Министерство образования и науки Российской Федерации
Федеральное агентство по образованию

Государственное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

Кафедра математических методов и моделей в экономике

Е. Н. Седова, О. С. Чудинова

ЛИНЕЙНАЯ МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ В ПАКЕТЕ GRETL

Методические указания к лабораторному практикуму и
самостоятельной работе студентов

Рекомендовано к изданию Редакционно-издательским советом Государственного
образовательного учреждения высшего профессионального образования
«Оренбургский государственный университет»

Оренбург
ИПК ГОУ ОГУ
2010

УДК 519.237.5 (076.5)
ББК 22.172я7
С28

Рецензент – доктор экономических наук, профессор Е. Г. Чмышенко

Седова, Е. Н.
С28 Линейная модель множественной регрессии в пакете GRETL : методические указания к лабораторному практикуму и самостоятельной работе студентов / Е.Н. Седова, О.С. Чудинова; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2010. – 46 с.

Методические указания содержат описание работы по оцениванию линейной регрессионной модели, ее исследованию на мультиколлинеарность и варианты индивидуальных заданий для проведения лабораторной работы. Методические указания предназначены для студентов экономических специальностей, изучающих такие дисциплины, как «Эконометрика», «Методы и модели в экономике» и др.

УДК 519.237.5 (076.5)
ББК 22.172я7

Содержание

Введение

Решение многих экономических задач требует привлечения аппарата регрессионного анализа. Это, например, изучение зависимости спроса на товар от его цены, характеристик и места продажи; изучение влияния процентной ставки на объем выдаваемых кредитов; построение производственных функций и др.

Проведение регрессионного анализа требует достаточного большого объема расчетов и проверки ряда гипотез, которые не реализованы в широко распространенных офисных пакетах MS Excel и OpenOffice. Кроме того, изучая реальные объекты и процессы, исследователь рано или поздно столкнется с проблемой существования тесных корреляционных зависимостей между изучаемыми переменными, а это часто приводит к неустойчивости модели и делает ее непригодной для использования. Между тем методы проверки наличия и особенно методы устранения мультиколлинеарности в офисных приложениях не реализованы.

Все перечисленное требует использования специализированных эконометрических пакетов типа Statistica, Eviews, Stata, распространяемых, однако, по платной лицензии. Предлагаемые методические указания демонстрируют проведение множественного регрессионного анализа на базе свободно распространяемого профессионального кросс-платформенного пакета GRETL. Цель методических указаний заключается в выработке практических навыков оценивания параметров и исследования линейных регрессионных моделей, в том числе в условиях мультиколлинеарности.

1 Описание лабораторной работы

Лабораторная работа включает в себя следующие этапы:

- постановку задачи;
- ознакомление с порядком выполнения работы;
- выполнение расчетов индивидуальных задач на компьютере и анализ результатов;
- подготовку письменного отчета с выводами по работе;
- защиту лабораторной работы.

2 Постановка задачи

По данным Приложения А:

1) построить МНК-оценки коэффициентов линейной модели множественной регрессии и провести ее анализ;

2) провести анализ построенной модели на мультиколлинеарность;

3) устранить мультиколлинеарность методом пошаговой регрессии.

3 Порядок выполнения работы

Целью проводимого исследования является изучение регрессионной зависимости ввода в действие жилых домов, построенных населением за свой счет и с помощью кредитов (у, кв. м), от ряда факторов:

$$\tilde{y} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

где $\tilde{y}$ – условное среднее значение ввода в действие жилых домов, соответствующее текущим значениям x_1, x_2, x_3, x_4 , x_5 (кв. м);

x_1 – инвестиции, направленные в жилищное хозяйство, на душу населения (руб.);

x_2 – обеспеченность населения собственными легковыми автомобилями в расчете на 1000 населения (штук);

x_3 – среднесписочная численность работников (человек);

x_4 – фонд оплаты труда работников (млн. руб.);

x_5 – среднемесячная начисленная заработная плата работников (руб.).

