



Челябинский государственный университет

Т.Б. Бигильдеева, Е.А. Постников

ЭКОНОМЕТРИКА

**Регрессионный анализ с использованием
пакета Gretl**

Лабораторный практикум

Челябинск
2014

УДК 330.43
ББК У.в631я73
Б 597

Бигильдеева Т.Б.

Эконометрика. Регрессионный анализ с использованием пакета *Gretl*: Лабораторный практикум / Т.Б. Бигильдеева, Е.А. Постников. – Челябинск: Центр Научного Сотрудничества. 2014. – 80 с.

Лабораторный практикум по дисциплине «Эконометрика» с использованием программного пакета *Gretl* (v. 1.9.90) содержит теоретический материал, практические задания и указания к их выполнению, а также вопросы для проверки знаний по изучаемой теме.

Лабораторный практикум предназначен для бакалавров и студентов экономических направлений и специальностей, преподавателей и научных работников, а также для всех интересующихся вопросами эконометрических исследований.

Рецензенты: канд. экон. наук, доцент кафедры оценки бизнеса и конкурентоспособности ФГБОУ ВПО «ЮУрГУ» (НИУ) Л.А. Ширшикова

канд. пед. наук, зав. кафедрой математических, технических и естественнонаучных дисциплин Южно-Уральского института управления и экономики ЧОУ ВПО «ЮУИУиЭ» И.Ю. Коробейникова

ББК У.в631я73-5

© ФГБОУ ВПО «Челябинский государственный университет», 2014

© Бигильдеева Т.Б., Постников Е.А., 2014

ОГЛАВЛЕНИЕ

Лабораторная работа № 1. Основы работы с пакетом <i>Gretl</i> ...	4
Лабораторная работа № 2. Анализ данных и их подготовка к эконометрическому исследованию	20
Лабораторная работа № 3. Парная линейная регрессия	33
Лабораторная работа № 4. Множественная линейная регрессия	45
Лабораторная работа № 5. Нелинейные регрессионные модели	63
Лабораторная работа № 6. Выбор регрессионной модели	71
Список рекомендуемой литературы	78

Лабораторная работа № 1.

ОСНОВЫ РАБОТЫ С ПАКЕТОМ GRETЛ

GRETЛ (Gnu Regression, Econometrics and Time-series Library) – это кросс-платформенный программный пакет для эконометрического анализа, написанный на языке С. Является открытым, свободным и бесплатным программным обеспечением.

Скачать пакет Gretl и получить инструкции по его установке можно по адресу: <http://gretl.sourceforge.net/ru.html>.

Основными возможностями и особенностями пакета являются:

- простой и интуитивно понятный интерфейс на русском, английском и других языках;
- широкий набор методов оценивания как одного уравнения, так и систем уравнений: метод наименьших квадратов (LS), метод максимального правдоподобия (ML), обобщённый метод моментов (GMM);
- удобный инструментарий для анализа временных рядов: ARIMA, GARCH, VAR и VECM, тесты на единичные корни и коинтеграцию, фильтр Калмана и прочее;
- методы оценки моделей с ограниченными значениями зависимой переменной (логит-модели, пробит-модели, модели Тобина и Хекмана);
- выдача результатов в формате LaTeX в форме таблицы или уравнения.

Пакет Gretl поддерживает базы данных в следующих форматах: XML; CSV; листы Excel, Gnumeric и Open Document; файлы .dta из Stata; файлы .sav из SPSS; рабочие файлы Eviews и ряд других.

Запуск приложения осуществляется двойным щелчком мыши по иконке Gretl на рабочем столе. В результате появится следующее окно (рис. 1.1).

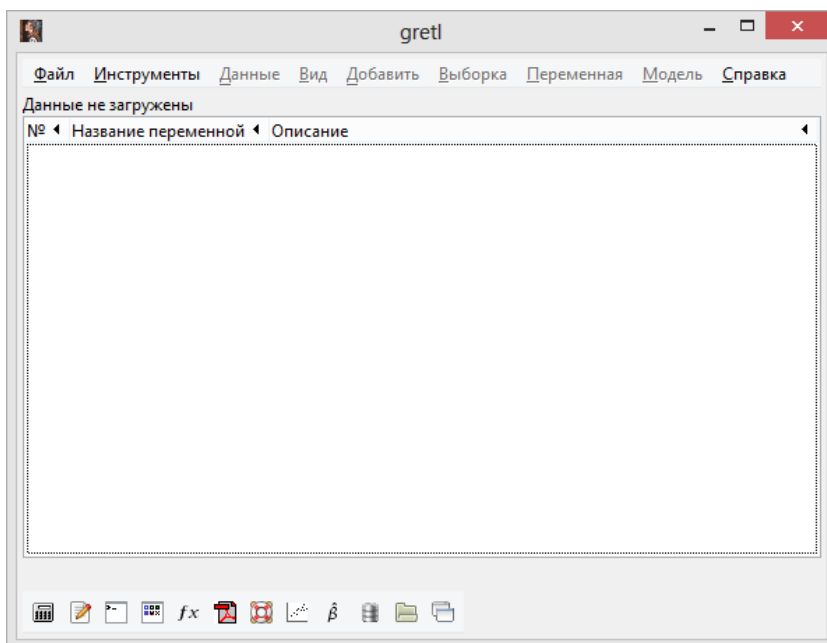


Рис. 1.1. Стартовое окно Gretl

Основными элементами интерфейса окна Gretl (рис. 1.1) являются главное меню, область переменных и набор иконок для некоторых полезных операций.

1. Создание, открытие и сохранение рабочего файла

Первым шагом при работе с приложением является создание нового или открытие существующего рабочего файла, в котором хранится вся информация.

Для **создания** нового *рабочего файла* необходимо выбрать в главном меню **Файл / Создать**, в появившемся окне указать максимальное количество наблюдений в выборке, нажать **ОК**, появится окно “Структура данных” (рис. 1.2).

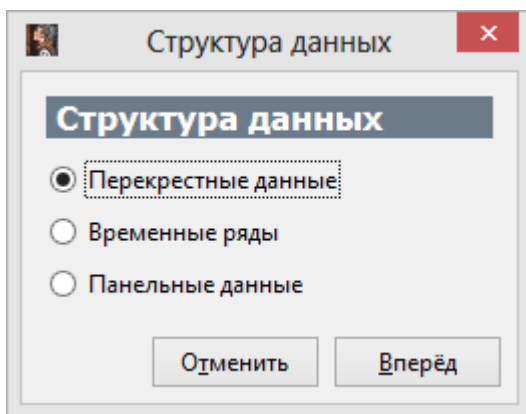


Рис. 1.2. Окно “Структура данных”

В окне “Структура данных” (рис. 1.2) выбирается тип данных.

Пространственными называются данные о совокупности объектов на определенный момент времени.

Временные данные (временной ряд) есть серия наблюдений об одном и том же объекте в последовательные моменты времени.

Панельные данные представляют собой совокупность наблюдений о нескольких объектах в разные моменты времени в течение некоторого периода.

При выборе **Перекрестных (пространственных) данных** необходимо задать размер (количество наблюдений) выборки, на которой будет проводиться исследование, например 100. Таким образом, будет создан файл с размером выборки, равным 100. В дальнейшем при необходимости размер выборки может быть уменьшен (часть наблюдений исключена из рассмотрения) или увеличен (добавлены новые наблюдения). Соответствующие команды будут описаны далее.

При выборе **Временных рядов** определяется периодичность данных – годовые, квартальные, ежемесячные, еженедельные и т.д.; далее указывается первый период, с которого начнется отсчет наблюдений и количество наблюдений.

Например, для **квартальных данных** (выборка из 100 наблюдений) – 1990:1 означает создание временного ряда кварталь-

ных данных на интервале от 1990:1 до 2014:4.

При выборе **Панельных данных** в поле «Количество кросс-секционных наблюдений» указывается число единиц совокупности, в поле «Количество временных периодов» – число периодов времени.

Открытие существующего рабочего файла осуществляется двойным щелчком мыши по файлу либо выбором в главном меню **Файл / Открыть / Пользовательские**.

В *Gretl* существуют встроенные примеры наборов данных, созданные разработчиками и исследователями. Для их открытия используется меню **Файл / Открыть / Примеры**, в которой выбирается на соответствующей вкладке имя открываемого файла двойным щелчком мыши. Например вкладка – *Dougherty*, файл – *EDUC* (рис. 1.3).

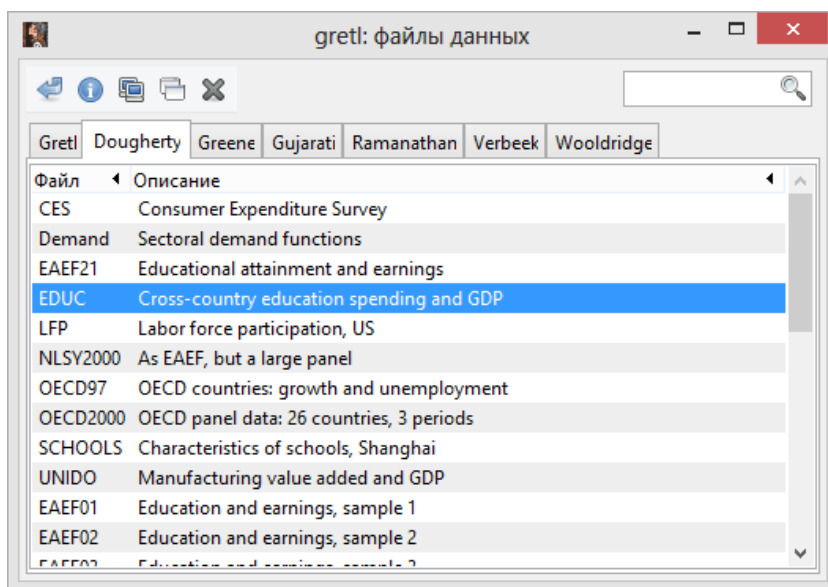


Рис. 1.3. Примеры файлов данных *Dougherty*

Исходные данные можно также **импортировать** из таблиц EXCEL или других популярных баз данных. Для этого необходимо выбрать в меню **Файл / Открыть / Пользовательские**.

Примечание. Для корректного импортирования данных все названия переменных, листов и файлов должны содержать только латинские буквы, цифры и подчеркивание. В файле не должно быть лишней информации: графиков, излишних пояснений, объединений ячеек и т.д. Удобно, если загружаемые в пакет данные расположены в вертикальных таблицах, где первая строка – это названия переменных.

После указания пути импортируемого файла и его открытия появится окно с вопросом о необходимости изменить структуру данных на временной ряд или панельные данные (рис. 1.4), при этом по умолчанию данные считаются пространственными.

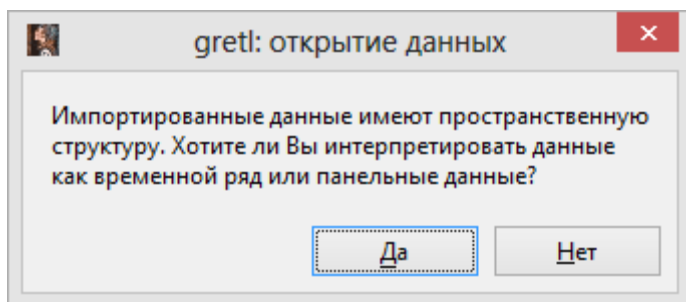


Рис. 1.4. Окно запроса структуры данных

В случае, если импортируемые данные являются пространственными (перекрестными), то необходимо нажать кнопку **Нет**.

Для **сохранения** рабочего файла необходимо выбрать в меню **Файл / Сохранить** или **Файл / Сохранить как....**. Сохраненный файл имеет расширение *.gdt.

2. Создание, просмотр и редактирование ряда данных

Ряды данных можно **создавать** несколькими способами.

1. Создание пустого ряда

Этот способ применяется для ручного ввода данных.

В этом случае для **создания ряда** необходимо выбрать в меню **Добавить / Добавить новую переменную....**; в появившемся окне ввести имя переменной, например X1 (рис. 1.5).

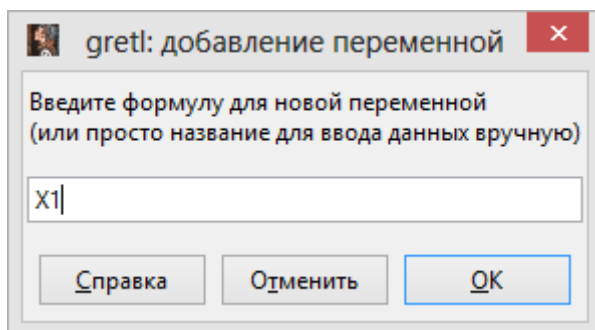


Рис. 1.5. Окно “Добавление переменной”

2. Создание ряда с помощью команды

В этом случае для создания ряда необходимо выбрать в меню **Добавить / Добавить новую переменную...** В появившемся окне вводится команда **genr имя_ряда=команда**, и нажать ОК.

Например, **genr y=x1+10*x2** (при условии, что уже созданы ряды **x1** и **x2**). В этом случае каждое значение ряда **y** будет равно сумме соответствующих значений ряда **x1** и значений ряда **x2**, умноженных на 10 (рис. 1.6).

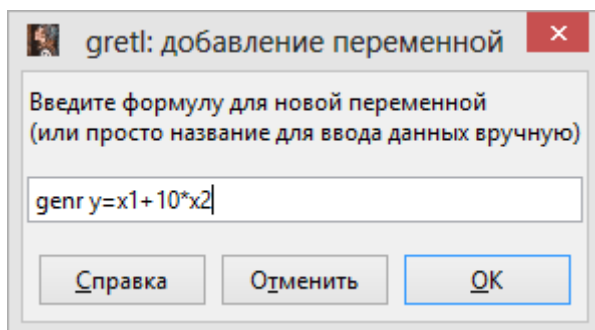


Рис. 1.6. Создание нового ряда данных с помощью команды

3. Генерирование случайных переменных

В ряде случаев требуется использование рядов данных, подчиняющихся какому-либо закону распределения (равномерному, нормальному, Стьюдента, Хи-квадрат, Фишера, Пуассона и др.).

В этом случае для создания ряда необходимо выбрать в ме-

ню **Добавить / Случайную переменную...**, в появившемся окне выбрать соответствующую вкладку вида распределения, указать параметры распределения и имя переменной (рис. 1.7).

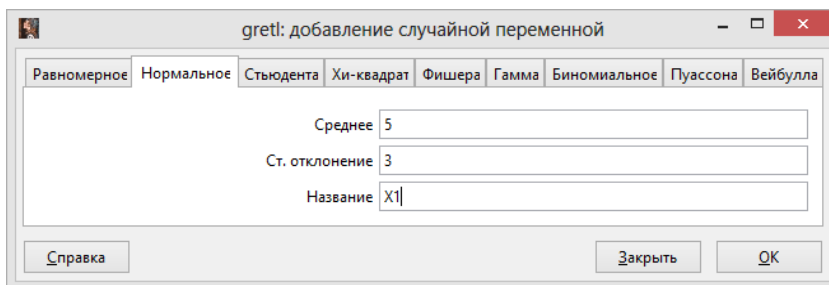



Рис. 1.7. Пример создания нормально распределенной случайной выборки

4. Создание индексной переменной

Индексная переменная – переменная, состоящая из значений номеров наблюдений 1, 2, 3, ..., n. Для ее создания необходимо выбрать в меню **Добавить / Индексную переменную**, после чего в области переменных появится переменная **index**.

Просмотр ряда осуществляется двойным щелчком мыши по его названию; либо щелчком правой кнопкой мыши по переменной, вызовом контекстного меню и выбором пункта **Показать значения**; либо выделением переменной и выбором в меню **Данные / Показать значения**.

Для **редактирования** ряда данных необходимо в режиме просмотра в окне этого ряда нажать кнопку  (рис. 1.8), либо правой кнопкой мыши вызвать контекстное меню и выбрать пункт **Изменить значения**, либо выбрать в меню **Данные / Изменить значения**.

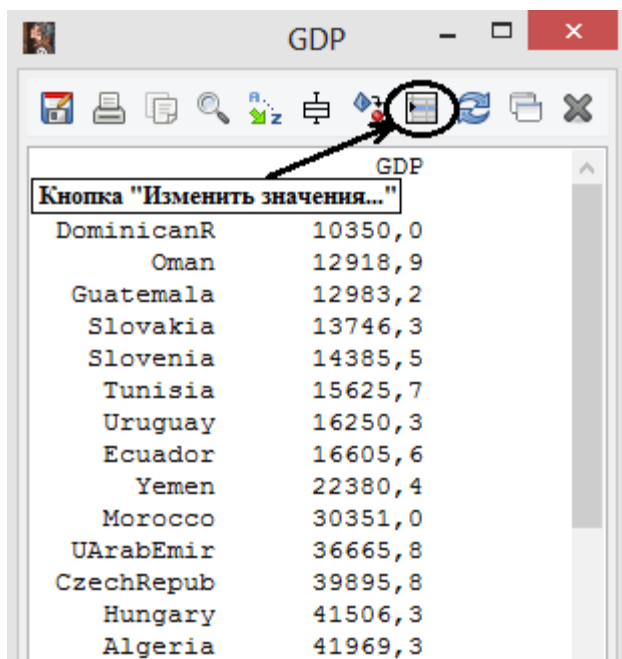


Рис. 1.8. Редактирование ряда данных

3. Удаление рядов данных

Для удаления переменной из рабочего файла нужно выделить ее в окне рабочего файла и выполнить любое из следующих действий

- нажать кнопку *Delete* на клавиатуре,
 - щелкнуть правой кнопкой мышки по переменной и выбрать в контекстном меню пункт *Удалить*,
- после чего нажать клавишу *Да*.

