\sigma^2} \\ 0 & \frac{2}{3\sigma^2} & 0 \\ \frac{3}{2\sigma^2} & 0 & 6 \end{bmatrix}.$$

Обратная матрица:

$$(\mathcal{I}_{cc})^{-1} = \frac{1}{N} \begin{bmatrix} 8\sigma^4 & 0 & -2\sigma^2 \\ 0 & \frac{3}{2}\sigma^2 & 0 \\ -2\sigma^2 & 0 & \frac{2}{3} \end{bmatrix}.$$

Тест множителя Лагранжа равен $LM = \tilde{\mathbf{g}}_c^\top (\tilde{\mathcal{I}}_{cc})^{-1} \tilde{\mathbf{g}}_c$ и распределен асимптотически как χ^2 с 2-мя степенями свободы. Градиенты здесь равны (\tilde{e}_i — нормированные остатки)

$$\tilde{\mathbf{g}}_c = \begin{pmatrix} 0 \\ \frac{1}{3\sigma} \sum_i (3\tilde{e}_i - \tilde{e}_i^3) \\ \frac{1}{4} \sum_i (\tilde{e}_i^4 - 3) \end{pmatrix}.$$

Поэтому

$$LM = \frac{1}{6N} (\sum_i (3\tilde{e}_i - \tilde{e}_i^3))^2 + \frac{1}{24N} (\sum_i (\tilde{e}_i^4 - 3))^2.$$

Два слагаемых, составляющих эту статистику, асимптотически независимы, и каждое распределено как $\chi^2(1)$. Первое слагаемое представляет собой тест на асимметрию, а второе — тест на эксцесс. Эту же статистику можно получить и с помощью других семейств распределений. Здесь мы опять сталкиваемся с локально эквивалентными альтернативами.

Точно такой же подход может быть использован в других моделях с нормально распределенными ошибками. Авторы теста Жарк и Бера (Jarque, Bera), применили этот подход к пробиту и моделям с усеченной и цензурированной зависимой переменной.

Регрессия с ошибками во всех переменных

В классической регрессионной модели ошибка содержится в единственной переменной — той, которая стоит в левой части (зависимой переменной). Рассмотрим более общий случай.

Пусть имеется матрица ненаблюдаемых исходных переменных $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_M)$, $\mathbf{X} = \{\mathbf{X}_{ij}\}$, $i = 1, \dots, N$, $j = 1, \dots, M$. Эти переменные связаны между собой линейной зависимостью: $\mathbf{X}\boldsymbol{\beta} = 0$. Требуется оценить коэффициенты $\boldsymbol{\beta}$. Наблюдаются только переменные \mathbf{Y} , которые представляют собой переменные \mathbf{X} измеренные с ошибками:

$$\mathbf{Y}_j = \mathbf{X}_j + \boldsymbol{\varepsilon}_j.$$

Предполагается, что ошибки некоррелированы для разных номеров наблюдений, но ошибки, относящиеся к наблюдениям с одним и тем же номером i коррелированы, причем матрица ковариаций Ω *точно известна*. Для того, чтобы можно было воспользоваться методом максимального правдоподобия, делаем предположение, что ошибки распределены нормально.

Логарифмическая функция правдоподобия равна

$$\ell = -\frac{N}{2} \ln |\Omega| - \frac{1}{2} \sum_i \mathbf{E}_i \Omega^{-1} \mathbf{E}_i^T + \text{const} \rightarrow \max_{\mathbf{X}, \boldsymbol{\beta}}$$

где $\mathbf{E}_i = \mathbf{Y}_j - \mathbf{X}_j$ — i -я строка матрицы остатков.

Максимизируем функцию правдоподобия при ограничении $\mathbf{X}\boldsymbol{\beta} = 0$.

Задачу максимизации можно записать с помощью лагранжиана:

$$\begin{aligned} \mathbf{L} &= -\frac{1}{2} \text{Tr}((\mathbf{Y} - \mathbf{X}) \Omega^{-1} (\mathbf{Y} - \mathbf{X})^T) - \boldsymbol{\lambda}^T \mathbf{X}\boldsymbol{\beta} = \\ &= -\frac{1}{2} \text{Tr}(\mathbf{Y}\Omega^{-1}\mathbf{Y}^T) + \text{Tr}(\mathbf{X}\Omega^{-1}\mathbf{Y}^T) - \frac{1}{2} \text{Tr}(\mathbf{X}\Omega^{-1}\mathbf{X}^T) - \boldsymbol{\lambda}^T \mathbf{X}\boldsymbol{\beta}. \\ \frac{\partial \mathbf{L}}{\partial \mathbf{X}} &= \Omega^{-1} \mathbf{Y}^T - \Omega^{-1} \mathbf{X}^T - \boldsymbol{\beta} \boldsymbol{\lambda}^T = 0. \end{aligned}$$

Используем

$$\frac{\partial \text{Tr}(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} = \mathbf{B} \qquad \frac{\partial \text{Tr}(\mathbf{A}\mathbf{B}\mathbf{A}^T)}{\partial \mathbf{A}} = 2\mathbf{B}\mathbf{A}^T \qquad \frac{\partial \text{Tr}(\mathbf{x}^T \mathbf{B}\mathbf{y})}{\partial \mathbf{A}} = \mathbf{y}\mathbf{x}^T.$$