Объектом исследования выступают районы Оренбургской области. Предметом исследования – взаимосвязи между вводом в действие жилых домов и указанными экономическими показателями. Информационная база представлена данными о значениях соответствующих показателей для 35 районов Оренбургской области за 2007 год. Таким образом, для оценки линейной функции множественной регрессии взята выборка объемом n=35. Результаты наблюдений над результативным признаком представлены вектором $Y = (y_1,..., y_n)^T$ и матрицей X типа «объект-свойство» наблюдаемых значений признаков $x_1, ...x_k$:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ . & . & . & ... & . \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix}$$

здесь x_{ij} – значение j-го признака на i-м объекте наблюдения; столбец из «1» можно считать столбцом «наблюденных» значений для признака $x_0^0 = 1$

Для оценки линейной функции (уравнения) множественной регрессии используется математическая модель:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \,,\ i = \overline{1, n}$$

где $\varepsilon_i = y_i - \tilde{y}_i$ - регрессионные остатки, характеризующие расхождение между наблюдаемым значением y_i и «осредненным» значением $\tilde{y}_i$ (значением линейной функции регрессии) и, учитывающие влияние всех прочих факторов, не включенных регрессионную модель.

Оценку коэффициентов β уравнения регрессии будем искать методом наименьших квадратов из принципа минимума суммы квадратов отклонений наблюдаемых значений y_i от «значений» функции регрессии:

$$F = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{in})^2 \to \min.$$

В результате решения данной оптимизационной задачи получаем, что интересующие нас оценки есть:

$$b_{МНК} \equiv b = (X^T X)^{-1} X^T Y$$

В рамках классической линейной модели множественной регрессии относительно регрессионных остатков и объясняющих переменных в дальнейшем предполагается выполнение пяти условий (условия Гаусса – Маркова):

1) $x_1, \ldots, x_к$ – детерминированные переменные;

2) ранг матрицы X равен "$к+1$" – отсутствие линейно зависимых признаков;

3) $M\varepsilon_i = 0$, $i = \overline{1,n}$ - отсутствие систематических ошибок в измерении y;

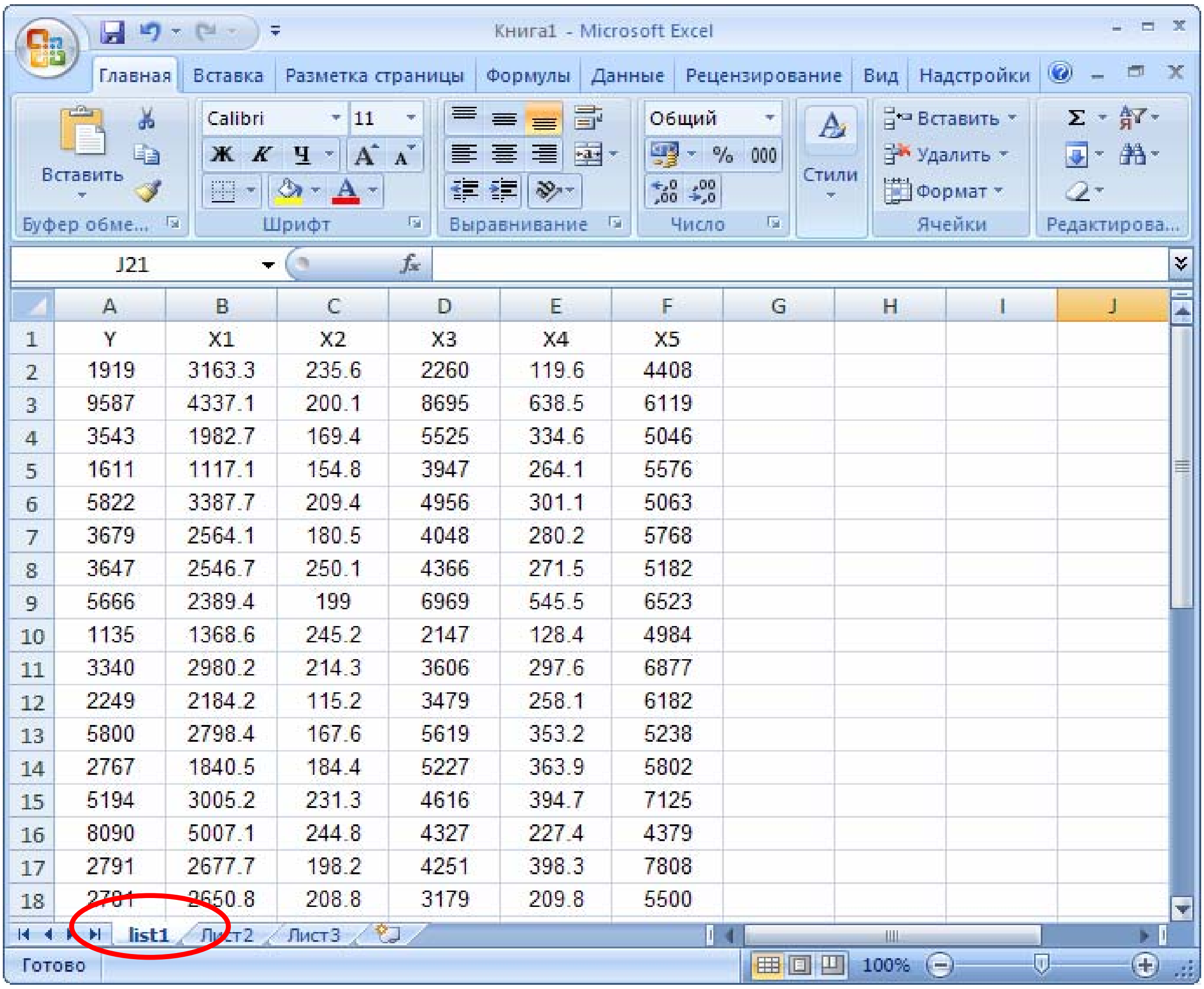
4) $D\varepsilon_i = M\varepsilon_i^2 = \sigma^2$, $i = \overline{1,n}$ - гомоскедастичность регрессионных остатков (равноточность измерений);