4. Работа с группой переменных

Для более удобной работы с данными переменные можно объединять в группу, в которой отображается содержимое двух или более рядов данных.

Для этого необходимо выделить нужные переменные с по-

мощью клавиши *Ctrl* и левой кнопки мыши, затем правой кнопкой мыши вызвать контекстное меню и выбрать соответствующую операцию (Показать значения, Изменить значения, Удалить, График разброса X-Y и др.).

5. Окно иконок текущей сессии

Для облегчения работы с *Gretl*, генерирующей большое количество окон с результатами, существует возможность работать в окне иконок текущей сессии. Все создаваемые объекты (графики, модели, тесты, скаляры) можно сохранять и затем просматривать и/или редактировать в окне «Значки» (рис. 1.9).

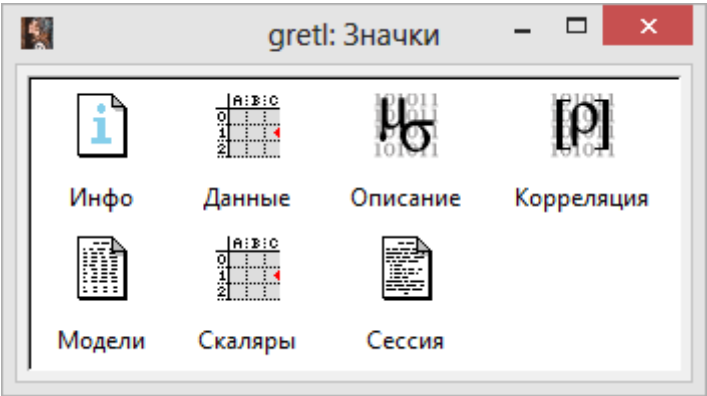










Рис. 1.9. Окно иконок текущей сессии

Для открытия окна иконок текущей сессии необходимо выбрать в меню **Вид / Сессия** либо нажать внизу окна программы иконку .

Иконки в окне текущей сессии отвечают за следующие функции.

 Инфо	открывает окно информации о рабочем файле
 Данные	открывает окно редактирования всех переменных

 Описание	открывает окно выборочных статистических характеристик всех переменных
 Корреляция	открывает окно матрицы коэффициентов корреляции всех переменных
 Модели	открывает окно табличного представления нескольких моделей с одинаковой зависимой переменной
 Скаляры	открывает окно созданных констант
 Сессия	открывает окно с инструкциями Gretl, реализованными в текущей сессии

Использование иконок существенно ускоряет работу в пакете Gretl.

6. Изменение диапазона рабочей выборки

Иногда необходимо работать (выполнять различные операции) с частью выборки (с выборкой меньшего размера, чем тот размер, который задан первоначально при создании рабочего файла). Например, максимальный размер выборки 1000 (заданный при создании рабочего файла), а далее требуется работать с частью выборки, а именно: с данными, имеющие номера от 1 до 10. В этом случае необходимо изменить рабочую выборку.

Для этого надо выбрать в меню пакета **Выборка / Установить диапазон....** Появится окно “Установить диапазон”, в котором в полях «Начало» и «Конец» надо указать требуемый соответственно 1-й и последний номера наблюдений рабочей выборки (рис. 1.10).

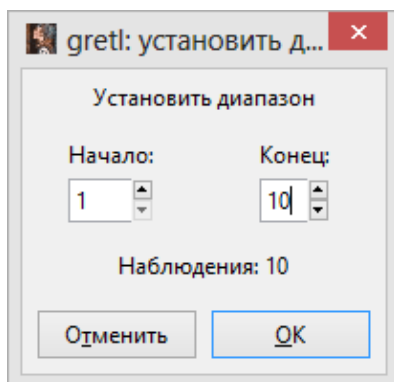


Рис. 1.10. Окно для изменения размера выборки

В окне рис. 1.10, согласно примеру, в поле «Начало» необходимо ввести 1, в поле «Конец» – 10 и нажать *ОК*. После этого внизу главного окна “Gretl” появившаяся надпись

“Перекрестные данные: Полный диапазон 1 – 1000”

изменится на

“Перекрестные данные: Полный диапазон 1–1000; выборка 1–10”

В данном случае поле «**Полный диапазон**» указывает размер всей выборки (от 1 до 1000) с максимальным количеством наблюдений 1000. В поле «**выборка**» указана рабочая выборка – та часть данных, над которой далее будут выполняться различные операции, это данные с номерами от 1 до 10, включающие 10 наблюдений.

Изменение рабочей выборки не приводит к преобразованию (потери) данных всей выборки, но далее при выполнении операций именно данные рабочей выборки будут предметом исследования и могут быть преобразованы.

Примечание: перед выполнением операций над рядами данных необходимо убедиться, что рабочая выборка установлена правильно.

7. Сортировка данных

Чтобы упорядочить в рабочем файле все данные по возрастанию по какому-либо из рядов, например *GDP*, необходимо вы-

брать в меню **Данные / Сортировать данные...** и выбрать переменную, по которой необходимо провести сортировку (рис. 1.11). В результате этого данные ряда *GDP* упорядочатся по возрастанию, а данные в других рядах упорядочатся относительно ряда *GDP*.

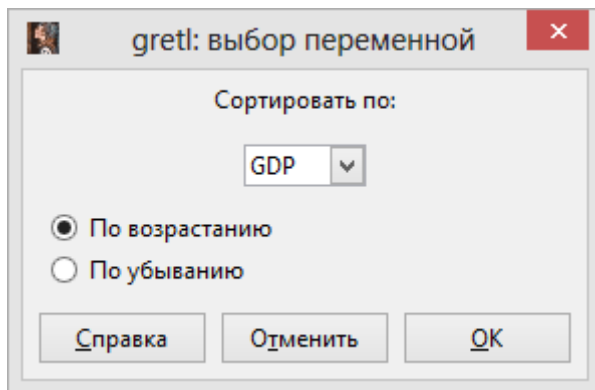



Рис. 1.11. Окно сортировки данных

Примечание: Для корректной работы с отсортированными данными (например, при построении графиков) необходимо заново создать индексную переменную **index**.

8. Создание и сохранение графиков

Для создания графика одной переменной необходимо сначала создать индексную переменную, затем нажать внизу окна программы иконку , либо выбрать в меню **Вид / График / Разброс X-Y**. В появившемся окне в поле «Ось X» перенести переменную **index**, а в поле «Ось Y» – переменную, для которой строится график, например **GDP**, затем нажать кнопку ОК (рис. 1.12).

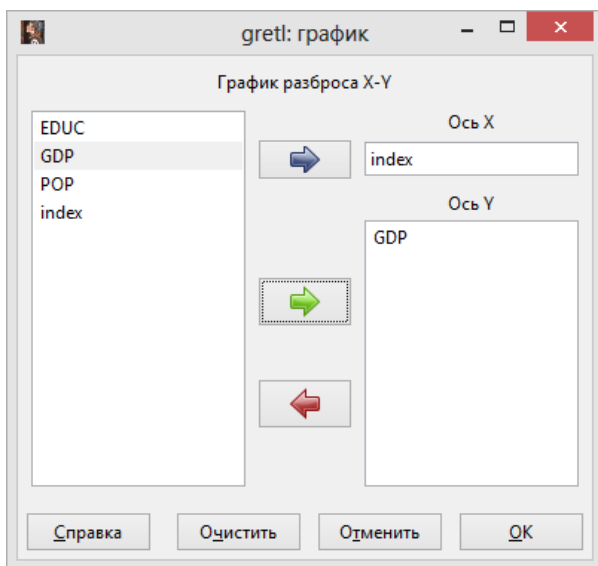


Рис. 1.12. Окно создания графика

В результате появится график переменной (маркеры на плоскости) (рис. 1.13).

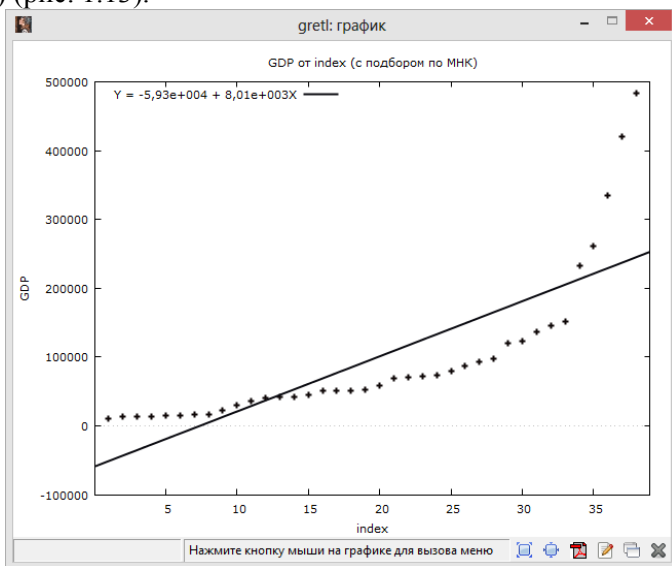


Рис. 1.13. График ряда данных

График зависимости одной переменной от другой строится аналогично графику одной переменной с одним отличием: в поле «Ось X» переносится переменная, зависимость от которой рассматривается. Например, на рис. 1.14 представлен график зависимости переменной **GDP** от переменной **EDUC**.

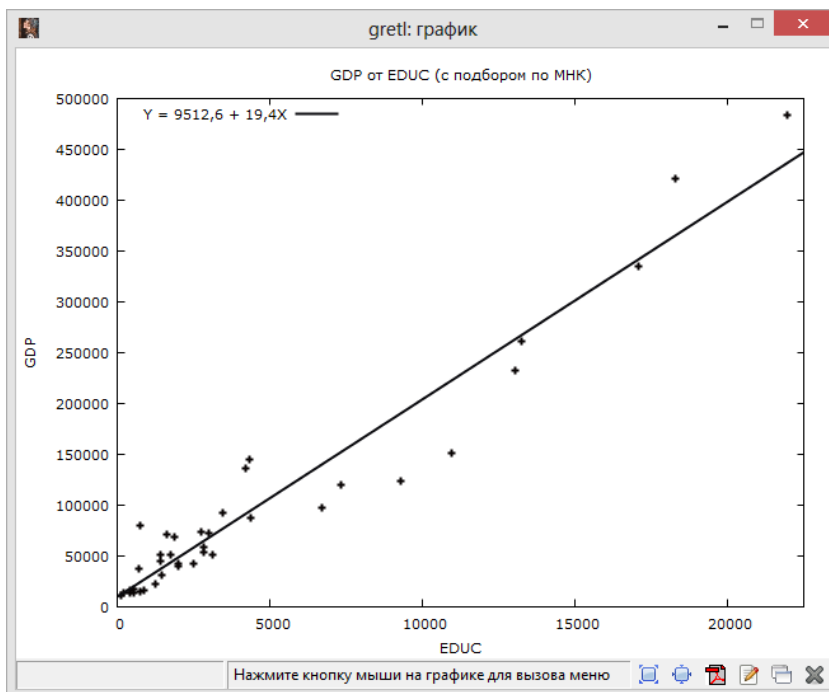


Рис. 1.14. Создание графика зависимости рядов

Замечание. На рисунках 1.13 и 1.14 помимо графиков переменных также отображаются линия и уравнение регрессии, которые подробнее будут рассмотрены в последующих лабораторных работах.

Для *сохранения* графика необходимо щелкнуть правой кнопкой мыши на самом графике и в появившемся контекстном меню выбрать «Сохранить в текущей сессии». В результате этого появится окно текущей сессии с иконкой **График**.

Задания

1. Создайте рабочий файл для пространственных данных с максимально возможным размером выборки 250 наблюдений.
2. Создайте ряды $X1$ и $X2$ размером (количеством) 250 и 100 наблюдений соответственно, представляющие собой равномерно распределенные на интервале $[5,10]$ случайные числа.
3. Создайте распределенные по стандартному нормальному закону ряды случайных чисел $Y1$ и $Y2$ размером 250 и 100 наблюдений соответственно.
4. Добавьте в отчет копию окна просмотра группы переменных $X1, X2, Y1, Y2$.
5. Создайте ряд Z , размером 200, равный сумме рядов $X1$ и $Y1$.
6. Создайте ряд R , размером 250, равный сумме рядов $X1$ и $X2$. проанализируйте полученный результат.
7. Создайте ряд T , размером 200, равномерно распределенных на интервале $[50,100]$ случайных чисел.
8. Создайте ряд случайных чисел H , размером 150, распределенных по закону $N(1,9)$.
9. Отсортируйте данные для всей выборки (от 1 до 250) по возрастанию ряда $X1$. Постройте и сохраните в текущей сессии графики рядов $X1$ и $Y1$, а также график зависимости $Y1$ от $X1$.
10. Добавьте следующую информацию о рабочем файле в текущей сессии: номер и название лабораторной работы, номер группы и ФИО студента.
11. Сохраните файл под именем *Фамилия студента_1.gdt*.

Вопросы для самоконтроля

1. Дайте определение эконометрики.
2. Каковы основные задачи эконометрического исследования?
3. Что такое эконометрическая модель?
4. Какие связи исследуются при эконометрическом моделировании? В чем их особенность?
5. Назовите этапы эконометрического исследования.
6. В чем суть первого этапа эконометрического моделирования– постановки задачи?
7. На какие вопросы необходимо ответить при выполнении

второго этапа эконометрического моделирования – анализа предметной области?

8. В чем суть выбора спецификации эконометрической модели?
9. Почему этап выбора спецификации модели необходимо выполнять до этапа сбора данных?
10. Что означает идентификация эконометрической модели?
11. Каково назначение этапа верификации эконометрической модели?
12. В чем суть этапа интерпретации результатов эконометрического моделирования?
13. Какие типы данных рассматриваются в эконометрике?
14. Что представляют собой пространственные данные?
15. Что такое временной ряд?
16. Что собой представляют панельные данные?
17. Каковы основные типы моделей в эконометрике?
18. Какие переменные в эконометрической модели называются эндогенными? Как иначе их можно назвать?
19. Какие переменные в эконометрической модели называются экзогенными? Как иначе их можно назвать?
20. Дайте определение случайной величины.
21. Что такое закон распределения случайной величины?
22. Дайте определение функции распределения случайной величины.
23. Дайте определение плотности распределения случайной величины.
24. Что показывает математическое ожидание случайной величины?
25. Что показывает дисперсия случайной величины?
26. В каких единицах измеряется среднеквадратическое отклонение случайной величины? Что оно показывает?
27. Какое распределение называется равномерным?
28. Что означает запись $X \sim N(1,9)$? $X \sim N(0,1)$?
29. Что такое *Индексная переменная* в Gretl?
30. Что означают числа 50 и 100 при создании ряда T (задание 7)?
31. Что означают числа 1 и 9 при создании ряда H (задание 8)?
32. Откройте ряд $X2$ и объясните, почему после сортировки в ряду числовые данные стали чередоваться с пустыми наблю-

дениями.

Лабораторная работа № 2.

АНАЛИЗ ДАННЫХ И ИХ ПОДГОТОВКА К ЭКОНОМЕТРИЧЕСКОМУ ИССЛЕДОВАНИЮ

Целью эконометрического исследования является определение модели, описывающей взаимосвязь двух и более переменных, или проверка гипотез относительно коэффициентов модели. Для получения качественных результатов исследования перед построением модели необходима тщательная подготовка и анализ данных.

Анализ данных достаточно часто позволяет сократить количество «лишней» работы, связанной с построением модели.

Предварительный анализ данных можно условно разделить на три этапа:

- 1) графический анализ данных;
- 2) анализ выборочных характеристик рассматриваемых рядов;
- 3) фильтрация (очистка) рядов данных.

На практике данные этапы не всегда выполняются в указанной последовательности. При работе с реальными данными они повторяются, меняются местами, в силу того, что каждое исследование требует своей схемы обработки информации. Эконометрическое исследование проводится как минимум для двух рядов данных, поэтому при подготовке и анализе данных они рассматриваются как по отдельности, так и совместно.

В качестве примера рассмотрим данные 38 стран по уровню ВВП (млн. долл.), расходам на образование (млн. долл.) и численности населения (тыс. чел.) в 1997 году. Требуется изучить влияние расходов на образование и численности населения страны на уровень ВВП.

1. Создание рабочего файла и создание рядов данных

Перед началом анализа необходимо собранные данные импортировать в *Gretl* (см. п. 1 лабораторной работы №1).

Каждое наблюдение относится к какой-либо стране, поэтому

желательно зафиксировать данную привязку, чтобы интерпретация выводов подразумевала некоторую страну. Для этого в файле *Excel*, откуда будет осуществляться импорт данных, необходимо создать столбец с наименованием стран на латинском языке (рис. 2.1).