Отсюда получим выражение для \mathbf{X} :

$$\mathbf{X} = \mathbf{Y} - \boldsymbol{\lambda} \boldsymbol{\beta}^T \Omega.$$

Если домножить это уравнение на $\boldsymbol{\beta}$ и вспомнить, что $\mathbf{X}\boldsymbol{\beta} = 0$, то

$$\mathbf{Y}\boldsymbol{\beta} = \boldsymbol{\lambda} \boldsymbol{\beta}^T \Omega \boldsymbol{\beta}, \quad \Rightarrow \quad \boldsymbol{\lambda} = \frac{1}{\boldsymbol{\beta}^T \Omega \boldsymbol{\beta}} \mathbf{Y}\boldsymbol{\beta}.$$

Подставим эти соотношения в функцию правдоподобия и получим концентрированную функцию правдоподобия (“концентрированный лагранжиан”):

$$\begin{aligned} \ell^c &= -\frac{1}{2} \text{Tr}(\boldsymbol{\lambda} \boldsymbol{\beta}^T \Omega \Omega^{-1} \Omega \boldsymbol{\beta} \boldsymbol{\lambda}^T) = -\frac{1}{2} \text{Tr}(\boldsymbol{\lambda}^T \boldsymbol{\lambda} \boldsymbol{\beta}^T \Omega \boldsymbol{\beta}) = -\frac{1}{2} \boldsymbol{\lambda}^T \boldsymbol{\lambda} \boldsymbol{\beta}^T \Omega \boldsymbol{\beta} = \\ &= -\frac{1}{2} \frac{\boldsymbol{\beta}^T \mathbf{Y}^T \mathbf{Y} \boldsymbol{\beta}}{\boldsymbol{\beta}^T \Omega \boldsymbol{\beta}} \end{aligned}$$

Таким образом, нахождение максимума правдоподобия равносильно минимизации следующей функции:

$$\varphi = \frac{\beta^T Y^T Y \beta}{\beta^T \Omega \beta} \rightarrow \min_{\beta}$$

Условие первого порядка для минимума:

$$\frac{\partial \varphi}{\partial \beta^T} = 2 \frac{1}{\beta^T \Omega \beta} Y^T Y \beta - 2 \frac{\beta^T Y^T Y \beta}{(\beta^T \Omega \beta)^2} \Omega \beta = 0$$

$$\Rightarrow (Y^T Y - \frac{\beta^T Y^T Y \beta}{\beta^T \Omega \beta} \Omega) \beta = 0$$

или $(\Omega^{-1} Y^T Y - \frac{\beta^T Y^T Y \beta}{\beta^T \Omega \beta}) \beta = (\Omega^{-1} Y^T Y - \varphi) \beta = 0.$

Таким образом, φ должно быть собственным числом матрицы $\Omega^{-1} Y^T Y$, а β — ее собственным вектором. Проверим, что эти два условия не противоречат друг другу. Пусть β_k — некоторый собственный вектор, φ_k — соответствующее собственное число этой матрицы:

$$(\Omega^{-1} Y^T Y - \varphi_k) \beta_k = 0.$$

Покажем, что $\varphi_k = \frac{\beta_k^T Y^T Y \beta_k}{\beta_k^T \Omega \beta_k}.$

Домножим слева на $\beta_k^T \Omega$:

$$\beta_k^T Y^T Y \beta_k - \varphi_k \beta_k^T \Omega \beta_k = 0.$$

Отсюда получаем требуемое равенство.

Поскольку требуется минимизировать φ , то нужно выбрать минимальное собственное число φ_{\min} . Оценкой вектора коэффициентов $\hat{\beta}$ будет соответствующий собственный вектор β_{\min} . Отсюда получаем оценки для матрицы исходных переменных X :

$$X = Y - \lambda \hat{\beta}^T \Omega = Y - \frac{1}{\hat{\beta}^T \Omega \hat{\beta}} Y \hat{\beta} \hat{\beta}^T \Omega = Y \left(I - \frac{1}{\hat{\beta}^T \Omega \hat{\beta}} \hat{\beta} \hat{\beta}^T \Omega \right).$$

В частном случае, когда ошибка имеется только в первой переменной

$$\Omega = \begin{bmatrix} \sigma^2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{O} \end{bmatrix}.$$

Нужно минимизировать величину

$$\varphi = \frac{\beta^T Y^T Y \beta}{\sigma^2 (\beta_1)^2} = (Y_1 + \frac{1}{\sigma \beta_1} Y_{(1)} \beta_{(1)})^T (Y_1 + \frac{1}{\sigma \beta_1} Y_{(1)} \beta_{(1)}).$$

где $Y_{(1)}$ — матрица Y без первого столбца, $\beta_{(1)}$ — вектор β без первого элемента. Если обозначить $\dot{Y} = Y_1$, $\dot{X} = Y_{(1)}$, $\dot{\beta} = \beta_{(1)}$ то получим ОМНК:

$$(\dot{Y} - \dot{X} \dot{\beta})^T (\dot{Y} - \dot{X} \dot{\beta}) \rightarrow \min$$

Внешне не связанные регрессионные уравнения

Пусть есть $k=1, \dots, K$ регрессионных уравнений и соответствующие выборки одинаковой длины $i=1, \dots, N$:

$$Y_k = X_k \beta_k + \varepsilon_k, \quad k = 1, \dots, K.$$

Предполагается, что коэффициенты β_k в разных уравнениях не связаны между собой какими-либо ограничениями, т.е. уравнения независимы с точки зрения коэффициентов, но существует связь с точки зрения ошибок наблюдений с одним и тем же номером i . Т.е. соответствующая корреляционная матрица не является диагональной, причем все элементы этой матрицы неизвестны. Ошибки, относящиеся к наблюдениям с разными номерами, считаются независимыми.