	A	B	C	D	E
1	NAMES	EDUC	GDP	POP	
2	DominicanR	122,7963562	10350	7684	
3	Oman	397,9193726	12918,85547	2116	
4	Guatemala	194,2203369	12983,20313	10322	
5	Slovakia	535,2473145	13746,29395	5325	
6	Slovenia	719,8353882	14385,5293	1925	
7	Tunisia	854,0925293	15625,74121	8820	
8	Uruguay	392,0578308	16250,27246	3168	
9	Ecuador	519,3915405	16605,5918	11221	
10	Yemen	1213,738525	22380,43359	14329	
11	Morocco	1452,461182	30350,97266	26025	
12	UArabEmir	700,354126	36665,75781	2157	

Рис. 2.1. Исходные данные для импорта из файла *Excel*

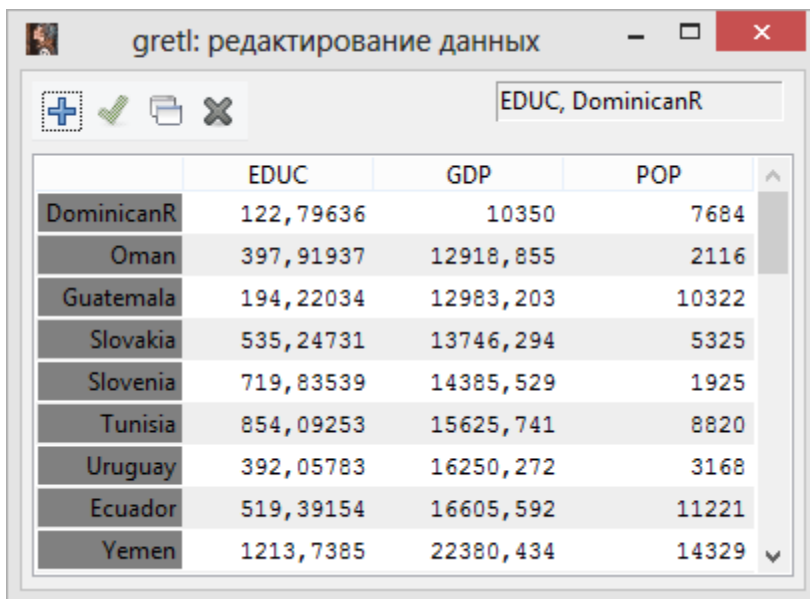
Данную базу также можно импортировать из встроенных примеров наборов данных *Gretl* (см. п. 1 лаб. раб. №1, рис. 1.3) вкладка – *Dougherty*, файл – *EDUC*.

После импорта ряды данных в *Gretl* будут выглядеть следующим образом (рис. 2.2). Для просмотра данных откройте в текущей сессии иконку «Данные». Обозначения переменных следующие:

EDUC – расходы на образование (млн. долл.);

GDP – уровень ВВП (млн. долл.);

POP – численность населения (тыс. чел.).



	EDUC	GDP	POP
DominicanR	122,79636	10350	7684
Oman	397,91937	12918,855	2116
Guatemala	194,22034	12983,203	10322
Slovakia	535,24731	13746,294	5325
Slovenia	719,83539	14385,529	1925
Tunisia	854,09253	15625,741	8820
Uruguay	392,05783	16250,272	3168
Ecuador	519,39154	16605,592	11221
Yemen	1213,7385	22380,434	14329

Рис. 2.2. Исходные данные после импорта в Gretl

2. Графический анализ данных

Однородность данных в статистическом анализе играет важную роль, так как она напрямую влияет на точность рассчитываемых показателей и качество аналитических выводов. Чем однороднее данные, тем надежнее и адекватнее реальным результатам результаты статистического анализа.

Под **однородными данными** обычно понимают данные с относительно невысоким уровнем рассеяния (разброса), при котором рассчитываемые статистические показатели будут давать надежную и качественную характеристику анализируемой совокупности.

Основным мериллом разброса (и однородности) данных являются показатели вариации: дисперсия, среднее квадратическое (стандартное) отклонение, а также коэффициент вариации V , который представляет собой отношение стандартного отклонения

σ к среднему значению исследуемой случайной величины \bar{X} :

$$V = \frac{\sigma}{\bar{X}} \cdot 100\% .$$

Выборка обычно считается однородной, если $V < 33\%$.

На практике достаточно часто исходные данные содержат ряд так называемых «аномальных» наблюдений, которые делают выборку неоднородной и способны существенно ухудшить качество эконометрической модели. Их появление может быть вызвано следующими причинами.

Во-первых, выборка может оказаться неоднородной из-за того, что при подготовке данных был проведен недостаточно глубокий их качественный анализ с точки зрения целей исследования. Так, например, при проведении эконометрического анализа факторов, влияющих на экономический рост, были объединены в одну выборку высокоразвитые страны и страны с низким уровнем развития (что нежелательно). Если это сделать необходимо (например, надо выявить общие для всех стран факторы экономического роста), то надо использовать специальные приемы, учитывающие неоднородность выборки.

Во-вторых, во многих странах (компаниях, агентствах и т.п.) статистическая информация собирается и анализируется при помощи разных инструментов и методов, что делает в ряде случаев данные несравнимыми.


В-третьих, при сборе и обработке информации могут случаться ошибки, вызванные человеческим фактором (опечатки и т.п.), которые также необходимо исключить для получения достоверных выводов.

Первичный анализ данных обязательно должен включать в себя **проверку данных на однородность, анализ «аномальных» наблюдений с точки зрения исследования и их исключение из выборки в случае необходимости**. Важно помнить, что каждое наблюдение очень дорого для исследования и исключение его из рассмотрения должно быть обосновано.

Замечание. Во многих случаях **исключение аномальных наблюдений** (наблюдений, числовые характеристики которых

существенно отличаются от других наблюдений), **имеет смысл «отложить» на следующие этапы эконометрического моделирования** (отметив аномальные наблюдения), если только эти наблюдения – не следствие явных ошибок или опечаток.

Первичный анализ данных можно провести визуально, путем анализа графиков и диаграмм. В п. 8 лабораторной работы №1 приведен пример построения графика одного ряда данных.

Для построения нескольких графиков на одной координатной плоскости также необходимо сначала создать *индексную переменную* (если она еще не создана), затем нажать внизу окна программы иконку , либо выбрать в меню **Вид / График / Разброс X-Y**. В появившемся окне в поле «Ось X» перенести переменную **index**, а в поле «Ось Y» – переменные, для которых строится график, например **GDP** и **POP**, затем нажать кнопку ОК (рис. 2.3).

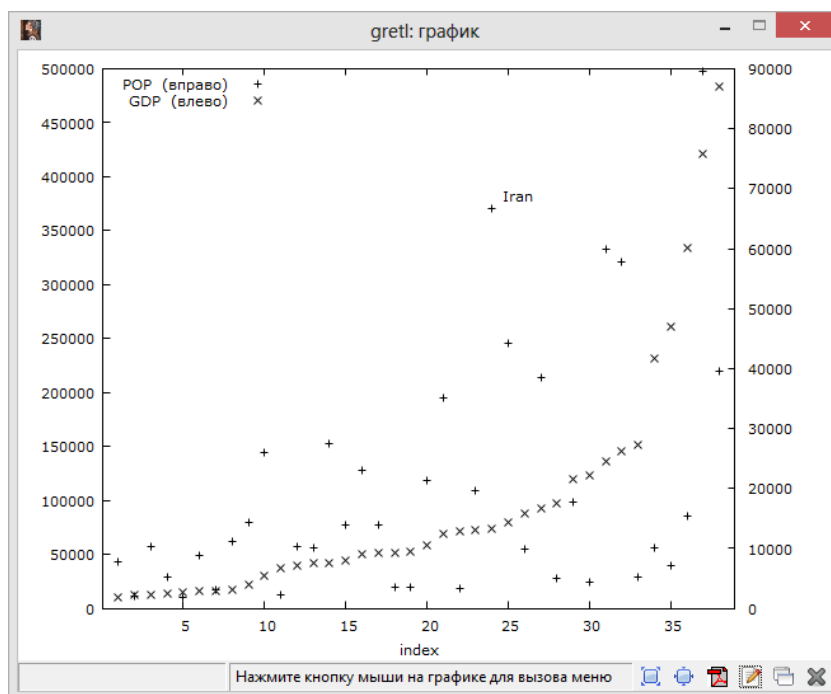


Рис. 2.3. Одновременное представление двух рядов данных

Обратите внимание, если переменные имеют существенно различающиеся размерности данных, то на графике отображаются две оси ординат. На рис. 2.3 слева ось для переменной **POP**, а справа – для **GDP**.

Графическое представление данных дает возможность увидеть наблюдения, «выпадающие» из общей картины. Так, значение численности населения для 24-го наблюдения (Iran) существенно отличается от основной массы (см. рис. 2.3). При этом соответствующих положительных/отрицательных колебаний в значениях ВВП нет. Если причину такого отклонения невозможно объяснить, то такие наблюдения из выборки, которая будет использоваться при моделировании, лучше исключить. Средства пакета *Gretl* позволяют не удалять из рабочего файла «лишние» наблюдения, а сократить используемую для моделирования выборку (см. п. 5).

3. Анализ графика зависимости рядов

Анализ графиков рядов позволяет легко увидеть наблюдения, значения которых существенно больше или меньше средних уровней ряда. Поскольку конечной целью эконометрического моделирования является построение уравнения, отражающего зависимость результативного признака от влияющих факторов, то на данном этапе необходимо определить, объясняются ли отклонения рассматриваемыми в исследовании факторами.

График зависимости рядов или диаграмма рассеяния является удобным инструментом для проведения такого анализа. Создание графика зависимости двух рядов описано в п. 8 лабораторной работы №1. Диаграмма рассеяния трех и более рядов строится аналогично графику зависимости двух рядов с одним отличием: в поле «Ось Y» переносятся все переменные, для которых рассматривается зависимость. Например, на рис. 2.4 представлен график зависимости переменных **EDUC** и **POP** от переменной **GDP**.

При исследовании, в первую очередь, рассматривают диаграммы рассеяния, отражающие зависимость результативного ряда и рядов факторов. Именно из их анализа можно сделать предположения о функциональной форме зависимости рядов.

Анализируя диаграммы рассеяния, можно увидеть «выпадающие» из общей массы значения. Так, на рис. 2.4. видно, что есть наблюдение (страна Mexico) с чрезмерно большой численностью населения при средних значениях уровня ВВП и расходов на образование.

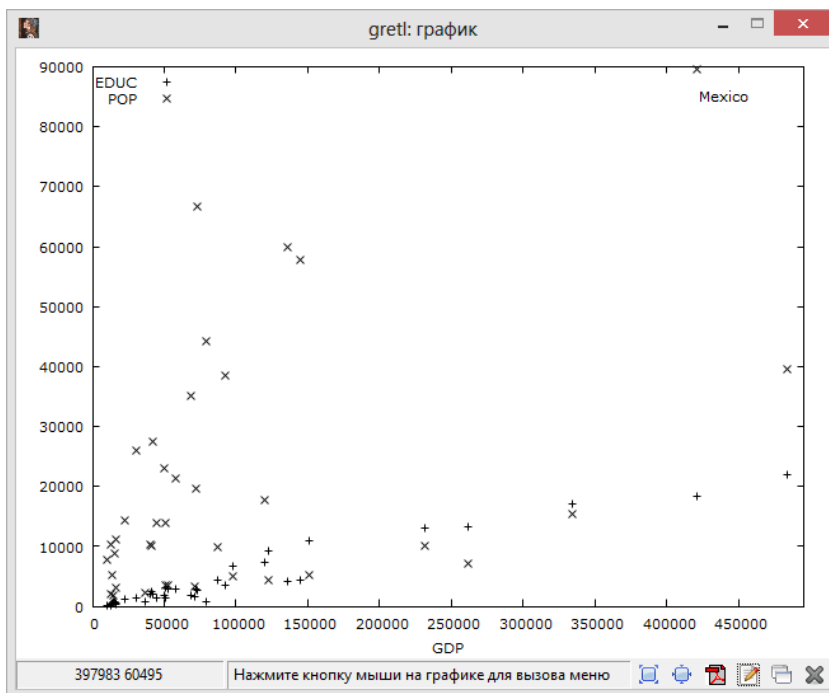


Рис. 2.4. График зависимости рядов

Идеальным вариантом является максимально близкое соответствие диаграммы рассеяния выбранной форме зависимости (линейной, квадратичной и т.д.). В противном случае, качество полученного уравнения может оказаться невысоким.

Подтвердить или опровергнуть предположения о наличии линейной связи можно при помощи коэффициента корреляции. Способы его расчета будут описаны ниже.

4. Расчет и анализ выборочных характеристик рядов данных

При первичном анализе рядов важную роль играет анализ выборочных статистических характеристик. Они позволяют исследовать как ряды данных по отдельности, так и взаимосвязь между ними. Эти характеристики могут служить математическим подтверждением выводов, построенных на анализе графиков и диаграмм рассеяния.

Для **расчета выборочных характеристик** необходимо выбрать в меню **Вид / Описательная статистика**, далее в открывшемся окне «Статистика» добавить переменные, для которых ведется расчет, и нажать ОК. Либо выделить в стартовом окне *Gretl* необходимые ряды, вызвать правой кнопкой мыши контекстное меню и выбрать пункт **Описательная статистика**. В результате появится следующее окно (рис. 2.5).

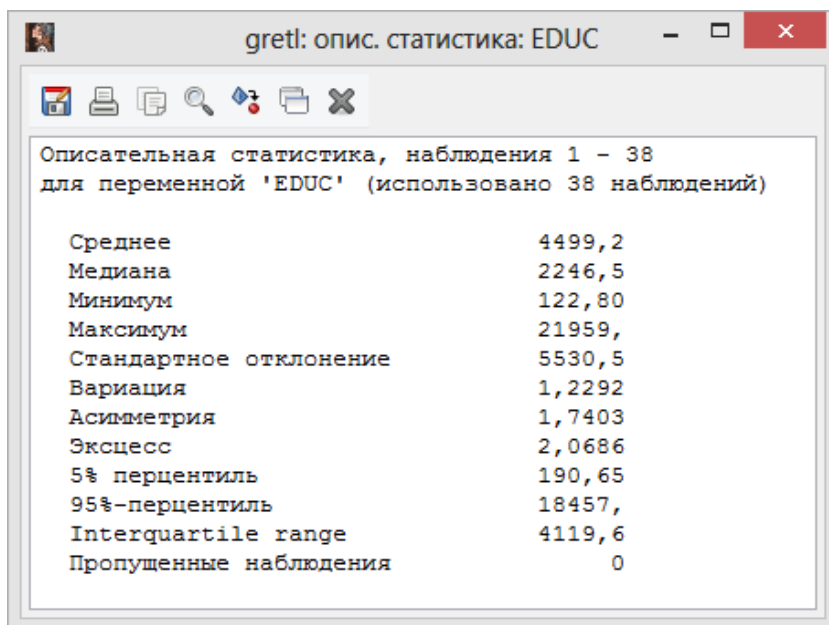


Рис. 2.5. Статистические характеристики одного ряда

Для **расчета линейного коэффициента корреляции** между

переменными необходимо выбрать в меню **Вид / Корреляционная матрица**, далее в открывшемся окне «Корреляция» добавить две или более переменных, для которых ведется расчет, и нажать ОК. В результате появится окно с коэффициентом корреляции (для двух переменных) или корреляционной матрицей (для трех и более переменных).

Для проверки ряда на нормальность распределения необходимо выделить нужную переменную, затем выбрать в меню **Переменная / Тест на нормальное распределение**. В результате появится окно с четырьмя рассчитанными статистиками, оценивающими нормальность распределения ряда (рис. 2.6).

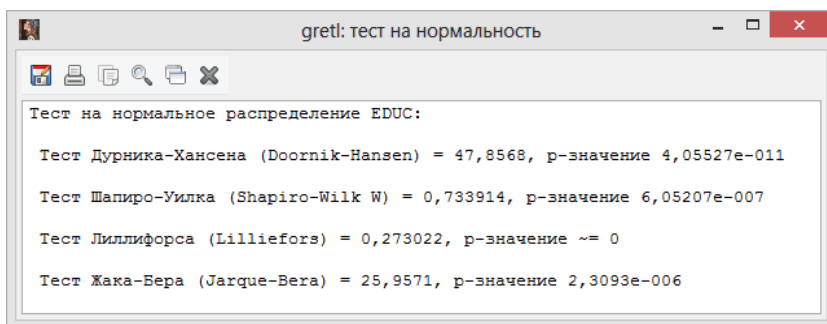


Рис. 2.6. Тест на нормальное распределение

Вывод о нормальности распределения делается на основе *p-значения* соответствующего теста. **Р-значение** – предельный уровень значимости, при котором тест находится на грани между отвержением и не отвержением нулевой гипотезы.

Если *p-значение* $> \alpha$, то имеет место нормальное распределение с доверительной вероятностью $1 - \alpha$; в противном случае распределение не является нормальным. Как правило, α принимается равным 0.05.

Одним из наиболее эффективных критериев проверки нормальности является тест Шапиро-Уилка.

Примечание: значение 2,3093e-006 означает число $2,3093 \cdot 10^{-6} = 0,0000023093$.

Проверить ряд на нормальность распределения также можно вместе с построением гистограммы распределения частот.

Для этого необходимо выбрать в меню **Переменная / Распределение частот...**, в появившемся окне «Распределение частот» указать количество столбцов гистограммы и выделить пункт **Тест на нормальное распределение**, поставить галочку **Показать график**. В появившемся графике *p*-значение теста на нормальное распределение указано в квадратных скобках (рис. 2.7).

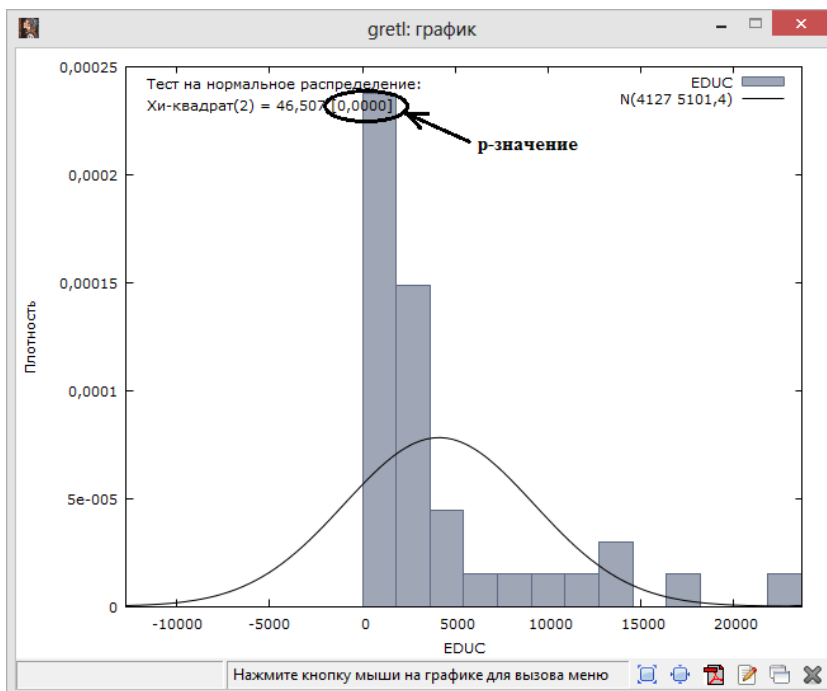


Рис. 2.7. Распределение частот с тестом на нормальное распределение

Анализ выборочных характеристик позволяет:

- проверить гипотезу о нормальности рядов данных, используемых при построении модели;
- оценить степень разброса значений результирующего признака и влияющих факторов
- выделить аномальные наблюдения;
- определить степень тесноты линейной связи исследуемых

рядов.

Анализ выборочных характеристик дает представление исследователю о выборке, с которой он работает. Так, например, анализ рис. 2.5, 2.6 и 2.7, позволяет сделать выводы о том, что в 1997 году:

- 1) средние расходы на образование в рассматриваемых 38 странах составляли 4499.2 млн. долларов;
- 2) максимальные и минимальные расходы на образование – 122.8 и 21959 млн. долл. соответственно;
- 3) переменная EDUC (расходы на образование) не является нормально распределенной.

Тщательный статистический анализ данных играет важную роль в эконометрическом исследовании, помогая на следующих этапах ответить на многие сложные вопросы.

5. Фильтрация данных

Если в результате предварительного анализа принято решение сократить используемую для моделирования выборку, это можно сделать при помощи средств фильтрации данных.

Если точно известны значения конкретного ряда, которые являются «нормальными» для рассматриваемой выборки, то можно исключить «лишние» наблюдения, задавая условие фильтрации данных, выбрав в меню **Выборка / Изменить на основе критерия...**; в появившемся окне «Ограничить выборку» ввести условие фильтрации (рис. 2.8).

Например, в рассматриваемой выборке Мексика (рис. 2.4) с численностью населения (**POP**), превышающей 70000 тыс. чел., сильно нарушает относительную однородность выборки. В результате фильтрации (рис. 2.8) в выборке останутся страны с численностью населения меньше 70000 тыс. чел., а Мексика и страны с численностью населения больше или равной 70000 тыс. чел. из рабочей выборки будут исключены.

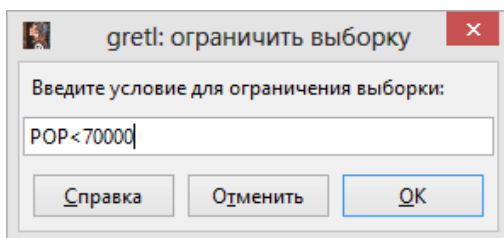


Рис. 2.8. Фильтрация данных

Замечание: название переменной необходимо вводить с учетом регистра.

Если требуется добавить или заменить условие фильтрации, то снова выбирается в меню **Выборка / Изменить на основе критерия...**

Для очистки фильтра необходимо выбрать в меню **Выборка/ Восстановить исходный диапазон.**

Задания

Требуется исследовать зависимость между несколькими экономическими показателями. Необходимо проверить собранные данные, провести их первичный статистический анализ и при необходимости скорректировать рабочую выборку.

Исходные данные по вариантам находятся в файле *lab 2.xls*, номер варианта соответствует номеру страницы в файле.

1. Импортируйте данные согласно варианту.
2. Проведите графический анализ данных, совместное поведение исследуемых величин; отразите в отчете ваши заключения. Здесь и далее построенные графики необходимо сохранять и подписывать в текущей сессии.
3. Проведите анализ диаграмм рассеяния, подготовьте выводы о форме функциональной зависимости между рядами, о тесноте связи (рассчитайте линейные коэффициенты корреляции).
4. Проведите анализ выборочных характеристик рядов данных, дайте характеристику выборки по каждому показателю.
5. Рассчитайте коэффициенты вариации для каждого показателя и оцените степень однородности выборки.
6. **Если необходимо**, проведите фильтрацию (очистку) рядов

данных.

7. Сохраните рабочий файл под именем *Фамилия студента_2.gdt*.
8. Подготовьте отчет о результатах исследования.

Вопросы для самоконтроля

1. Что такое генеральная совокупность? Что представляет собой случайная выборка?
2. Что такое гистограмма распределения?
3. Что означает однородность данных?
4. Что является мерой однородности данных?
5. Как рассчитывается коэффициент вариации?
6. Как определить однородность выборки по коэффициенту вариации?
7. Каковы основные причины появления аномальных наблюдений в выборке?
8. Какими способами можно проверить наличие аномальных наблюдений в выборке?
9. Каким образом в *Gretl* можно исключить аномальные наблюдения из рабочей выборки?
10. Что такое поле корреляции? Как иначе называется поле корреляции?
11. Какую информацию можно получить, анализируя диаграмму рассеяния?
12. Как вычисляется коэффициент корреляции?
13. Как по коэффициенту корреляции определить направление связи?
14. Как по коэффициенту корреляции определить тесноту связи?
15. Какие значения может принимать коэффициент корреляции?
16. Пусть стандартное отклонение для некоторого показателя равно 2.5, а среднее значение равно 5.1 для данной выборки. Что можно сказать о значениях данного показателя в этой выборке?
17. Что показывает стандартное отклонение в окне «Описательная статистика» на рис. 2.5?
18. Как проверить ряд данных на нормальность распределения?
19. Что такое «оценка параметра»? От чего она зависит?
20. Какая оценка называется несмещенной?

21. Какая оценка называется эффективной?
22. Какая оценка называется состоятельной?
23. Что такое статистическая гипотеза?
24. В чем состоит схема проверки гипотез?
25. Что такое уровень значимости теста?
26. Что такое мощность теста?
27. В чем суть статистического теста?
28. Что называется Р-значением?
29. Что такое доверительный интервал?

Лабораторная работа № 3. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Эконометрическое исследование проводится на основе **выборочных данных**, которые отбираются из части всей совокупности по определенным правилам выборки и обеспечивают получение данных, характеризующих всю совокупность в целом.

Наиболее простым и распространенным предположением о взаимосвязи факторов является предположение об их линейной зависимости.

Парная линейная регрессия представляет собой линейную зависимость между двумя переменными: y и x , т.е. модель вида

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad t = \overline{1, n},$$

где y_t — значение зависимой переменной для наблюдения t ;
 x_t — значение независимой переменной для наблюдения t ;
 β_0 и β_1 — коэффициенты (параметры) регрессии;
 ε_t — значение случайной компоненты для наблюдения t ;
 n — число наблюдений.

Оценки коэффициентов парной линейной регрессии $\hat{\beta}_0$ и $\hat{\beta}_1$ определяются **методом наименьших квадратов (МНК)**. Полученные значения $\hat{\beta}_0$ и $\hat{\beta}_1$ позволяют определить \hat{y}_t — расчетное (прогнозное) значение зависимой (объясняемой, эндогенной) переменной для наблюдения t .

Метод наименьших квадратов минимизирует сумму квадра-

тов ошибок регрессии (случайных компонент):

$$\sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n \left[y_t - (\hat{\beta}_0 + \hat{\beta}_1 x_{1t}) \right]^2 \rightarrow \min_{\beta_0, \beta_1}.$$

При оценке качества парной линейной регрессии необходимо обратить внимание на следующие моменты.

1. Значимость коэффициентов регрессии

Истинное значение коэффициентов регрессии неизвестно. На основе имеющихся наблюдений можно рассчитать только их оценки (приближенные значения). Заметим, что если коэффициент $\beta_1 = 0$ (даже если значение оценки этого коэффициента $\hat{\beta}_1 \neq 0$), то фактически переменная y не зависит от x и уравнение регрессии не имеет смысла.

Коэффициент регрессии значим на заданном уровне доверия (значимости), если можно принять гипотезу, что его истинное значение отлично от нуля. **Коэффициент регрессии незначим на заданном уровне доверия (значимости)**, если принимается гипотеза, что его истинное значение равно нулю.

Для проверки значимости коэффициента регрессии (проверки гипотезы $\beta_i = 0$) необходимо:

- 1) задать **уровень значимости** $\alpha \in (0,1)$; обычно это значение 0,05 (величина $1 - \alpha$ называется **уровнем доверия**);
- 2) рассчитать **стандартные ошибки оценок** — среднеквадратические отклонения коэффициентов регрессии от их истинных значений:

$$S_{\hat{\beta}_0} = \sqrt{\frac{1}{n-2} \cdot \frac{\sum_i \left(y_i - \hat{y}_i \right)^2}{\sum_i \left(x_i - \bar{x} \right)^2} \cdot \overline{x^2}}, S_{\hat{\beta}_1} = \sqrt{\frac{1}{n-2} \cdot \frac{\sum_i \left(y_i - \hat{y}_i \right)^2}{\sum_i \left(x_i - \bar{x} \right)^2}};$$

- 3) рассчитать значение критерия Стьюдента $t_{\beta_i} = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$;
- 4) определить $t_{кр}$ — табличное (критическое) значение t -критерия Стьюдента для уровня значимости α и числа степеней свободы $n - 2$ (n — число наблюдений);
- 5) сравнить t_{β_i} и $t_{кр}$: если $|t_{\beta_i}| > t_{кр}$, то коэффициент β_i значим с уровнем доверия $(1 - \alpha)$, иначе — незначим.

Замечания

1) Значение $t_{кр}$ определяются по специальным таблицам для распределения Стьюдента. Так, при числе степеней свободы более 200

$t_{кр} = 1.65$ для $\alpha = 0.10$ (10%);

$t_{кр} = 1.96$ для $\alpha = 0.05$ (5%);

$t_{кр} = 2.58$ для $\alpha = 0.01$ (1%)

2) Чем меньше значение стандартной ошибки по сравнению со значением соответствующего коэффициента, тем ближе оценки коэффициентов регрессии к их истинным значениям.

3) Аналогично проверяется значимость и в случае многих переменных (множественной регрессии).

2. Доверительные интервалы коэффициентов показывают, что истинное значение параметра с вероятностью $1 - \alpha$ находится в пределах

$$\beta_i \in \left(\hat{\beta}_i - t_{кр} \cdot S_{\hat{\beta}_i}; \hat{\beta}_i + t_{кр} \cdot S_{\hat{\beta}_i} \right),$$

где α — уровень значимости.

Чем меньше доверительный интервал относительно коэффициента, тем точнее полученная оценка.

Пример. Пусть значение $\hat{\beta}_1 = 2,1$, $S_{\hat{\beta}_1} = 0.8$, число наблюдений 250. Тогда:

1) для уровня значимости 0,05 критическое значение критерия Стьюдента $t_{кр} = 1,96$;

2) значение критерия Стьюдента для данного коэффициента равно $t_{\beta_1} = \frac{2,1}{0,8} = 2,625 > t_{кр} = 1,96$, следовательно, коэффициент β_1 значим на уровне доверия 0,95 (на уровне значимости 0,05), т.е. с вероятностью 0,95 истинное значение этого коэффициента отлично от нуля;

3) истинное значение коэффициента β_1 находится в интервале

$$\beta_1 \in (2,1 - 1,96 \times 0,8; 2,1 + 1,9 \times 0,8) = (2,1 - 1,568; 2,1 + 1,568) = (0,532; 3,668).$$

Величина интервала $(3,668 - 0,532 = 3,136)$ относительно оценки коэффициента $\hat{\beta}_1 = 2,1$ большая, следовательно, точность оценки низкая.

3. Коэффициент детерминации R^2 показывает степень соответствия найденного уравнения фактическим данным (качество подгонки уравнения):

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2},$$

где \hat{y}_i — прогнозные (расчетные) значения зависимой переменной y_i , \bar{y} — среднее значение зависимой переменной y .

R^2 изменяется в пределах $[0; 1]$, и чем ближе его значение к 1, тем лучше модель согласуется с выборочными данными.

Например, если $R^2 = 0,75$, то говорят, что на 75 % изменение переменной y описывается полученным уравнением и влиянием переменной x , а 25 % изменения y — следствие влияния неучтенных в уравнении регрессии факторов.

Коэффициент детерминации не используется, если в уравнении отсутствует константа β_0 (в этом случае его значения могут выйти за пределы $[0; 1]$).

4. Стандартная ошибка регрессии S_e является оценкой величины квадрата ошибки, приходящейся на одну степень свободы модели

$$S_e = \sqrt{\frac{\sum_{i=1}^n \left(\hat{y}_i - y_i \right)^2}{n-2}}.$$

Она показывает, насколько прогнозные значения зависимой переменной отличаются от фактических значений. Чем меньше S_e по сравнению с зависимой переменной, тем лучше качество модели.

5. Значимость уравнения регрессии

Уравнение значимо на заданном уровне доверия (уровне значимости), если можно принять гипотезу о том, что существует хотя бы один коэффициент при независимых переменных, отличный от нуля. Проверяется по F-критерию Фишера:

$$F = \frac{R^2}{1-R^2} \cdot (n-2).$$

Для заданного уровня значимости и по числу степеней свободы, равному $(n-2)$, определяется критическое значение критерия Фишера $F_{кр}$. Если $F > F_{кр}$, то уравнение статистически значимо, иначе — незначимо.

***Замечание.** Для парной регрессии значимость уравнения в целом равносильна значимости коэффициента при независимой переменной.*

6. Средняя абсолютная процентная ошибка (ошибка аппроксимации) показывает в процентах среднее отклонение расчетных значений зависимой переменной \hat{y}_i от фактических значений y_i :

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%.$$

Если $MAPE \leq 10 \%$, то качество подгонки уравнения считается хорошим. Чем меньше значение $MAPE$, тем лучше.

Экономическая интерпретация парной линейной регрессии

Параметр $\hat{\beta}_1$ показывает, насколько изменится среднее значение y при увеличении x на единицу.

Параметр $\hat{\beta}_0$ является значением y при $x = 0$. Он может не иметь экономического содержания.

Пример. Пусть по 300 наблюдениям получено следующее уравнение регрессии:

$$\hat{y} = 2,01 + 3,46x, R^2 = 0,64.$$

(s.e) (0,51) (1,28)

В скобках под значениями коэффициентов приведены их стандартные ошибки. По их значениям можно утверждать, что:


- 1) на 5 %-м уровне значимости коэффициенты значимы;
- 2) качество подгонки удовлетворительное, на 64 % изменение y описывается полученным уравнением и влиянием переменной x ;
- 3) уравнение значимо в целом;
- 4) с увеличением переменной x на единицу значение переменной y увеличится в среднем на 3,46 единиц.

Построение уравнения парной регрессии в Gretl

Для построения уравнения парной линейной регрессии необходимо иметь два ряда выборочных данных, характеризующих значения зависимой и независимой переменных.

1. Оценка параметров модели методом наименьших квадратов

В *Gretl* оценка модели парной линейной регрессии осуществляется следующим образом.

Необходимо нажать внизу окна программы иконку , либо выбрать в меню **Модель / Метод наименьших квадратов...** В появившемся окне «*Спецификация модели*» в поля «Зависимая переменная» и «Регрессоры» переносятся соответствующие переменные. На рис. 3.1 приведен пример построения модели зависимости **GDP** от **EDUC**.

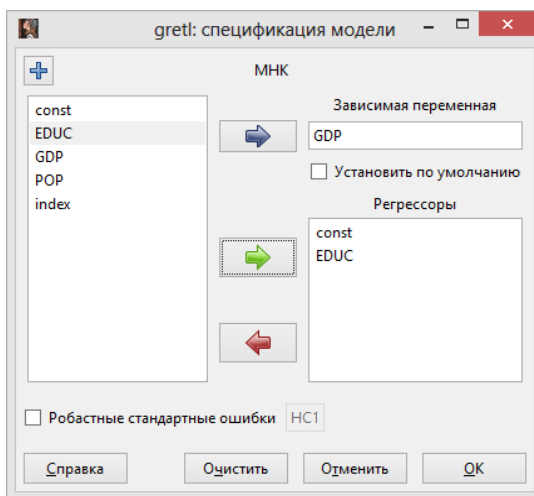


Рис. 3.1. Окно Спецификация модели для оценки парной линейной регрессии

В результате оценки появится окно Модель (рис. 3.2).