В экономике такое может быть если наблюдения относятся к одному и тому же периоду времени. В этом случае “одновременные” ошибки могут быть сильно коррелированы (как прибыль предприятий из-за экономического цикла). Еще один пример — регрессии, относящиеся к мужьям и женам.

Если обозначить

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_K \end{pmatrix} \quad X = \begin{bmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & X_K \end{bmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_K \end{pmatrix},$$

тогда модель можно переписать в более компактном виде:

$$Y = X\beta + \varepsilon.$$

Пусть остатки k -й регрессии равны $e_k(\beta_k) = Y_k - X_k \beta_k$.

Составим матрицу $E(\beta)$, столбцами которой являются e_k :

$$E_{ki} = Y_{ki} - X_{ki} \beta_{ki}.$$

Строки матрицы E будем обозначать E_i .

Предположение модели об ошибках состоит в том, что “одновременные” ошибки коррелированы, а “разновременные” — нет. Таким образом,

$$E(\varepsilon_{ki} \varepsilon_{li}) = \omega_{kl} \quad \text{и} \quad E(\varepsilon_{ki} \varepsilon_{ls}) = 0 \quad (i \neq s)$$

или, в матричной записи,

$$E(E_i(\beta_0) E_i(\beta_0)^T) = \Omega \quad \text{и} \quad E(E_i(\beta_0) E_s(\beta_0)^T) = 0 \quad (i \neq s).$$

Составим из ошибок вектор-столбец по другому:

$$\mathring{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{K1} \\ \vdots \\ \varepsilon_{1N} \\ \vdots \\ \varepsilon_{KN} \end{pmatrix} = \begin{pmatrix} \mathbf{E}_1^\top(\boldsymbol{\beta}_0) \\ \vdots \\ \mathbf{E}_N^\top(\boldsymbol{\beta}_0) \end{pmatrix},$$

где ошибки, относящиеся к одному и тому же наблюдению i стоят рядом.

Ковариационная матрица этого вектора ошибок является блочно-диагональной и равна

$$\mathring{V} = E(\mathring{\boldsymbol{\varepsilon}} \mathring{\boldsymbol{\varepsilon}}^\top) = \begin{bmatrix} \boldsymbol{\Omega} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \boldsymbol{\Omega} \end{bmatrix}.$$

Обратная к ней тоже блочно-диагональная и равна

$$\mathring{V}^{-1} = \begin{bmatrix} \boldsymbol{\Omega}^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}^{-1} & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \boldsymbol{\Omega}^{-1} \end{bmatrix}.$$

Если использовать символ Кронекера, то можно записать ковариационную матрицу не переставляя наблюдения:

$$V = E(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}) = \boldsymbol{\Omega} \otimes \mathbf{I}_K$$

(в этих обозначениях $\mathring{V} = \mathbf{I}_K \otimes \boldsymbol{\Omega}$), где \mathbf{I}_K — единичная матрица $K \times K$. Аналогично

$$V^{-1} = \boldsymbol{\Omega}^{-1} \otimes \mathbf{I}_K.$$

Чтобы оценить модель, мы должны указать распределение ошибок. Предполагаем, как обычно, что ошибки имеют нормальное распределение.

$\boldsymbol{\beta}$ и $\boldsymbol{\Omega}$ — неизвестные параметры модели.

Так же как и из ошибок, составим один вектор-столбец и из остатков:

$$\mathbf{e}(\boldsymbol{\beta}) = \begin{pmatrix} e_1(\boldsymbol{\beta}) \\ \vdots \\ e_K(\boldsymbol{\beta}) \end{pmatrix}, \quad \mathring{e}(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{E}_1^\top(\boldsymbol{\beta}) \\ \vdots \\ \mathbf{E}_N^\top(\boldsymbol{\beta}) \end{pmatrix}$$

Логарифмическая функция правдоподобия равна

$$\ell = -\frac{1}{2} \ln |V| - \frac{1}{2} \mathbf{e}^\top V^{-1} \mathbf{e} + \text{const} =$$

$$\begin{aligned}
&= -\frac{1}{2} \ln |\mathring{V}| - \frac{1}{2} \mathring{e}^\top \mathring{V}^{-1} \mathring{e} + \text{const} \rightarrow \max_{\beta} \\
\ell &= -\frac{N}{2} \ln |\mathbf{\Omega}| - \frac{1}{2} \sum_{i=1}^N \mathbf{E}_i \mathbf{\Omega}^{-1} \mathbf{E}_i^\top + \text{const} = \\
&= -\frac{N}{2} \ln |\mathbf{\Omega}| - \frac{1}{2} \text{Tr}(\mathbf{E} \mathbf{\Omega}^{-1} \mathbf{E}^\top) + \text{const}.
\end{aligned}$$

В максимуме правдоподобия производная по $\mathbf{\Omega}^{-1}$ равна нулю:

$$\frac{\partial \ell}{\partial \mathbf{\Omega}^{-1}} = \frac{N}{2} \mathbf{\Omega} - \frac{1}{2} \mathbf{E}^\top \mathbf{E} = 0.$$

Откуда получаем $\mathbf{\Omega}(\beta) = \frac{1}{N} \mathbf{E}(\beta)^\top \mathbf{E}(\beta)$.