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	11495,0	5824,27	1,974	0,0564	*
EDUC	18,5844	0,894957	20,77	2,91e-021	***
Среднее зав. перемен	88191,92	Ст. откл. зав. перемен	98578,35		
Сумма кв. остатков	2,63e+10	Ст. ошибка модели	27393,08		
R-квадрат	0,924927	Испр. R-квадрат	0,922782		
F(1, 35)	431,2124	P-значение (F)	2,91e-21		
Лог. правдоподобие	-429,5404	Крит. Акаике	863,0808		
Крит. Шварца	866,3026	Крит. Хеннана-Куинна	864,2166		

Рис. 3.2. Результаты оценки модели парной линейной регрессии

В данном окне (рис. 3.2) столбец **Коэффициент** — это полученные оценки коэффициентов, т. е. для данного примера уравнение регрессии будет выглядеть следующим образом:

$$GDP = 11495 + 18,5844 \cdot EDUC.$$

Для просмотра оцененного уравнения необходимо выбрать в окне *Модель* меню **Файл / Просмотреть как уравнение**.

Для сохранения результатов оценки в текущей сессии выберите в окне *Модель* меню **Файл / Сохранить в текущей сессии**. Для переименования модели в текущей сессии надо щелкнуть правой кнопкой мыши по иконке модели и в контекстном меню выбрать **Переименовать**.

2. Оценка качества парной регрессии

Оценить качество регрессии можно по следующей информации, представленной в окне *Модель* (рис. 3.2).

1. Величины стандартных ошибок коэффициентов регрессии приведены в столбце **Ст. ошибка**.

В столбце **t-статистика** представлены значения *t*-статистик для коэффициентов. Для проверки значимости коэффициентов регрессии указанное значение *t*-статистики сравнивается с $t_{кр}$, определяемым по специальным таблицам для заданного уровня значимости.

В столбце **P-значение** показана вероятность того, что гипотеза о незначимости коэффициента верна. Для вывода *P*-значение соответствующего коэффициента сравнивается с уровнем значимости α : если *P-значение* < 0.01 , коэффициент значим на уровне значимости 0,01 (1%) (на уровне доверия 99 %); если *P-значение* $< 0,05$, коэффициент значим на уровне значимости 0,05 (5%) (на уровне доверия 95 %).

Справа от *P-значений* есть подсказки об оценке значимости коэффициентов в виде *. Если напротив коэффициента стоит * — коэффициент значим на уровне значимости 10%, ** — значим на уровне значимости 5%, *** — значим на уровне значимости 1%.

2. Доверительные интервалы коэффициентов на 95% уровне доверия строятся с помощью меню **Анализ / Доверительные интервалы для коэффициентов**.
3. Коэффициент детерминации R^2 приводится в поле **R-квадрат**.
4. Стандартная ошибка регрессии S_e (поле **Ст. ошибка модели**)

может сравниваться со средним значением зависимой переменной в поле **Среднее зав. перемен.** Чем меньше S_e по отношению к среднему значению зависимой переменной, тем лучше качество модели.

5. В поле **F(1,n-2)** указывается значение F -критерия Фишера для уравнения регрессии. Здесь n – количество наблюдений. Значимость уравнения регрессии определяется путем сравнения с табличным (критическим) значением F -критерия Фишера для выбранного уровня значимости.
В поле **Р-значение (F)** приводится вероятность того, что гипотеза о незначимости уравнения верна. При проверке значимости уравнения регрессии **Р-значение (F)** сравнивается с уровнем значимости α : если $P\text{-значение (F)} < 0.01$, уравнение значимо на уровне значимости $\alpha = 0.01$ (на уровне доверия 99 %); если $P\text{-значение (F)} < 0.05$, уравнение значимо на уровне значимости $\alpha = 0.05$ (на уровне доверия 95 %).
6. Для определения **средней абсолютной процентной ошибки** необходимо в окне *Модель* (рис. 3.2) выбрать в меню **Анализ / Наблюдаемые и расчетные значения**, в появившемся окне под расчетными данными приведена статистика для оценки прогноза (рис. 3.3) с рассчитанным значением средней абсолютной процентной ошибки (MAPE).

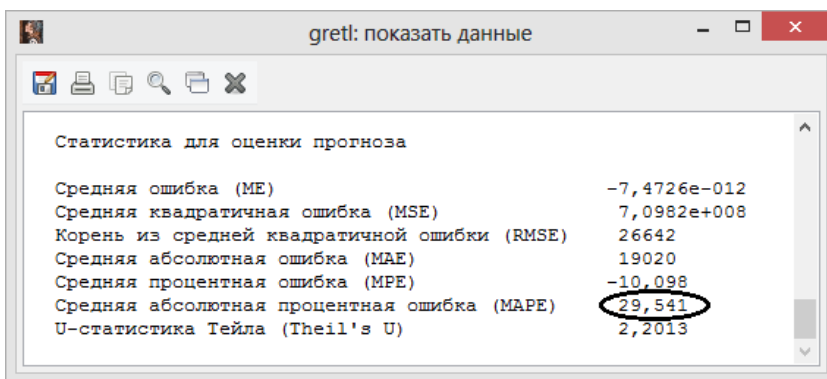


Рис. 3.3. Расчет средней абсолютной процентной ошибки

На рис. 3.3 значение средней абсолютной процентной ошиб-

ки равно $MAPE = 29,54\% > 10\%$, что является достаточно высоким значением и указывает на плохое качество подгонки уравнения к выборочным данным.

Чтобы проверить качество построенного уравнения регрессии, можно также провести анализ фактических, теоретических значений зависимой переменной и остатков регрессии. Для этого в окне *Модель* (рис. 3.2) необходимо выбрать меню

Графики / График наблюдаемых и расчетных значений / По номеру наблюдения (рис. 3.4) либо

Графики / График остатков / По номеру наблюдения

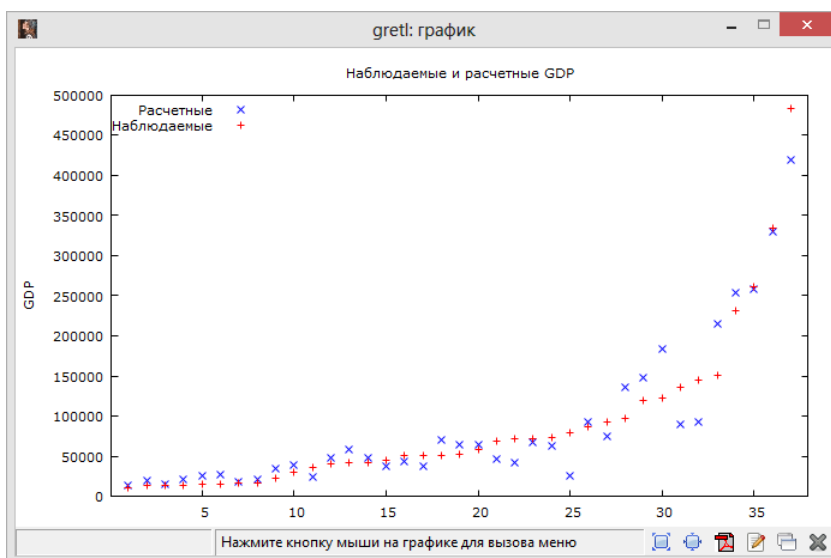


Рис. 3.4. График фактических и прогнозных значений

Качество уравнения считается хорошим, если графики фактических и прогнозных данных (рис. 3.4) близки, а значения на графике остатков невелики относительно значений зависимой переменной.

Задания

Исследовать влияние факторов на зависимую переменную путем построения уравнения парной линейной регрессии. Исходные данные по вариантам находятся в файле *lab 3.xls*.

1. Проведите анализ данных и подготовьте выборку к эконометрическому моделированию.
2. Выберите один из объясняющих факторов и выполните следующие действия для построения одного уравнения регрессии:
 - 1) по исходным данным найдите оценки коэффициентов регрессии β_0 и β_1 , используя МНК;
 - 2) оцените качество построенной модели (выясните, значимы ли параметры регрессии, значимо ли уравнение в целом, оцените качество подгонки к выборочным данным по коэффициенту детерминации и коэффициенту аппроксимации);
 - 3) постройте доверительные интервалы для коэффициента β_1 на уровне значимости 5%;
 - 4) рассчитайте коэффициент корреляции между зависимой и независимой переменной, дайте его интерпретацию;
 - 5) запишите полученное уравнение регрессии;
 - 6) дайте экономическую интерпретацию коэффициентов.
3. Постройте линейное уравнение регрессии для второй влияющей величины. Оцените качество уравнения.
4. Сравните полученные модели и выберите лучшую модель по степени объяснения значений зависимой переменной (качеству подгонки уравнения к выборочным данным).
5. Сохраните рабочий файл в вашем разделе под именем *Фамилия студента_3.gdt*.
6. Для лучшей регрессии рассчитайте параметры регрессии в Excel, используя построение линии тренда и используя статистическую функцию «Регрессия», сравните полученные результаты. Сохраните файл Excel.
7. Оформите отчет о проделанной работе.

Вопросы для самоконтроля

1. Что такое парная линейная регрессия?
2. В чем суть метода наименьших квадратов?
3. По каким формулам рассчитываются оценки параметров регрессии для парной линейной регрессии?
4. В чем отличие β_1 и $\hat{\beta}_1$?

5. Сформулируйте условия теоремы Гаусса-Маркова.
6. Какие оценки параметров называются несмещенными?
7. Какие оценки параметров называются эффективными?
8. Какие оценки параметров называются состоятельными?
9. Какими свойствами обладают оценки параметров парной линейной регрессии, найденные методом наименьших квадратов при выполнении условий теоремы Гаусса-Маркова?
10. В чем суть условия гомоскедастичности?
11. В чем состоит условие отсутствия автокорреляции?
12. В чем отличие классической линейной регрессионной модели от нормальной классической линейной регрессионной модели?
13. Что показывает полученный доверительный интервал для параметра β_1 на уровне значимости 5%?
14. Что такое значимость параметра регрессии?
15. Что такое t-статистика и как она используется для проверки значимости параметра регрессии?
16. Как используется t-статистика для построения доверительного интервала для параметра регрессии?
17. Перечислите основные этапы проверки значимости параметра линейной регрессии.
18. Что такое Р-значение и как оно используется при анализе значимости параметров регрессии?
19. Как связан наклон линии регрессии и значение коэффициента β_1 при объясняющей переменной в парной линейной регрессии?
20. Чем отличаются ошибки регрессии от остатков регрессии?
21. В чем отличие теоретической регрессии от выборочной?
22. Что такое сумма квадратов (дисперсия), объясненная регрессией?
23. Что такое сумма квадратов (дисперсия), не объясненная регрессией?
24. Как можно представить вариацию (дисперсию, разброс) значений зависимой переменной через дисперсию, объясненную регрессией и дисперсию, не объясненную регрессией?
25. Как рассчитывается коэффициент детерминации?
26. Что показывает коэффициент детерминации?

27. Какие значения может принимать коэффициент детерминации?
28. Как связаны коэффициент парной линейной корреляции и коэффициент детерминации для уравнения парной линейной регрессии?

Лабораторная работа № 4. **МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ**

На практике часто возникает ситуация, когда нужно проанализировать влияние ряда факторов на исследуемый показатель. В этом случае необходимо рассматривать обобщение парной регрессии — модель множественной регрессии. Линейная модель множественной регрессии выглядит следующим образом:

$$y_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_i x_{it} + \dots + \beta_k x_{kt} + \varepsilon_t, t = 1, 2, \dots, n, \quad (3.1)$$

где y_t — значение зависимой (эндогенной) переменной для наблюдения t ;

x_{it} — значение независимой (экзогенной) переменной с номером i для наблюдения t ;

ε_t — случайная компонента для наблюдения t (она учитывает влияние неучтенных в модели факторов);

k — количество независимых переменных (регрессоров в уравнении);

n — число наблюдений.

Общая последовательность построения множественной линейной регрессионной модели состоит из следующих этапов.

1. Оценка параметров (коэффициентов) уравнения.
2. Оценка значимости параметров регрессии и уравнения регрессии в целом.
3. Оценка качества подгонки регрессионного уравнения к выборочным данным.
4. Проверка переменных модели на мультиколлинеарность, ее исключение.
5. Проверка модели на гетероскедастичность, коррекция на гетероскедастичность в случае необходимости.
6. Проведение тестов на функциональную форму, корректиров-

ка вида модели в случае необходимости (подробно будет рассмотрено в лабораторной работе № 6).

7. Экономическая интерпретация параметров (коэффициентов) уравнения регрессии.

Рассмотрим отдельные этапы построения множественной линейной регрессии.

1. Оценка параметров уравнения

Для определения параметров уравнения множественной линейной регрессии, как и в случае парной линейной регрессии, используется метод наименьших квадратов, который минимизирует сумму квадратов ошибок регрессии (случайных компонент):

$$\sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n [y_t - (\hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + \dots + \hat{\beta}_k x_{kt})]^2 \rightarrow \min_{\beta_0, \beta_1, \dots, \beta_k}.$$

Оптимальные значения параметров $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ являются оценками, приближенными значениями истинных (неизвестных нам) параметров $\beta_0, \beta_1, \dots, \beta_k$.

Решение этой системы (искомые оценки параметров $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$) в матричном виде осуществляется следующим образом:

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

где $X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix}$ — матрица объясняющих (экзо-

генных) переменных (она получена путем выписывания значений всех переменных для имеющихся наблюдений и добавлением слева единичного столбца, матрица является прямоугольной и имеет размерность $n \times (k + 1)$, т. е. имеет n строк и $(k + 1)$ столбец);

$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$ — вектор значений зависимой (эндогенной) переменной;

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}$ — вектор параметров (коэффициентов) уравнения;

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$ — вектор ошибок регрессии (случайных компонент);

X^T — транспонированная матрица объясняющих переменных.

2. Оценка значимости параметров регрессии и уравнения регрессии в целом

Важным этапом анализа качества построенного уравнения регрессии является проверка значимости параметров регрессии и уравнения регрессии в целом.

Для определения значимости параметров β_i (проверки гипотезы $H_0 : \beta_i = 0$ против гипотезы $H_1 : \beta_i \neq 0$) используется

тот факт, что случайная величина $t_{\beta_i} = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$ имеет распределение

Стьюдента с $(n - k - 1)$ степенями свободы, где $S_{\hat{\beta}_i}$ — стандартная ошибка оценки параметра $\hat{\beta}_i$. Величина $S_{\hat{\beta}_i}$ показывает, насколько в среднем значение $\hat{\beta}_i$ отклоняется от истинного значе-

ния β_i . Значение $S_{\hat{\beta}_i}$ вычисляется по формулам, аналогичным приведенным в лабораторной работе № 3.

Незначимость параметра β_i означает, что $\beta_i = 0$ для выбранного уровня доверия (уровня значимости) и что переменная y не зависит фактически от переменной x_i . Это может быть основанием для исключения переменной x_i из модели множественной линейной регрессии. Однако незначимость β_i может также означать, что связь между переменными y и x_i носит более сложный, нелинейный характер и имеет смысл исследовать нелинейные модели.

При определении значимости параметров уравнения и уравнения в целом необходимо выполнить следующие действия.

1. Рассчитать $S_{\hat{\beta}_i}$ стандартные ошибки оценок.

2. Рассчитать $t_{\beta_i} = \frac{\hat{\beta}_i}{S_{\beta_i}}$.

3. Определить по таблицам распределения Стьюдента $t_{кр}$ для выбранного уровня значимости и числа степеней свободы $(n - k - 1)$, где n — число наблюдений, k — число независимых переменных модели.

4. Сравнить t_{β_i} с $t_{кр}$: если $|t_{\beta_i}| > t_{кр}$, то коэффициент β_i значим на выбранном уровне значимости, в противном случае — незначим.

Для определения возможных значений коэффициентов уравнения необходимо построить доверительные интервалы коэффициентов

$$\beta_i \in \left(\hat{\beta}_i - t_{кр} \cdot S_{\hat{\beta}_i}; \hat{\beta}_i + t_{кр} \cdot S_{\hat{\beta}_i} \right), i = \overline{1, k}.$$

Оценка значимости уравнения в целом состоит в проверке гипотезы $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$, т. е. проверке того факта, что все коэффициенты при независимых переменных в уравнении регрессии равны нулю и переменные модели не оказывают ника-

кого влияния на зависимую переменную. В этом случае уравнение не имеет смысла, т. е. незначимо.

Оценка значимости уравнения в целом осуществляется при альтернативной гипотезе H_1 , состоящей в том, что гипотеза H_0 неверна, т.е. существует хотя бы один коэффициент, отличный от нуля.

Для проверки гипотезы H_0 рассчитывается F -статистика:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k},$$

где n — число наблюдений, k — число независимых переменных (не считая свободного члена). F -статистика имеет распределение Фишера со степенями свободы $(k, n - k - 1)$.

Чтобы проверить значимость уравнения в целом, надо:

- 1) задать уровень значимости α ;
- 2) для выбранного уровня значимости и числа степеней свободы по таблицам, в которых приведены квантили распределения Фишера, найти критическое значение F -статистики;
- 3) сравнить рассчитанное значение F с критическим значением $F_{кр}$: если $F > F_{кр}$, то гипотеза H_0 отвергается, принимается альтернативная гипотеза – уравнение в целом значимо для выбранного уровня доверия; в противном случае, если $F \leq F_{кр}$, гипотеза H_0 принимается, делается вывод, что уравнение в целом незначимо.

F -критерии в разных моделях с разным числом наблюдений и (или) переменных несравнимы.

3. Оценка качества подгонки регрессионного уравнения к данным

Для оценки качества подгонки уравнения к выборочным данным, т. е. проверки близости к фактическим рассчитанных по модели значений, используется *коэффициент детерминации* R^2 . Он определяется аналогично тому, как это было для парной линейной регрессии:

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2},$$

где $\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$ — среднее значение зависимой переменной,

$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \dots + \hat{\beta}_k x_{kt}$ — рассчитанное по модели (прогнозное) значение зависимой переменной.

Справедливо следующее соотношение: $0 \leq R^2 \leq 1$. Чем ближе R^2 к единице, тем в выше качество подгонки, тем ближе расчетные значения к фактическим. Близость R^2 к нулю означает, что в качестве прогноза лучше использовать среднее значение зависимой переменной, а не расчетные значения \hat{y}_t .

Значение коэффициента детерминации R^2 повышается, если число независимых переменных возрастает, независимо от «ценности» вклада дополнительной переменной.

Чтобы исключить при оценке качества подгонки влияние на величину коэффициента детерминации увеличения числа переменных в модели, рассчитывают *скорректированный или нормированный (adjusted) коэффициент детерминации* R_{adj}^2 :

$$R_{adj}^2 = 1 - \frac{n-1}{n-k-1} \cdot (1 - R^2),$$

где n — число наблюдений, k — число независимых переменных (не считая свободного члена).

Отметим некоторые свойства скорректированного коэффициента детерминации R_{adj}^2 :

- 1) $R^2 \geq R_{adj}^2$;
- 2) $R_{adj}^2 \leq 1$, но может быть меньше нуля!

Если при добавлении переменной в модель линейной регрессии увеличивается не только значение коэффициента детерминации R^2 , но и значение скорректированного коэффициента

детерминации R_{adj}^2 , то можно утверждать, что вклад этой переменной в объяснение изменений зависимой переменной (повышение качества подгонки к данным) существенен. Это используется при выполнении пошагового регрессионного анализа, основанного на последовательном включении (или исключении) переменных в модель.

Качество подгонки уравнения к данным также характеризуют следующие показатели: **средняя абсолютная процентная ошибка (MAPE)** (см. для парной регрессии, лаб. работа № 3) и **стандартная ошибка регрессии**

$$S_e = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n - k - 1}}.$$

Значения S_e в однотипных моделях с разным числом наблюдений и (или) переменных сравнимы.

4. Проверка переменных модели на мультиколлинеарность, исключение мультиколлинеарности

Мультиколлинеарность — наличие сильной линейной корреляционной связи между объясняющими переменными.

При наличии эффекта мультиколлинеарности матрица $X^T X$, используемая при вычислении оценок параметров уравнения, будет близка к вырожденной.

Последствия мультиколлинеарности:

1. Стандартные ошибки оценок коэффициентов завышены (больше реальных).
2. Неустойчивость оценок. Добавление или исключение малого количества наблюдений может привести к очень сильному изменению оценок коэффициентов, при этом резко уменьшается и точность предсказания по модели.
3. Высокая коррелированность коэффициентов лишает смысла их интерпретацию.

Признаки мультиколлинеарности:

- 1) некоторые из оценок параметров регрессии имеют неверный с точки зрения экономической теории (здравого смысла) знак;

- 2) небольшое изменение исходных данных (исключение или добавление небольшой порции наблюдений) приводит к существенному изменению оценок параметров уравнения регрессии;
- 3) большинство коэффициентов уравнения регрессии незначимо, хотя уравнение в целом значимо и имеет достаточно высокий коэффициент детерминации;
- 4) высокие парные коэффициенты корреляции между объясняющими переменными;
- 5) значения коэффициента $VIF > 10$. Коэффициент VIF (*variance inflation factor*) характеризует силу мультиколлинеарности. Вычисляется на основе значений R^2 во вспомогательных регрессиях одного регрессора на другие:

$$x_i^{(k)} = \beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_{k-1} x_i^{(k-1)} + u_i$$

$$VIF = \frac{1}{1 - R^2}$$

Одним из способов устранения эффекта мультиколлинеарности является **метод включения-исключения** переменных. При этом выполняются следующие действия:

1. Строится регрессионная модель методом наименьших квадратов.
2. Оценивается значимость параметров регрессии.
3. Выявляется наличие зависимости между факторными признаками путем анализа матрицы парных коэффициентов корреляции и коэффициента VIF .
4. Строится новое уравнение регрессии с исключением незначимых и части взаимно коррелирующих переменных. При этом из числа коррелирующих переменных в модели оставляют те, которые более соответствуют ее экономическому содержанию, либо те, которые имеют наибольшее значение частной корреляции с зависимой переменной. При необходимости включаются уже исключенные переменные, если этого требует экономический смысл.
5. Повторяются 3-й и 4-й шаги до тех пор, пока мультиколлинеарность не будет исключена.

5. Проверка модели на гетероскедастичность, коррекция на

гетероскедастичность в случае необходимости

На практике часто встречаются модели, в которых не выполняется условие теоремы Гаусса — Маркова о том, что все случайные компоненты уравнения регрессии имеют одинаковую дисперсию (одинаковый разброс относительно среднего, нулевого значения). Это условие называется *гомоскедастичностью*. Нарушение условия постоянства дисперсий называется *гетероскедастичностью*.

Гетероскедастичность не приводит к смещению оценок уравнений, т. е. оценки остаются несмещенными, но они не будут эффективными. Гетероскедастичность может привести к заниженным значениям стандартных ошибок, получаемых обычным МНК, вследствие чего завышаются t -статистики и дается неправильное (завышенное) представление о точности оценок. Заметим, что гетероскедастичность остатков может быть вызвана неправильным выбором модели (например, рассмотрением линейной модели в случае, когда истинная связь между переменными носит нелинейный характер).

Для проверки модели на гетероскедастичность чаще всего используется **тест Вайта**, который основан на следующем предположении: если в модели присутствует гетероскедастичность, то это может быть связано с тем, что дисперсии ошибок зависят от регрессоров, а гетероскедастичность должна отражаться в остатках обычной регрессии исходной модели.

В данном методе тестирования гипотезы H_0 не делается предположений относительно структуры гетероскедастичности.

Для проведения теста необходимо:

- провести обычную регрессию и получить вектор регрессионных остатков $e = (e_1, e_2, \dots, e_n)$, где $e_t = \hat{y}_t - y_t$, представляющее отклонение расчетных (прогнозных) значений зависимой переменной от фактических значений;
- провести регрессию e_t^2 на все независимые переменные, их квадраты, попарные произведения и свободный член;
- рассчитать статистику nR^2 .

Если верна гипотеза H_0 : $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2$ (отсутствие гетероскедастичности), то величина имеет распределение $\chi^2(N - 1)$,

где N — количество регрессоров во второй модели.

При наличии гетероскедастичности рассчитываются **состоятельные (робастные к гетероскедастичности) стандартные ошибки в форме Вайта**, которые не устраняют гетероскедастичность, но корректируют значения стандартных ошибок оценок коэффициентов.

6. Тест Вальда

«Улучшить» вид построенной множественной линейной регрессии можно, используя тест Вальда. Он проводится для проверки гипотез равенства коэффициентов какому-либо значению или соотношения коэффициентов между собой. Например, $\beta_2 = 0$ или $\beta_3 = 2 \cdot \beta_4$.

Наиболее распространенным является проверка равенства коэффициентов между собой, то есть одинакового вклада в результат двух или более переменных. Например, если для уравнения $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ принимается гипотеза $H_0: \beta_2 = \beta_4$, то переменные x_2 и x_4 можно объединить (суммировать) в модели, то есть

$$y = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + x_4) + \beta_3 x_3 + \varepsilon.$$

7. Экономическая интерпретация параметров уравнения регрессии


Коэффициент регрессии $\hat{\beta}_i$ при переменной x_i показывает величину прироста зависимой переменной Y при изменении переменной x_i при неизменных других переменных; $\hat{\beta}_i$ показывает, на сколько в среднем увеличится Y при увеличении x_i на единицу.

Построение множественной линейной регрессии в Gretl

Рассмотрим процесс построения уравнения множественной линейно регрессии с использованием пакета *Gretl*.

1. Оценка параметров модели методом наименьших квадратов в *Gretl*.

В *Gretl* оценка линейной модели множественной регрессии

осуществляется аналогично парной регрессии (лаб. работа № 3): необходимо нажать внизу окна программы иконку , либо выбрать в меню **Модель / Метод наименьших квадратов...** В появившемся окне «*Спецификация модели*» в поля «Зависимая переменная» и «Регрессоры» необходимо перенести соответствующие переменные.

2. Оценка значимости параметров регрессии и уравнения регрессии в целом и качества подгонки множественной линейной регрессии к данным.

Осуществляется аналогично парной регрессии (лаб. работа № 3).

В поле **Испр. R-квадрат** приведено значение скорректированного коэффициента детерминации (R_{adj}^2).

3. Проверка на мультиколлинеарность.

В *Gretl* отображение корреляционной матрицы осуществляется следующим способом: необходимо выбрать в меню **Вид / Корреляционная матрица**, далее в открывшемся окне «Корреляция» добавить две или более переменных, для которых ведется расчет, и нажать ОК. В результате появится окно с корреляционной матрицей (рис. 4.1).

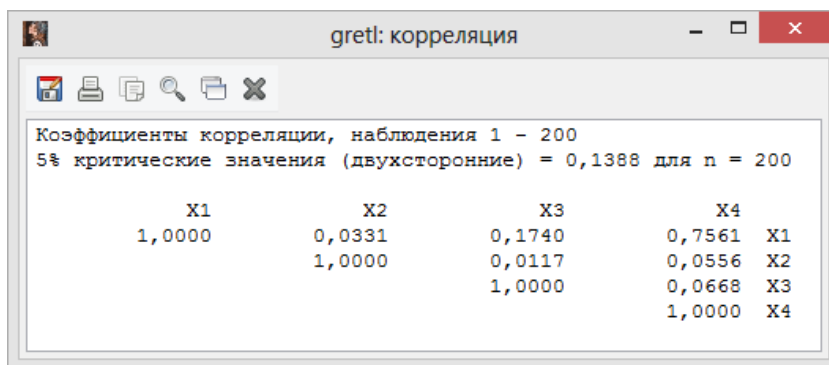


Рис. 4.1. Корреляционная матрица

Дополнительно на мультиколлинеарность можно проверить

методом инфляционных факторов. Для этого в окне *Модель* необходимо выбрать **Тесты / Мультиколлинеарность**. В появившемся окне *Мультиколлинеарность* будет рассчитан коэффициент *VIF*, характеризующий силу мультиколлинеарности.

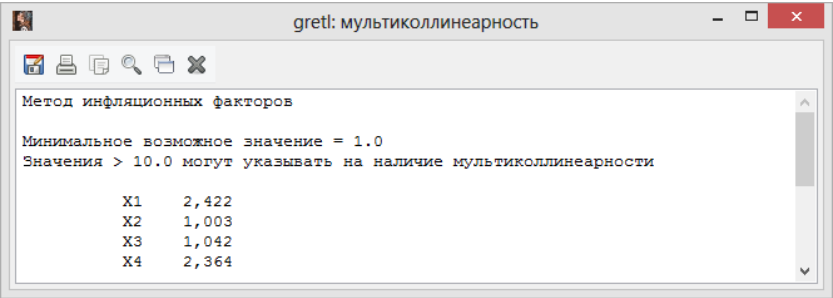


Рис. 4.2. Расчет коэффициента *VIF*

Пример. Для рядов *Y*, *X1*, *X2*, *X3* и *X4* была построена и оценена регрессия $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \varepsilon$ (рис. 4.3).

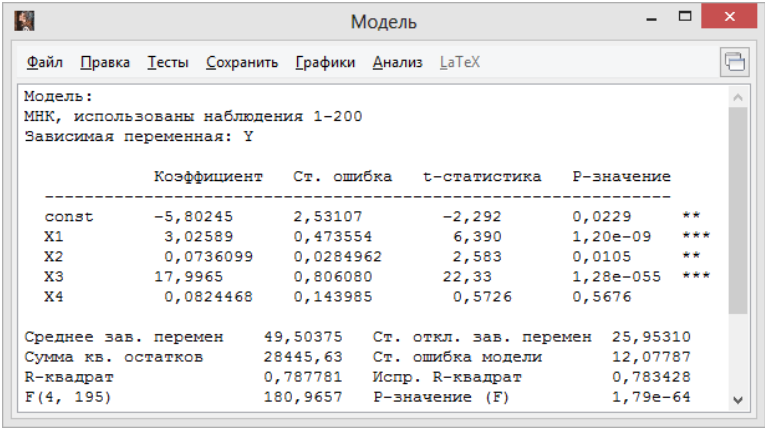


Рис. 4.3. Результаты оценки множественной регрессии

На рис. 4.2 все значения коэффициента *VIF* < 10, что не подтверждает наличие мультиколлинеарности. Однако из рис. 4.1 видно, что *X1* коррелирует с *X4* (коэффициент парной линейной корреляции $r = 0,7561$), следовательно, имеет место мультиколлинеарность.

Из рис. 4.3 следует, что коэффициент β_4 незначим на 10%-м уровне значимости.

Удалим из модели фактор X_4 как менее существенный и коррелирующий со значимым (X_1) фактором. Заметим, что значение скорректированного коэффициента детерминации **Испр. R-квадрат** при этом возросло, что также свидетельствует о несущественности фактора X_4 (рис. 4.4).

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-6,25535	2,40019	-2,606	0,0099	***
X1	3,23131	0,308599	10,47	1,16e-020	***
X2	0,0743862	0,0284151	2,618	0,0095	***
X3	17,9499	0,800581	22,42	5,32e-056	***
Среднее зав. перемен	49,50375	Ст. откл. зав. перемен	25,95310		
Сумма кв. остатков	28493,46	Ст. ошибка модели	12,05715		
R-квадрат	0,787424	Испр. R-квадрат	0,784171		

Рис. 4.4. Результаты оценки после удаления X_4

Если при удалении незначимого фактора значение скорректированного коэффициента детерминации снизится, то, возможно, этот фактор является существенным и исключить нужно какой-то другой фактор.

В итоговой модели (рис. 4.4) все коэффициенты при факторах значимы, между факторами X_1 , X_2 и X_3 нет высокой корреляции (все $r < 0,1$), что позволяет сделать вывод об отсутствии мультиколлинеарности в полученной модели.

4. Проверка на гетероскедастичность.

В *Gretl* тестирование линейной модели множественной регрессии на гетероскедастичность осуществляется следующим способом:

- Осуществляется оценка регрессии обычным МНК.
- Для проверки ошибок на гетероскедастичность в окне *Модель* выбирается **Тесты / Гетероскедастичность / Тест Вай-**

та (или другой тест). Если $P\text{-значение} = P(\text{Хи-квадрат})$ теста меньше 0,05, то гипотеза о гомоскедастичности отвергается (принимается гетероскедастичность) на уровне значимости 5% (рис. 4.5).

Модель 2:
МНК, использованы наблюдения 1-200
Зависимая переменная: Y

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-6,25535	2,40019	-2,606	0,0099	***
X1	3,23131	0,308599	10,47	1,16e-020	***
X2	0,0743862	0,0284151	2,618	0,0095	***
X3	17,9499	0,800581	22,42	5,32e-056	***
Среднее зав. перемен	49,50375	Ст. откл. зав. перемен	25,95310		
Сумма кв. остатков	28493,46	Ст. ошибка модели	12,05715		
R-квадрат	0,787424	Испр. R-квадрат	0,784171		
F(3, 196)	242,0082	P-значение (F)	1,24e-65		
Лог. правдоподобие	-779,6990	Крит. Акаике	1567,398		
Крит. Шварца	1580,591	Крит. Хеннана-Куинна	1572,737		

Тест Вайта (White) на гетероскедастичность -
Нулевая гипотеза: гетероскедастичность отсутствует
Тестовая статистика: LM = 11,3078
p-значение = P(Хи-квадрат(8) > 11,3078) = 0,184862

Рис. 4.5. Тест Вайта на гетероскедастичность

Также в окне *Модель* появится информация, о том, что был проведен соответствующий тест на гетероскедастичность с $p\text{-значением}$.

- Если гетероскедастичность подтверждается, то необходимо выполнить коррекцию стандартных ошибок оценок коэффициентов; для этого в окне *Модель* выбирается **Правка / Изменить модель...**, в появившемся окне *Спецификация модели* ставится галочка напротив поля **Робастные стандартные ошибки**, нажимается ОК (рис. 4.6).

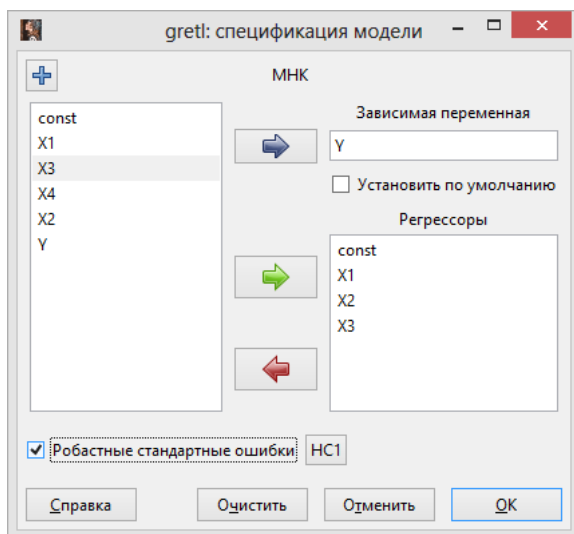


Рис. 4.6. Коррекция на гетероскедастичность

После проведения коррекции стандартных ошибок на гетероскедастичность в окне модели появится указание, что рассчитаны *робастные оценки стандартных ошибок (с поправкой на гетероскедастичность)*.