При дифференцировании мы использовали правила

$$\frac{\partial \ln |\mathbf{A}|}{\partial \mathbf{A}^{-1}} = \mathbf{A}^\top \quad \text{и} \quad \frac{\partial \text{Tr}(\mathbf{ABC})}{\partial \mathbf{B}} = \mathbf{CA}.$$

Подставим $\mathbf{\Omega}(\beta)$ в $\ell(\cdot)$, чтобы получить концентрированную функцию правдоподобия:

$$\begin{aligned}
\ell^c &= -\frac{N}{2} \ln |\mathbf{\Omega}(\beta)| - \frac{1}{2} \text{Tr}(\mathbf{E} \mathbf{\Omega}(\beta)^{-1} \mathbf{E}^\top) = -\frac{N}{2} \ln |\mathbf{\Omega}(\beta)| - \frac{1}{2} \text{Tr}(\mathbf{E}^\top \mathbf{E} \mathbf{\Omega}(\beta)^{-1}) = \\
&= -\frac{N}{2} \ln |\mathbf{\Omega}(\beta)| - \text{Tr}(\mathbf{I}_K) = -\frac{N}{2} \ln |\mathbf{\Omega}(\beta)| - \frac{NK}{2}.
\end{aligned}$$

Задача $\ell^c \rightarrow \max_{\beta}$ эквивалентна задаче

$$\begin{aligned}
|\mathbf{\Omega}(\beta)| &\rightarrow \min_{\beta} \quad \text{или} \\
|\mathbf{E}(\beta)^\top \mathbf{E}(\beta)| &\rightarrow \min_{\beta}.
\end{aligned}$$

Выражение $|\mathbf{E}(\beta)^\top \mathbf{E}(\beta)|$ получило название **обобщенной дисперсии**.

Найдем условия первого порядка максимума концентрированной функции правдоподобия.

$$\frac{\partial}{\partial \beta_{kj}} \ln |\mathbf{\Omega}(\beta)| = \text{Tr}(\mathbf{\Omega}(\beta)^{-1} \frac{\partial}{\partial \beta_{kj}} \mathbf{\Omega}(\beta)).$$

Здесь мы используем, что

$$\frac{ds(\mathbf{A})}{dt} = \text{Tr}\left(\frac{\partial s}{\partial \mathbf{A}} \frac{d\mathbf{A}}{dt}\right),$$

где $s(\mathbf{A})$ — скалярная функция от матрицы \mathbf{A} , а также что

$$\frac{\partial \ln|A|}{\partial A} = (A^{-1})^T.$$

$$\begin{aligned} \frac{\partial}{\partial \beta_{kj}} \Omega(\beta) &= \frac{\partial}{\partial \beta_{kj}} \left(\frac{1}{N} E^T E \right) = \frac{2}{N} E^T \frac{\partial E}{\partial \beta_{kj}} = -\frac{2}{N} E^T (0, \dots, 0, X_{kj}, 0, \dots, 0) = \\ &= -\frac{2}{N} (0, \dots, 0, E^T X_{kj}, 0, \dots, 0). \end{aligned}$$

Поэтому,

$$\text{Tr} \left(\Omega(\beta)^{-1} \frac{\partial}{\partial \beta_{kj}} \Omega(\beta) \right) = -\frac{2}{N} \Omega(\beta)_k^{-1} E^T X_{kj},$$

где $\Omega(\beta)_k^{-1}$ — k -я строка матрицы $\Omega(\beta)^{-1}$.

Отсюда

$$\frac{\partial}{\partial \beta_k} \ln|\Omega(\beta)| = -\frac{2}{N} \Omega(\beta)_k^{-1} E^T X_k = -2 (E^T E)_k^{-1} E^T X_k.$$

Получим условия первого порядка:

$$(E(\beta)^T E(\beta))_k^{-1} E(\beta)^T X_k = 0.$$

Эта система уравнений нелинейна относительно β . Один из возможных способов решения состоит в использовании последовательности вспомогательных регрессий. Он основан на том, что эти уравнения для оценок МП $\hat{\beta}$ (хотя показать это технически сложно) эквивалентны уравнениям

$$\hat{\beta}_k = (X_k^{*T} X_k^*)^{-1} X_k^{*T} Y_k,$$

где $X_k^* = (X_k, E(\hat{\beta})_{-k})$, $E(\hat{\beta})_{-k}$ — матрица остатков всех уравнений, кроме k -го.

То есть в каждую регрессию надо добавить остатки из других регрессий.

Оценки вычисляются итеративно, начиная, например, с оценок ОМНК. Стандартные ошибки полученных в результате итераций уравнений надо скорректировать: вычислять не на основе суммы квадратов остатков из вспомогательной регрессии, а на основе суммы квадратов исходных остатков, т. е.

$$e_k(\hat{\beta})^T e_k(\hat{\beta}) = (Y_k - X_k \hat{\beta})^T (Y_k - X_k \hat{\beta}).$$

В частном случае, когда регрессоры во всех уравнениях одни и те же, оценки МП для коэффициентов β совпадут с оценками ОМНК. Различие будет только в оценке ковариационной матрицы ошибок Ω . Если в условии первого порядка

$$(E(\beta)^T E(\beta))_k^{-1} E(\beta)^T X_k = 0$$

взять $X_k = X \forall k$, то должно выполняться $(E(\beta)^T E(\beta))^{-1} E(\beta)^T X = 0$, откуда следует, что $E(\beta)^T X = 0$, то есть $e_k(\beta)^T X = 0$ — в каждом уравнении остатки ортогональны матрице регрессоров X . А это и есть условие минимума суммы квадратов в соответствующем уравнении.