5. Тест Вальда.

Для проверки равенства коэффициентов в модели необходимо в окне *Модель* выбрать **Тесты / Сумма коэффициентов** и в появившемся окне *Тестирование моделей* выбрать переменные, для которых проверяется гипотеза (рис. 4.7) и нажать *ОК*.

На рис. 4.7 проверяется гипотеза о равном влиянии переменных X2 и X4 на результирующую переменную.

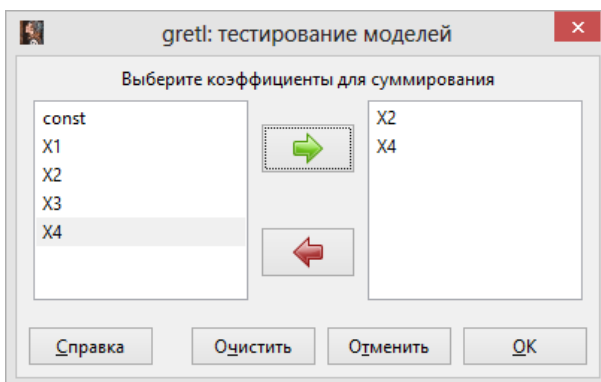


Рис. 4.7. Тест Вальда

В результате проведения теста появится следующее окно (рис. 4.8).

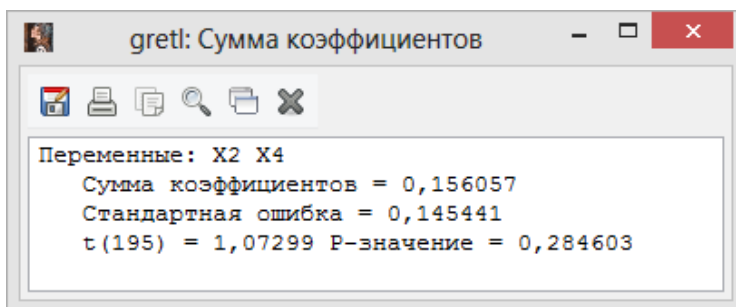


Рис. 4.8. Результаты теста Вальда

Если полученное P -значение $< 0,05$ (рис. 4.8), то гипотеза отвергается на уровне доверия 95%. В данном примере влияние переменных $X2$ и $X4$ на результирующую переменную неравное.

Задания

Требуется исследовать влияние факторов на зависимую переменную путем построения уравнения множественной линейной регрессии. Исходные данные по вариантам находятся в файле *lab 4.xls*.

1. Проведите анализ данных и подготовьте выборку к проведению эконометрического моделирования.

2. Оцените параметры начального уравнения множественной линейной регрессии методом наименьших квадратов.
3. Оцените качество построенной модели (значимость коэффициентов и уравнения в целом, качество подгонки к выборочным данным).
4. Проверьте модель на мультиколлинеарность и при необходимости исключите ее.
5. Проведите тест на гетероскедастичность и при необходимости скорректируйте стандартные ошибки.
6. Предложите свои варианты влияющих величин, являющихся некоторой комбинацией исходных величин, которые бы позволили сделать более качественные выводы по модели (проверьте линейные гипотезы о параметрах уравнения регрессии).
7. Дайте экономическую интерпретацию полученной модели.
8. Сохраните рабочий файл под именем *Фамилия студента_4.gdt*.
9. По результатам исследования оформите отчет.

Вопросы для самоконтроля

1. Чем отличается модель множественной линейной регрессии от парной линейной регрессии?
2. Запишите модель множественной линейной регрессии в матричном виде.
3. Каковы основные предположения относительно модели множественной линейной регрессии?
4. Что утверждает теорема Гаусса-Маркова?
5. Каковы свойства оценок множественной линейной регрессии при выполнении условий теоремы Гаусса-Маркова?
6. Как проверяется значимость параметров регрессии?
7. Как построить доверительный интервал для параметра регрессии?
8. Как проверяются линейные гипотезы о параметрах уравнения регрессии в *Gretl*?
9. Как рассчитывается коэффициент детерминации для множественной линейной регрессии? Каковы его свойства?
10. Что показывает коэффициент детерминации для множественной линейной регрессии?

11. Что такое степень свободы? Как рассчитывается общая дисперсия, скорректированная на число степеней свободы?
12. Как рассчитывается остаточная дисперсия, скорректированная на число степеней свободы?
13. Как рассчитывается факторная дисперсия, скорректированная на число степеней свободы?
14. Как рассчитывается скорректированный коэффициент детерминации?
15. Каковы свойства скорректированного коэффициента детерминации?
16. Как можно использовать скорректированный коэффициент детерминации (наряду с обычным коэффициентом детерминации) при выполнении регрессионного анализа?
17. Что означает значимость уравнения регрессии в целом?
18. Как проверяется значимость уравнения регрессии в целом?
19. Что называется эластичностью зависимой переменной по фактору?
20. Как рассчитать эластичность для факторов линейной модели регрессии?
21. Как записывается уравнение линейной регрессии в стандартизированном виде?
22. Что показывают стандартизированные коэффициенты линейного уравнения регрессии?
23. Как рассчитываются стандартизированные коэффициенты линейного уравнения регрессии?
24. Как сравнить переменные в линейной регрессии по их вкладу в изменение зависимой переменной?
25. Что такое мультиколлинеарность? Каковы ее последствия?
26. Как избавиться от мультиколлинеарности?
27. Что такое гетероскедастичность? Каковы ее последствия?
28. Что изменится в модели при коррекции на гетероскедастичность?
29. Какие вы знаете основные тесты на гетероскедастичность?
30. В чем суть теста Гольдфелда-Квандта?
31. Каковы основные этапы проведения теста Вайта?
32. В чем суть обобщенного метода наименьших квадратов? В каком случае необходимо его использовать?
33. Каковы основные причины неоднородности данных?

34. Какие переменные называются фиктивными? Какие значения они принимают?
35. Как интерпретируются параметры при фиктивных переменных?
36. Для чего проводится тест Вальда?
37. Что такое стохастические регрессоры? Какими свойствами обладают МНК-оценки в случае стохастических регрессоров?

Лабораторная работа № 5.

НЕЛИНЕЙНЫЕ РЕГРЕССИОННЫЕ МОДЕЛИ

Часто на практике между зависимой и независимыми переменными наблюдается нелинейная форма взаимосвязи.

В этом случае существует два выхода:

- 1) подобрать к анализируемым переменным преобразование, которое бы позволило представить существующую зависимость в виде линейной функции;
- 2) применить нелинейный метод наименьших квадратов.

В эконометрическом моделировании (кроме линейной) чаще всего используются следующие виды зависимостей: экспоненциальная, логарифмическая, степенная.

1. Приведение к линейной форме нелинейных регрессионных моделей

1. Экспоненциальная зависимость $y = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon}$.

Если прологарифмировать левую и правую части данного уравнения, то получится

$$\ln y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon.$$

Это уравнение является линейным относительно β_i , для $i = \overline{0, k}$, но вместо y в левой части стоит $\ln y$.

В данном случае параметр β_i имеет следующий экономический смысл: при увеличении переменной x_i на единицу переменная y в среднем увеличится в e^{β_i} раз.

Если $\beta_i < 0,1$, то можно использовать следующую интерпрета-

цию: при увеличении переменной x_i на единицу переменная y в среднем увеличится примерно на $(100 \cdot \beta_i)\%$.

2. Логарифмическая зависимость $y = \beta_0 + \beta_1 \ln x_1 + \dots + \beta_k \ln x_k + \varepsilon$.

Переход к линейному уравнению осуществляется заменой переменных x_i на $X_i = \ln x_i$, $i = \overline{1, k}$.

Параметр β_i имеет следующий экономический смысл: для увеличения y на единицу необходимо увеличить переменную x в

$$e^{\frac{1}{\beta_i}} \text{ раз, т. е. примерно на } \frac{100}{\beta_i} \% .$$

или

при увеличении переменной x_i на 1% y в среднем увеличится примерно на $\frac{\beta_i}{100}$ единиц.

3. Степенная зависимость $y = \beta_0 x_1^{\beta_1} \dots x_k^{\beta_k} \cdot e^\varepsilon$.

Прологарифмировав левую и правую части данного уравнения, получим

$$\ln y = \ln \beta_0 + \beta_1 \ln x_1 + \dots + \beta_k \ln x_k + \varepsilon .$$

Заменив соответствующие ряды их логарифмами, получим линейную регрессию.

Экономический смысл параметра β_i : если значение переменной x_i увеличить на 1%, то y в среднем увеличится на $\beta_i\%$.

Замечания

1. Кроме указанных выше зависимостей при эконометрическом моделировании используются различные их комбинации, например, достаточно часто встречается так называемая **полулогарифмическая зависимость**, в которой часть переменных стоит под знаком логарифма, а часть — нет. К полулогарифмической модели можно отнести модель вида

$$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 x_2 + \varepsilon ,$$

где переменная x_1 находится под знаком логарифма, а остальные переменные (y и x_2) — нет.

2. Не все существующие зависимости путем преобразования

можно привести к линейному виду. В этом случае для получения оценок параметров регрессии приходится использовать более сложный метод оценки — нелинейный метод наименьших квадратов.

2. Пример построения нелинейной регрессионной модели — оценка параметров производственной функции Кобба-Дугласа

Производственная функция Кобба-Дугласа с двумя факторами производства (трудом и капиталом) имеет вид

$$Q = \beta_0 K^{\beta_1} L^{\beta_2},$$

где β_0 , β_1 , β_2 — параметры модели, Q — объем выпуска, K — капитал (стоимость основных фондов), L — затраты труда.

Параметр β_0 представляет собой масштабный множитель и зависит от единиц измерения Q , K и L .

Параметры β_1 и β_2 являются *коэффициентами эластичности выпуска по капиталу и труду* соответственно. Параметр β_1 показывает, на сколько процентов в среднем изменится Q , если капитал увеличить на 1 %. Параметр β_2 показывает, на сколько процентов в среднем изменится выпуск, если затраты труда увеличить на 1 %.

Сумма параметров β_1 и β_2 характеризует отдачу от увеличения масштабов производства.

Если $\beta_1 + \beta_2 = 1$, то *отдача от увеличения масштабов постоянна (эффект масштаба равен 0)*, т. е. увеличение затрат труда и капитала в несколько раз приведет к увеличению выпуска производства в это же число раз (средние издержки на единицу продукции не изменяются с ростом выпуска);

если $\beta_1 + \beta_2 < 1$, то *отдача от масштаба убывает (эффект масштаба отрицательный)*, т. е. с увеличением затрат труда и капитала в несколько раз выпуск увеличится в меньшее число раз (средние издержки, рассчитанные на единицу продукции, растут при увеличении выпуска);

если $\beta_1 + \beta_2 > 1$, то *отдача от масштаба возрастает (эффект масштаба положительный)*, т. е. с увеличением затрат труда и капитала в несколько раз выпуск увеличится в большее число раз (средние издержки, рассчитанные на единицу продук-

ции, падают при увеличении выпуска).

Для оценки параметров функции Кобба-Дугласа используют предварительное ее приведение к линейной зависимости путем логарифмирования:

$$\ln Q = \ln \beta_0 + \beta_1 \cdot \ln K + \beta_2 \cdot \ln L.$$

По имеющимся данным о значениях Q , K , L оценивают (определяют) параметры β_0 , β_1 , β_2 методом наименьших квадратов.

Заметим, что необходимость построения нелинейных регрессий возможна в двух случаях:

1) до проведения исследования известно (из теории), что имеет место нелинейная зависимость, как в примере с функцией Кобба-Дугласа;

2) до проведения исследования неизвестен вид зависимости, поэтому надо найти форму взаимосвязи переменных, наилучшим образом описывающую существующую зависимость.

Во втором случае приходится строить регрессии разных видов (как правило, используются описанные выше зависимости). Полученные регрессии сравниваются между собой и выбирается лучшая. Хорошим подспорьем при выборе вида модели может оказаться графический анализ зависимостей.

3. Сравнение нелинейных моделей

Модели с разными зависимыми переменными (например, y и $\ln y$) некорректно сравнивать, используя *стандартную ошибку регрессии* S_e или *скорректированный коэффициент детерминации* R_{adj}^2 . Для достаточно больших значений переменной всегда $y \gg \ln y$, поэтому S_e во второй модели будет меньше, даже если эта модель хуже. В данном случае корректнее использовать *среднюю абсолютную процентную ошибку прогноза переменной y (оценивающую отклонение фактических значений переменной y от ее прогнозных, рассчитанных по модели значений)*.

Модели с одинаковыми зависимыми переменными, но разными правыми частями (например, линейная и логарифмическая модели) можно сравнивать, используя S_e или R_{adj}^2 .

Построение нелинейной регрессионной модели в Gretl

Тест на правильность функциональной формы

Одним из тестов, позволяющих понять необходимость перехода от линейной модели к нелинейной, является тест Рамсея на правильность линейной спецификации модели (**RESET-тест**). Идея этого теста заключается в том, что если спецификация модели $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$ верна, то добавление нелинейных функций $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$ не должно помогать объяснять y . В частности, можно добавлять степени:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \alpha_2 (\hat{y})^2 + \dots + \alpha_m (\hat{y})^m + \varepsilon. \quad (1)$$

Тестируется гипотеза $H_0: \alpha_2 = \dots = \alpha_m = 0$ с помощью F -статистики. Обычно тест применяется при небольших значениях $m = 2, 3, 4$. Однако он может отвергать нулевую гипотезу не потому, что в истинной модели есть нелинейные члены, а в силу того что в уравнении пропущена переменная, влияние которой частично учтено нелинейными членами в (1).

Проведение данного теста помогает понять необходимость включения нелинейных регрессоров в модель. Так, включение в модель квадрата переменной x может быть обусловлено тем, что влияние x на y в какой-то момент достигает максимума или минимума и форма взаимосвязи меняется. Например, рост заработной платы в зависимости от возраста замедляется ближе к пенсионному возрасту, а с какого-то момента (возраста) начинает снижаться. Поэтому при исследовании зависимости заработной платы (y) от возраста (x) в модель обычно вводится дополнительная переменная x^2 (переменная «возраст» в квадрате), которая позволяет учесть изменение характера влияния возраста при приближении к пенсионному возрасту. Заметим, что введение нелинейных слагаемых должно иметь экономический смысл.

Для проведения **RESET**-теста необходимо в окне *Модель* выбрать в меню **Тесты / Тест Рамсея (RESET)**, в появившемся окне *Тест Рамсея (RESET)* выбрать вариант теста, например «*Квадраты и кубы*». В результате появится окно с результатами теста (рис. 5.1).

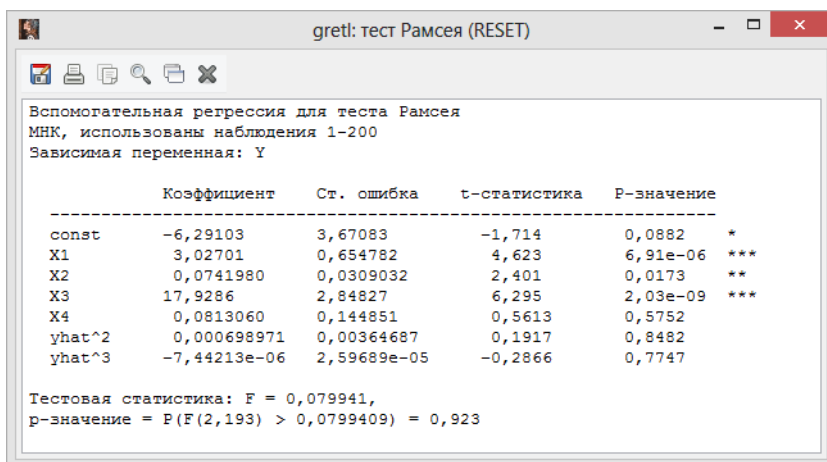


Рис. 5.1. Результаты проведения теста Рамсея

Если $p\text{-значение} = P(F(q,n-k-1)) < 0,05$ (рис. 5.1), то делается вывод о неверной функциональной форме модели (с уровнем доверия 95%), поэтому в уравнение необходимо добавить нелинейные члены (например, $X1^2$) или другие переменные. Например, на рис. 5.1 $p\text{-значение} = P(F(2,193)) = 0,923$, следовательно, функциональная форма модели верна.

Для изменения функциональной формы модели в окне *Модель* выбирается в меню **Правка / Изменить модель...**

Оценка нелинейной регрессионной модели

Построение нелинейных зависимостей, которые в результате определенных преобразований могут быть представлены как линейные, осуществляется в *Gretl* путем добавления логарифмов или квадратов переменных. Для этого необходимо в стартовом окне выделить переменные, для которых рассчитываются логарифмы, затем выбрать в меню **Добавить / Логарифмы выделенных переменных** (или **Квадраты выделенных переменных**).

В результате в стартовом окне появятся новые переменные. Например, для исходной переменной Y:

l_Y – логарифм Y,

sq_Y – квадрат Y.

Набор переменных, необходимых для оценки некоторых нелинейных регрессий, следующий.

1. Экспоненциальное уравнение $y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$.
Зависимая переменная: **l_y**. Регрессоры: **x1** и **x2**.
2. Логарифмическое уравнение $y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$.
Зависимая переменная: **y**. Регрессоры: **l_x1** и **l_x2**.
3. Степенное уравнение $y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} e^{\varepsilon}$.
Зависимая переменная: **l_y**. Регрессоры: **l_x1** и **l_x2**.