Как и в любой модели обобщенного метода наименьших квадратов информационная матрица параметров β вычисляется по формуле

$$\mathcal{I}_{\beta\beta}^N = X^T V^{-1} X,$$

т.е.

$$\mathcal{I}_{\beta\beta}^N = X^T (\Omega^{-1} \otimes I_k) X.$$

Другой подход к вычислению оценок МП состоит в использовании итеративного обобщенного МНК.

$$\Omega^t = \frac{1}{N} E(\beta^t)^T E(\beta^t).$$

$$\beta^{t+1} = (X^T ((\Omega^t)^{-1} \otimes I_k) X)^{-1} X^T ((\Omega^t)^{-1} \otimes I_k) Y.$$

При использовании этого метода приходится иметь дело с матрицами большой размерности.

Системы одновременных уравнений

Пусть Y — матрица $(N \times m)$ эндогенных переменных, X — матрица $(N \times k)$ экзогенных переменных. Предполагается, что они связаны между собой m уравнениями:

$$Y\Gamma = XB + E.$$

Здесь Γ — квадратная невырожденная матрица $(m \times m)$ и B $(k \times m)$ — матрицы коэффициентов.

Такое представление системы одновременных уравнений называют **структурной формой**. Коэффициенты здесь можно определить только с точностью до множителей. Если каждое уравнение пронормировать, то изменятся только неизвестные параметры в Γ , B и матрице ковариаций ошибок. Один из способов нормировки заключается в том, чтобы взять диагональные элементы матрицы Γ равными 1:

$$\Gamma_{ll} = 1, \quad l = 1, \dots, m.$$

Предполагается кроме того, что часть коэффициентов в матрицах Γ и B могут быть равны нулю. Если оставить в уравнениях только неизвестные ненулевые коэффициенты, то можно переписать их в виде

$$Y_l = Y_{-l} \gamma_l + X_l \beta_l + \varepsilon_l, \quad l = 1, \dots, m.$$

В каждом уравнении в левой части остается только одна эндогенная переменная, имеющая тот же номер, что и уравнение. Остальные эндогенные переменные с ненулевыми коэффициентами перенесены в правую часть. Y_{-i} — составленная из них матрица. X_i — это экзогенные переменные с ненулевыми коэффициентами.

Систему одновременных уравнений можно записать также в приведенной форме, где каждая эндогенная переменная представлена как функция только экзогенных переменных:

$$Y = X\Pi + U.$$

Структурной форме соответствует ограниченная приведенная форма, называемая так потому, что на коэффициенты накладываются ограничения $\Pi = B\Gamma^{-1}$, так что:

$$Y = X\Pi + U = XB\Gamma^{-1} + E\Gamma^{-1}.$$

Если ограничения не учитываются, то имеем **неограниченную приведенную форму**. Неограниченная приведенная форма представляет собой систему внешне не связанных уравнений с одной и той же матрицей регрессоров во всех уравнениях. Таким образом, коэффициенты в ней можно оценить с помощью ОМНК.

Рассмотрим два подхода к оцениванию системы одновременных уравнений методом максимального правдоподобия.

FIML

Метод максимального правдоподобия, использующий полную информацию, (full information maximum likelihood) называется так, потому что он использует всю информацию об ограничениях, в том числе информацию о том, что часть коэффициентов равна нулю и $\Pi = B\Gamma^{-1}$.

Для применения ММП предположим, что $E_i \sim \text{NID}(\mathbf{0}, \Omega)$, то есть ошибки имеют нормальное распределение; ошибки, относящиеся к одному и тому же номеру наблюдения i коррелированы с матрицей ковариаций Ω , а относящиеся к наблюдениям с разными номерами — некоррелированы.

Матрицы Γ , B , Ω нужно оценить. Функция правдоподобия для i -го наблюдения равна

$$\ell_i = -\frac{m}{2} \ln 2\pi - \frac{1}{2} \ln |\Omega| - \frac{1}{2} E_i \Omega^{-1} E_i^\top.$$

Заменяем E_i на $Y_i \Gamma - X_i B$, при этом в функцию правдоподобия нужно добавить якобианный член, соответствующий преобразованию Y_i в E_i . Якобиан этого преобразования равен

$$\mathbf{J} = \frac{dY_i}{dE_i} = \Gamma.$$

Таким образом, якобианный член равен

$$\ln (\text{abs}|\Gamma|).$$

Просуммируем функции правдоподобия отдельных наблюдений:

$$\begin{aligned} \ell = \sum_i \ell_i = & -\frac{Nm}{2} \ln 2\pi + N \ln (\text{abs}|\Gamma|) - \frac{N}{2} \ln |\Omega| - \\ & -\frac{1}{2} \sum_i E_i(\Gamma, B) \Omega^{-1} E_i(\Gamma, B)^\top. \end{aligned}$$

Концентрируем функцию правдоподобия по Ω . В максимуме

$$\frac{\partial \ell}{\partial \Omega^{-1}} = \frac{N}{2} \Omega - \frac{1}{2} \sum_i E_i^\top E_i = \frac{N}{2} \Omega - \frac{1}{2} E^\top E = 0.$$

Отсюда имеем:

$$\Omega(\Gamma, B) = \frac{1}{N} (Y\Gamma - XB)^\top (Y\Gamma - XB).$$

Концентрированная функция правдоподобия равна

$$\ell^c = N \ln (\text{abs}|\Gamma|) - \frac{N}{2} \ln |(Y\Gamma - XB)^\top (Y\Gamma - XB)| + \text{const}.$$

Можно переписать ее в другом виде:

$$\begin{aligned} \ell^c = & -\frac{N}{2} \ln \left| \frac{1}{|\Gamma|^2} (Y\Gamma - XB)^\top (Y\Gamma - XB) \right| + \text{const} = \\ = & -\frac{N}{2} \ln |(Y - XB\Gamma^{-1})^\top (Y - XB\Gamma^{-1})| + \text{const}. \end{aligned}$$

Последнее выражение совпадает с концентрированной по матрице ковариаций функцией правдоподобия для ограниченной приведенной формы. Причина этого заключается в том, что структурная и ограниченная приведенная формы являются только различными способами записи одной и той же модели.

Методы оценивания систем одновременных уравнений довольно громоздки. Хорошие результаты дает применение метода Ньютона, но выражение для гессиана имеет довольно сложный вид.

Опишем здесь один из возможных методов, который имеет интуитивно понятную интерпретацию, и который несложно реализовать в виде компьютерной программы.

“Очищенные” от ошибок переменные Y можно определить как

$$\hat{Y} = XB\Gamma^{-1}.$$

Матрица соответствующих остатков

$$\hat{U} = Y - \hat{Y} = Y - XB\Gamma^{-1}.$$

Используя эти обозначения, уравнения системы

$$Y_l = Y_{-l} \gamma_l + X_l \beta_l + \varepsilon_l$$

можно переписать в виде

$$Y_l - \hat{U}_{-l} \gamma_l = \hat{Y}_{-l} \gamma_l + X_l \beta_l + \varepsilon_l,$$

где \hat{Y}_{-l} — это “очищенные” переменные Y_{-l} , и $\hat{U}_{-l} = Y_{-l} - \hat{Y}_{-l}$. Если рассматривать в этих уравнениях $\hat{U}_{-l} \gamma_l$ и \hat{Y}_{-l} как известные, то получаем систему внешне не связанных регрессионных уравнений. Все случайные компоненты как бы переносятся в левую часть регрессионных уравнений, так как если переменные \hat{Y}_{-l} вычисляются на основании состоятельных оценок параметров Γ и B , то они асимптотически некоррелированы с ошибками.

К этим уравнениям можно применить один из итеративных методов оценивания внешне не связанных регрессий. Величины $\hat{U}_{-l} \gamma_l$ и \hat{Y}_{-l} вычисляются на основании оценок параметров Γ и B , полученных на предыдущих итерациях. Как можно показать, этот алгоритм сходится к оценкам максимального правдоподобия.

После того, как получены оценки FIML, интересно сравнить структурную форму с неограниченной приведенной формой: $Y = X\Pi + U$.

Как уже говорилось, неограниченную форму можно оценить, применяя ОМНК к каждому уравнению, т. е., $\hat{\Pi} = (X^T X)^{-1} X^T Y$, только в качестве оценки ковариационной матрицы ошибок следует взять

$$\begin{aligned} \hat{\Omega} &= \frac{1}{N} U(\hat{\Pi})^T U(\hat{\Pi}) = \frac{1}{N} (Y - X\hat{\Pi})^T (Y - X\hat{\Pi}) \\ &= \frac{1}{N} Y^T (I - X(X^T X)^{-1} X^T) Y = \frac{1}{N} Y^T M_X Y. \end{aligned}$$

Логарифмическая функция правдоподобия в максимуме для ограниченной и неограниченной модели равна соответственно

$$\hat{\ell} = -\frac{N}{2} \ln |(Y - XB\hat{\Gamma}^{-1})^\top (Y - XB\hat{\Gamma}^{-1})| + \text{const}$$

$$\text{и } \tilde{\ell} = -\frac{N}{2} \ln |Y^\top M_X Y| + \text{const}.$$

Константа в обеих формулах одна и та же и равна $-\frac{Nm}{2} \ln 2\pi + \frac{N}{2} \ln N$.

Статистика отношения правдоподобия, равная $LR = 2(\tilde{\ell} - \hat{\ell})$ имеет распределение χ^2 с числом степеней свободы равным количеству ограничений. Этот тест называется **тестом на сверхидентифицирующие ограничения**.

Поскольку неограниченную форму оценивать легче, то имеет смысл использовать ее для проверки различных гипотез. Если ограничения выполнены, то оценивая неограниченную модель мы теряем в эффективности, но оценки все же будут состоятельными. Удобно проверять таким образом регрессию на автокорреляцию остатков, гетероскедастичность, функциональную форму.

LIML

Метод максимального правдоподобия, использующий ограниченную информацию, (limited information maximum likelihood) предназначен для оценивания одного уравнения из системы одновременных уравнений. Остальные уравнения оцениваются только в той степени, в какой это требуется для оценивания первого уравнения. Первое уравнение оценивается в структурной форме, а остальные — в неограниченной приведенной (тем самым, используется не вся имеющаяся информация):

$$Y_1 - Y_{-1} \gamma_1 = X_1 \beta_1 + \varepsilon_1,$$

$$Y_{-1} = X_1 B_1 + X_{-1} B_{-1} + E_{-1}.$$

Здесь Y_1 — “зависимая” переменная в первом уравнении, Y_{-1} — другие эндогенные переменные, входящие в первое уравнение, X_1 — экзогенные переменные, входящие в первое уравнение, X_{-1} — остальные экзогенные переменные системы.