Примечание: переменные, содержащие в наблюдениях значения «0», нельзя логарифмировать и брать от них обратную величину. В этом случае *Gretl* автоматически исключит из выборки нулевые значения и выборка существенно сократится, что приведет к искажению результатов.

Задания

Требуется исследовать влияние факторов на зависимую переменную путем построения уравнения множественной нелинейной регрессии.

1. Для построенной при выполнении лабораторной работы № 4 линейной модели множественной регрессии проведите тест Рамсея. Если тест укажет на ошибочность спецификации модели, постройте эконометрические модели, включающие наряду с линейными слагаемыми квадраты независимых переменных или их произведения.
2. Постройте нелинейные регрессионные модели (экспоненциальную, степенную, логарифмическую, полулогарифмическую модели).
3. Для построенных нелинейных моделей проведите тест на гетероскедастичность и при необходимости скорректируйте стандартные ошибки.
4. Дайте экономическую интерпретацию полученных моделей.
5. Сравните построенные модели по скорректированному коэффициенту детерминации, стандартной ошибке регрессии (**когда это возможно**) и/или по величине средней абсолютной процентной ошибки, выберите лучшую модель.
6. Сохраните рабочий файл под именем *Фамилия*

студента_5.gdt.

7. Оформите отчет по результатам исследований.

Вопросы для самоконтроля

1. Всегда ли можно ли применить метод наименьших квадратов для оценки коэффициентов нелинейной регрессии?
2. Приведите пример эконометрической модели, линейной по параметрам, но нелинейной по переменным.
3. Приведите пример эконометрической модели, нелинейной по параметрам, но приводимой к линейной модели регрессии.
4. Приведите пример нелинейной эконометрической модели, которую нельзя преобразовать к линейной модели.
5. Какие преобразования надо выполнить, чтобы привести логарифмическое уравнение регрессии к линейному? Как связаны между собой параметры исходной и полученной в результате преобразования модели?
6. Какие преобразования надо выполнить, чтобы привести степенное уравнение регрессии к линейному? Как связаны между собой параметры исходной и полученной в результате преобразования модели?
7. Какие преобразования надо выполнить, чтобы привести экспоненциальное уравнение регрессии к линейному? Как связаны между собой параметры исходной и полученной в результате преобразования модели?
8. Какие преобразования надо выполнить, чтобы привести гиперболическое уравнение регрессии к линейному? Как связаны между собой параметры исходной и полученной в результате преобразования модели?
9. Каким образом интерпретируются параметры степенной регрессии?
10. Каким образом интерпретируются параметры экспоненциальной регрессии?
11. Каким образом интерпретируются параметры логарифмической регрессии?
12. Каким образом интерпретируются параметры гиперболической регрессии?
13. Каким образом интерпретируется параметр регрессии при квадрате независимой переменной?

14. Какая модель называется полулогарифмической?
15. Какая функция называется функцией Кобба-Дугласа?
16. Что означает возрастающая отдача от масштаба для производственной функции?
17. Что означает убывающая отдача от масштаба для производственной функции?
18. Что означает постоянная отдача от масштаба для производственной функции?
19. Что означает нулевой эффект масштаба для производственной функции?
20. Что означает отрицательный эффект масштаба для производственной функции?
21. Что означает положительный эффект масштаба для производственной функции?
22. Что называется коэффициентом эластичности?
23. Дайте определение индекса детерминации. Каков диапазон его изменения?
24. Как вычисляется индекс корреляции? Чем он отличается от коэффициента парной линейной корреляции?
25. Какой тест помогает выявить ошибочность спецификации модели?
26. Что означает термин "спецификация модели"?
27. Каковы последствия исключения из уравнения регрессии существенной переменной?
28. Каковы последствия включения в уравнение регрессии не-существенной переменной?
29. Чем может быть обусловлена необходимость добавления в модель квадрата независимой переменной?

Лабораторная работа № 6.

ВЫБОР РЕГРЕССИОННОЙ МОДЕЛИ

Выбор регрессоров и вида модели называется *спецификацией модели*. Ранее предполагалось, что мы имеем дело с правильной спецификацией модели, т. е. считалось, что зависимая переменная y , регрессоры x_i и оцениваемые параметры β_i связаны соотношением

$$y_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_i x_{it} + \dots + \beta_k x_{kt} + \varepsilon_t, t = 1, 2, \dots, n$$

и выполняются условия Гаусса — Маркова.

При этом часто утверждается, что такое соотношение описывает «процесс, порождающий данные» или что оно является «истинной моделью». Как правило, на практике истинная модель неизвестна, исследователь оценивает модель, которая лишь приближенно соответствует процессу, порождающему данные. Поэтому возникает естественный вопрос о соотношении между МНК-оценками параметров в истинной и выбранной моделях.

Рассматривается два основных случая.

1. В оцениваемой модели отсутствует часть независимых переменных, имеющих в истинной модели (исключение существенных переменных):

истинная модель — длинная регрессия

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 z_{1t} + \beta_4 z_{2t} + \varepsilon_t;$$

оцениваемая модель — короткая регрессия

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t.$$

В этой ситуации оценка, полученная в короткой регрессии, в общем случае смещенная, но обладает меньшей вариацией.

2. В оцениваемой модели присутствуют независимые переменные, которых нет в истинной модели (включение несущественных переменных):

истинная модель — короткая регрессия

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t;$$

оцениваемая модель — длинная регрессия

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 z_{1t} + \beta_4 z_{2t} + \varepsilon_t.$$

В этом случае оценка $\hat{\beta}$ не смещенная, но дисперсия оценки увеличивается от включения в модель несущественных переменных.

Основным критерием выбора вида модели является ее экономический смысл, а не качество подгонки модели под выборочные данные.

Если в процессе моделирования было получено две модели, качественно описывающие исследуемый процесс, то проводятся

различные тесты на выбор лучшей модели.

1) *F-тест.*

Пусть имеется две модели, которые содержат как одинаковые, так и различные правые части.

$$\text{модель } A: y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \varepsilon_t;$$

$$\text{модель } B: y_t = \gamma_0 + \gamma_1 x_{1t} + \gamma_2 x_{2t} + \gamma_3 z_{1t} + \gamma_4 z_{2t} + v_t.$$

Рассматриваются две регрессии:

$$y_t = \gamma_0 + \gamma_1 x_{1t} + \gamma_2 x_{2t} + \gamma_3 z_{1t} + \gamma_4 z_{2t} + \delta_{1A} x_{3t} + \delta_{2A} x_{4t} + v_t;$$

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \delta_{1B} z_{1t} + \delta_{2B} z_{2t} + \varepsilon_t$$

Далее проверяются две гипотезы.

$H_0: \delta_{1A} = \delta_{2A} = 0$. Если эта гипотеза не отвергается, то и не отвергается модель B .

$H_0: \delta_{1B} = \delta_{2B} = 0$. Если эта гипотеза не отвергается, то и не отвергается модель A .

Если обе гипотезы либо принимаются, либо отвергаются, то ситуация остается неопределенной.

2) *РЕ-тест.*

Применяется, когда в левой части сравниваемых уравнений – различные зависимые переменные, например:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t,$$

$$\ln y_t = \gamma_0 + \gamma_1 \ln x_{1t} + \gamma_2 \ln x_{2t} + u_t.$$

Коэффициент детерминации R^2 здесь не может применяться для сравнения и выбора лучшей модели, т. к. левые части уравнений различны.

Содержательный смысл *РЕ*-теста заключается в проверке, улучшится ли модель при включении в нее прогноза конкурирующей модели.

Реализация *РЕ*-теста состоит в следующем: оцениваются обе модели МНК и получаются соответствующие прогнозные значения \hat{y}_t и $\ln \hat{y}_t$. Далее оцениваются модели

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \delta_{LIN} \left(\ln \hat{y}_t - \ln y_t \right) + v_t,$$

$$\ln y_t = \gamma_0 + \gamma_1 \ln x_{1t} + \gamma_2 \ln x_{2t} + \delta_{LOG} \left(\hat{y}_t - e^{\ln \hat{y}_t} \right) + w_t.$$

Затем тестируются следующие гипотезы.

$H_0: \delta_{LIN} = 0$. Если эта гипотеза не отвергается, то и не отвергается линейная модель.

$H_0: \delta_{LOG} = 0$. Если эта гипотеза не отвергается, то и не отвергается полулогарифмическая модель.

Если обе гипотезы либо принимаются, либо отвергаются, то ситуация остается неопределенной.

РЕ-тест может применяться в значительно более общей ситуации.

Выбор регрессионной модели в Gretl

РЕ-тест.

Пусть имеется две значимых модели:

$$y = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2,$$

$$\ln y = \hat{\beta}_0 + \hat{\beta}_1 \ln x_1 + \hat{\beta}_2 \ln x_2,$$

где $\ln y$, $\ln x_1$, $\ln x_2$ — логарифмы рядов y , x_1 и x_2 соответственно.

Для проведения *РЕ*-теста выполняются следующие шаги.

1. Строятся прогнозные значения для y и $\ln y$.

Для линейной модели (y) в окне *Модель* выберите в меню **Сохранить / Расчетные значения** и в появившемся окне в поле *Название переменной* введите имя прогноза, например *yf*. Нажмите *ОК*. В результате в рабочем файле появится прогнозный ряд *yf*.

Для логарифмической модели ($\ln y$) процедура построения прогнозных значений (*lyf*) проводится аналогично.

2. Оцениваются модели

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \delta_{LIN} \left(\ln \hat{y} - \ln \hat{y} \right) + v,$$

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \delta_{LOG} \left(\hat{y} - e^{\ln \hat{y}} \right) + w.$$

Для этого для линейной модели добавляется логарифм переменной *yf*, который назовем *l_yf*, затем добавляется переменная

разности $lin = l_Yf - lYf$ (рис. 6.1).

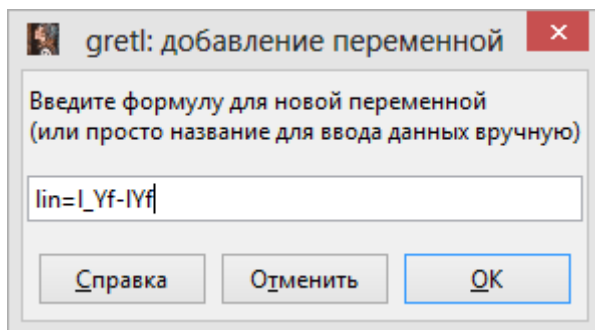


Рис. 6.1. Добавление тестовой переменной для линейной модели

Затем оценивается линейная модель с добавлением переменной lin .

Аналогичная процедура проводится для логарифмической модели:

- рассчитывается экспонента переменной $lyf : elYf = \exp(lYf)$,
 - добавляется переменная $log = yf - elYf$,
 - оценивается логарифмическая модель с добавлением переменной log .
3. Оценивается значимость коэффициентов при добавленных регрессорах в обеих моделях (рис. 6.2).

Если оба коэффициента значимы (рис. 6.2) или оба незначимы, то ситуация неопределенная и выбирать модель необходимо другими методами.

Если в одной модели коэффициент при добавленном регрессоре значим, а в другой — незначим, то лучшей считается модель с незначимым коэффициентом.

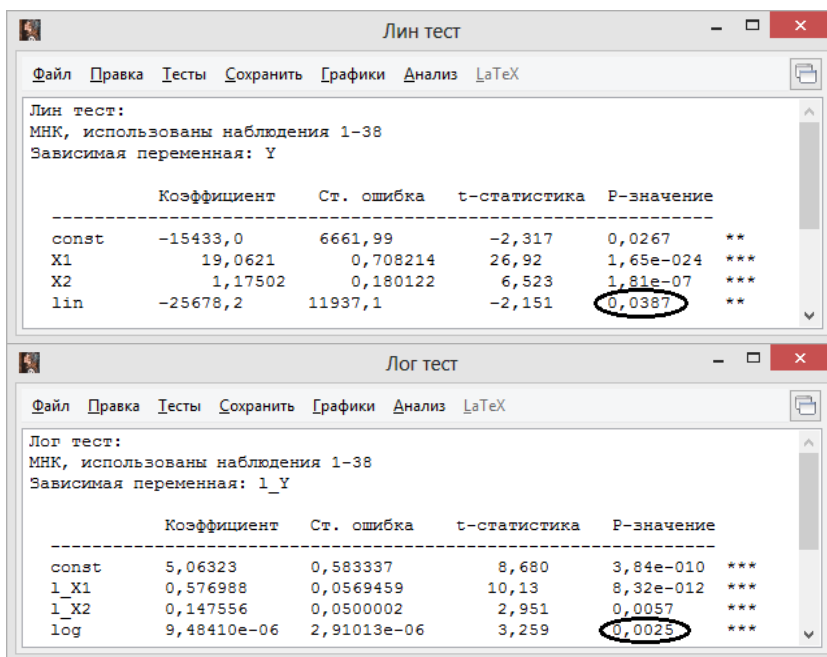


Рис. 6.2. Результаты РЕ-теста

Задания

Используя построенные итоговые модели в лабораторных работах №4 и №5, выполните исследование на выбор лучшей модели.

1. Парно сравните все модели, самостоятельно определив для каждого случая корректный способ выбора лучшей модели (коэффициент детерминации, ошибка аппроксимации, РЕ-тест).
2. Интерпретируйте полученные результаты оценки лучшей модели.
3. Сохраните рабочий файл под именем *Фамилия студента_6.gdt*.
4. Оформите отчет по результатам исследований.

Вопросы для самоконтроля

1. Что означает термин «спецификация модели»?
2. Каковы последствия исключения из уравнения регрессии существенной переменной?
3. Каковы последствия включения в уравнение регрессии не-существенной переменной?
4. Что такое «Замещающая переменная»?
5. Назовите основные критерии для включения переменной в модель.
6. Почему для выбора двух конкурирующих моделей с разными зависимыми переменными не может использоваться R^2 ?
7. Почему в PE -тесте, в случае если гипотеза $H_0: \delta_{\text{LIN}} = 0$ не отвергается, то не отвергается и линейная модель?
8. Почему в PE -тесте, в случае если гипотеза $H_0: \delta_{\text{LOG}} = 0$ не отвергается, то не отвергается и логарифмическая модель?
9. Почему в PE -тесте, если оба коэффициента при добавленных переменных значимы (рис. 6.2), то ситуация неопределенная?
10. Почему в PE -тесте, если оба коэффициента при добавленных переменных незначимы, то ситуация неопределенная?

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Аистов, А. В. Эконометрика шаг за шагом : учеб. пособие для вузов / А. В. Аистов, А. Г. Максимов. М. : Изд. дом ГУ ВШЭ, 2006. – 180 с.
2. Берндт, Э. Р. Практика эконометрики: классика и современность / Э. Р. Берндт. М.: ЮНИТИ-ДАНА, 2013. – 848 с.
3. Бигильдеева, Т. Б. Эконометрика / Т. Б. Бигильдеева, Е. А. Постников. Челябинск: Челяб. гос. ун-т, 2007. – 110 с.
4. Бородич, С. А. Эконометрика. Минск: Новое знание, 2006. — 407 с.
5. Бородич, С. А. Эконометрика. Практикум [Текст]: учебное пособие / С. А. Бородич. —М.: ИНФРА-М, 2014. — 329 с.
6. Валентинов, В. А. Эконометрика / В. А. Валентинов. М. : Дашков и К°, 2012. – 448 с.
7. Доугерти К. Введение в эконометрику: Учебник. 3-е изд. / К. Доугерти. Пер. с англ. М.: ИНФРА-М, 2009. – 466 с.
8. Елисеева, И. И. Практикум по эконометрике / И. И. Елисеева. М.: Финансы и статистика, 2007. – 344 с.
9. Кремер, Н. Ш. Эконометрика. Учебник / Н. Ш. Кремер, Б. А. Путко. М.: ЮНИТИ-ДАНА, 2010. – 328 с.
10. Колемаев, В. А. Эконометрика / В. А. Колемаев. М.: ИНФРА-М, 2010. – 160 с.
11. Магнус, Я. Р. Эконометрика. Начальный курс: учебник / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. М.: Дело, 2008 – 576.
12. Носко, В. П. Эконометрика / В. П. Носко. М.: Издательский дом "Дело" РАНХиГС, 2011.
13. Эконометрика. Учебник для бакалавров / под ред. И.И.Елисеевой. М.: Проспект, 2014. – 288 с.
14. Официальный сайт GRETЛ (Gnu Regression, Econometrics and Time-series Library) <http://gretl.sourceforge.net/ru.html>
15. Международный эконометрический журнал на русском языке «Квантиль» <http://quantile.ru/>

Лабораторный практикум

**Татьяна Борисовна БИГИЛЬДЕЕВА
Евгений Анатольевич ПОСТНИКОВ**

ЭКОНОМЕТРИКА

Регрессионный анализ с использованием пакета Gretl

Редактор Т. Г. Марчевская

Подписано в печать 29.10.14.

Формат 60×84/16. Бумага офсетная.

Печать трафаретная. Усл. печ. л. 2,23. Уч.-изд. л. 2,23.

Тираж 70 экз. Заказ № 2910-557.

Отпечатано в типографии

«Центр Научного Сотрудничества»

ООО «БизнесКом»

454048, г. Челябинск, ул. Воровского, 50 Б

Тел.: (351) 215-12-23; e-mail: cns74@mail.ru