Удобно рассматривать данную систему уравнений как структурную форму с ограничениями на матрица Γ и B .

Обозначим

$$Y = (Y_1, Y_{-1}), X = (X_1, X_{-1}), \gamma = \begin{bmatrix} 1 \\ -\gamma_1 \end{bmatrix}.$$

$$M_X = I - X(X^\top X)^{-1} X^\top, \quad M_1 = I - X_1(X_1^\top X_1)^{-1} X_1^\top.$$

Задача нахождения оценок максимального правдоподобия сводится к задаче нахождения **наименьшего дисперсионного отношения**:

$$\frac{\gamma^T Y^T M_X Y \gamma}{\gamma^T Y^T M_1 Y \gamma} \rightarrow \min_{\gamma}$$

Условие первого порядка минимума:

$$2 \frac{1}{\gamma^T Y^T M_1 Y \gamma} Y^T M_1 Y \gamma - 2 \frac{\gamma^T Y^T M_X Y \gamma}{(\gamma^T Y^T M_1 Y \gamma)^2} Y^T M_X Y \gamma = 0.$$

$$\Rightarrow \left(Y^T M_1 Y - \frac{\gamma^T Y^T M_X Y \gamma}{\gamma^T Y^T M_1 Y \gamma} Y^T M_X Y \right) \gamma = 0.$$

Отсюда следует, что $\frac{\gamma^T Y^T M_X Y \gamma}{\gamma^T Y^T M_1 Y \gamma}$ — собственное число матрицы $Y^T M_1 Y (Y^T M_X Y)^{-1}$, а γ — соответствующий собственный вектор. Поскольку удобно искать собственные числа симметричной матрицы, то лучше взять матрицу

$$(Y^T M_X Y)^{-1/2} Y^T M_1 Y (Y^T M_X Y)^{-1/2},$$

которая имеет те же собственные числа и вектора.

Таким образом, метод наименьшего дисперсионного отношения сводится к задаче отыскания минимального собственного числа вещественной симметричной матрицы.

Использованная литература

- Дрейпер, Н., Смит, Г. *Прикладной регрессионный анализ*: В 2-х кн. — М: Финансы и статистика, 1986.
- Песаран, М., Слейтер, Л. *Динамическая регрессия: теория и алгоритмы*. — М: Финансы и статистика, 1984.
- *Статистические методы в экспериментальной физике*. — М: Атомиздат, 1976.
- Amemiya, T. "Selection of Regressors," *International Economic Review*, **21** (1980), 331-354.
- Beggs, J.J. "Diagnostic Testing in Applied Econometrics," *Economic Record*, **64** (1988), 81-101.
- Bera A.K., C.M. Jarque, and L.-F. Lee. "Testing the Normality Assumption in Limited Dependent Variable Models," *International Economic review*, **25** (1984), 563-578.
- Bollerslev, T. "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, **31** (1986), 307-327.
- Cramer, J.S. *The Logit Model for Economists*. Adward Arnold, 1991.
- Dagenais, M.G. "The Computation of FIML Estimates as Iterative Generalized Least Squares Estimates in Linear and Nonlinear Simultaneous Equations Models," *Econometrica*, **46** (1978), 1351-1362.
- Davidson, R., and J.G. MacKinnon. "Convenient Specification Tests for Logit and Probit Models," *Journal of Econometrics*, **25** (1984), 241-262.
- Davidson, R., and J.G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, 1993.
- Enders, W. *Applied Econometric Time Series*. John Wiley & Sons inc., 1995.
- Engle, R.F. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, **50** (1982), 987-1007.
- Engle, R.F. "Band Spectrum Regression," *Economic Review*, **15** (1974), 1-11.

- Fiorentini, G., G. Calzolari, and L. Panattoni. “Analytical Derivatives and the Computation of GARCH Estimates,” *Journal of Applied Econometrics*, **11** (1996), 399-417.
- Godfrey, L.G. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge University Press, 1988.
- Granger, C.W.J., and P. Newbold. “Spurious Regressions in Econometrics,” *Journal of Econometrics*, **21** (1974), 111-120.
- Jarque, C.M., and A.K. Bera. “Efficient Tests for Normality, Homoskedasticity, and Serial Independence of Regression Residuals,” *Economic Letters*, **6** (1980), 255-259.
- Krdmer, W., and H. Sonnberger. *The Linear Regression Model Under Test*. Physica-Verlag, 1986.
- MacKinnon, J.G., A.A. Haug, and L. Michelis. “Numerical Distribution Functions of Likelihood Ratio Tests for Cointegration” (discussion paper), 1996.
- Pagan, A.R., and D.F. Nicholls. “Estimating Predictions, Prediction Errors and Their Standard Deviations Using Constructed Variables,” *Journal of Econometrics*, **24** (1984), 293-310.
- Pesaran, M.H., and B. Pesaran. *Microfit 3.0. An Interactive Econometric Software Package (User Manual)*. Oxford University Press, 1991.
- Telser L.G. “Iterative Estimation of a Set of Linear Regression Equations,” *Journal of American Statistical Association*, **59**, 845-862.

Предметный указатель

- A**
ADF • 44
ADL • 33
AIC • 14
ARCH • 100
- B**
ВННН • 69; 71
- D**
DF • 41
- E**
ECM • 35
- G**
GARCH • 101
- J**
J-тест • 14
- M**
method of scoring • 71
- P**
P_E-тест • 15
- A**
авторегрессионная модель с
 распределенным лагом • 33
авторегрессия • 34; 36
 векторная • 49
адаптивных ожиданий модель • 34
аддитивная ошибка • 10
асимптотическая информационная
 матрица • 58
асимптотическая эквивалентность трех
 классических тестов • 76
асимптотические t- и F-статистики • 79
- B**
бинарная зависимая переменная • 23
Бокса-Кокса модель • 107
Бокса-Кокса преобразование • 11; 107
- B**
Вейбулла распределение • 29
векторная авторегрессия • 49
- взаимодействия члены • 11
 взвешенная регрессия • 88
 взрывной процесс • 37
 вкладов в градиент матрица • 65
 вклады отдельных наблюдений в
 функцию правдоподобия • 57
 внешне не связанные регрессионные
 уравнения • 114
 внешнее произведение градиента • 69
 внутренне линейная модель • 10
 вспомогательная регрессия • 63
 для модели Бокса-Кокса • 108
 выброс • 18
- Г**
Гаусса-Ньютона регрессия
 определение • 94
гессиян логарифмической функции
 правдоподобия • 57
 эмпирический • 69
гетероскедастичность • 9; 88
градиент логарифмической функции
 правдоподобия • 57
градиентные методы • 70
- Д**
Дики-Фуллера тест • 41
 дополненный • 44
динамическая регрессия • 33
долгосрочное стационарное состояние • 35
дрейф • 37
- E**
единичный корень • 37
- И**
инвариантность ММП • 62
интегрированный процесс • 38
информационная матрица

вычисление • 65
определение • 58
способы оценивания • 68
**информационный критерий Акаике • 14;
27**
искусственная регрессия • 63
для логита • 81
для пробита • 83
для пуассоновой регрессии • 85
для регрессии с мультипликативной
гетероскедастичностью • 93
метод ВННН (OPG) • 71
метод Гаусса-Ньютона • 94
исправления ошибок модель • 35
векторная • 50
истинное распределение • 56
истинный параметр • 56
Й
Йохансена метод • 49
К
качественная зависимая переменная • 22
квази-МП методы • 62; 70
ковариационная матрица оценок МП • 68
коинтеграции ранг • 49
коинтегрированные процессы • 46
коинтегрирующий вектор • 47
Койка преобразование • 32
концентрированная функция
правдоподобия • 72
Кронекера произведение • 116
Л
лаг • 29
лаговый оператор • 30
линейная по виду модель • 9
линейная по параметрам модель • 9
линейный тренд • 20
логарифмическая функция правдоподобия
• 56

логистическое распределение • 24
логит • 24
логлинейная форма модели • 13
ложная регрессия • 39
**локально эквивалентные альтернативы •
92**
М
максимального правдоподобия метод • 56
максимального правдоподобия оценки
альтернативное определение • 58
асимптотическое распределение • 68
вычисление • 70
определение • 56
максимального собственного числа
статистика • 51
множественный логит • 27
модель выбора • 23
мультипликативная гетероскедастичность
• 92
мультипликативная ошибка • 10
Н
наименьшего дисперсионного отношения
метод • 123
Ньютона метод • 71
О
обобщенная дисперсия • 117
П
**пирсоновское семейство распределений •
109**
полиномиальная форма модели • 11
полиномиальный лаг • 31
полулогарифмическая форма модели • 13
порождающий данные процесс • 56
приведенная форма системы
одновременных уравнений • 120
пробит • 25
производственная функция
Кобба-Дугласа • 10

с постоянной эластичностью замены • 10
транслоговая • 11
Прэйса-Винстена преобразование • 96
псевдо- R^2 для моделей с бинарной
зависимой переменной • 27
Пуассона распределение • 83
Р
распределенный лаг • 30
регрессия
пуассонова • 84
с МА-ошибкой • 98
с мультипликативной
гетероскедастичностью • 92
с AR-ошибкой • 94
С
сезонные переменные • 20
семейство распределений • 56
следа статистика • 51
случайное блуждание • 37
спектральный анализ • 22
стационарность • 36
тестирование • 41
стохастический тренд • 21; 40
структурная форма системы
одновременных уравнений • 119
Т
тест
Вальда • 75
Дики-Фуллера • 41
множителя Лагранжа • 74

множителя Лагранжа в градиентной форме
• 74
на авторегрессионную условную
гетероскедастичность • 105
на адекватность предсказаний • 19
на гетероскедастичность • 91
на гетероскедастичность • 9
на нормальность • 112
на сверхидентифицирующие ограничения
• 122
на стационарность • 41
отношения правдоподобия • 75
Чоу • 18
у
упорядоченный логит • 28
уравнения правдоподобия • 57
Ф
фиктивные переменные • 15
функциональная форма
неправильная спецификация • 8
функция правдоподобия • 56
Х
Холецкого разложение • 86
ч
частичного приспособления модель • 34
Э
Энгла-Грейнджера метод • 48
Я
якобиан преобразования плотности
распределения • 106