5) $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i \cdot \varepsilon_j) = 0$, $i \neq j$, $i = \overline{1,n}$ $j = \overline{1,n}$ - некоррелированность регрессионных остатков.

При выполнении условий Гаусса-Маркова линейная модель множественной регрессии называется классической, а полученные методом наименьших квадратов оценки будут несмещенными, состоятельными и эффективными.

Эконометрический пакет GRETL позволяет как создавать рабочие файлы сразу в своем формате и вводить данные непосредственно с клавиатуры, так и осуществлять импорт данных из большинства распространенных офисных и специализированных статических пакетов – Eviews, Stata, SPSS, SAS. Покажем, как импортировать данные из Excel.

Создадим новую книгу Excel и на ее первом листе введем (или скопируем) данные, причем в первой строке сразу укажем названия анализируемых переменных. Переименуем лист с данными из Лист1 в list1 (для этого нужно щелкнуть левой клавишей мыши на ярлычке листа и ввести новое имя), сохраним файл под именем kniga1.xls, используя пункт меню Файл – Сохранить как (рисунок 1).



	A	B	C	D	E	F
1	Y	X1	X2	X3	X4	X5
2	1919	3163.3	235.6	2260	119.6	4408
3	9587	4337.1	200.1	8695	638.5	6119
4	3543	1982.7	169.4	5525	334.6	5046
5	1611	1117.1	154.8	3947	264.1	5576
6	5822	3387.7	209.4	4956	301.1	5063
7	3679	2564.1	180.5	4048	280.2	5768
8	3647	2546.7	250.1	4366	271.5	5182
9	5666	2389.4	199	6969	545.5	6523
10	1135	1368.6	245.2	2147	128.4	4984
11	3340	2980.2	214.3	3606	297.6	6877
12	2249	2184.2	115.2	3479	258.1	6182
13	5800	2798.4	167.6	5619	353.2	5238
14	2767	1840.5	184.4	5227	363.9	5802
15	5194	3005.2	231.3	4616	394.7	7125
16	8090	5007.1	244.8	4327	227.4	4379
17	2791	2677.7	198.2	4251	398.3	7808
18	2781	2650.8	208.8	3179	209.8	5500

Рисунок 1 – Исходные данные в Excel

Памятка по импорту данных в GRETL из Excel

1. Имена переменных, листы рабочих книг и сами книги называйте ЛАТИНСКИМИ буквами. Старайтесь не использовать символы национальных алфавитов.
2. Импортируемые данные должны находиться на первом из листов рабочей книги (если их несколько).
3. На листе с импортируемыми данными не должно находиться **никаких** других посторонних объектов (графиков, рисунков и т.д.).

Запустим GRETL. После запуска на экране откроется основное окно программы. Выберем пункт главного меню **Файл – Открыть – Импорт – Excel** (рисунок 2).

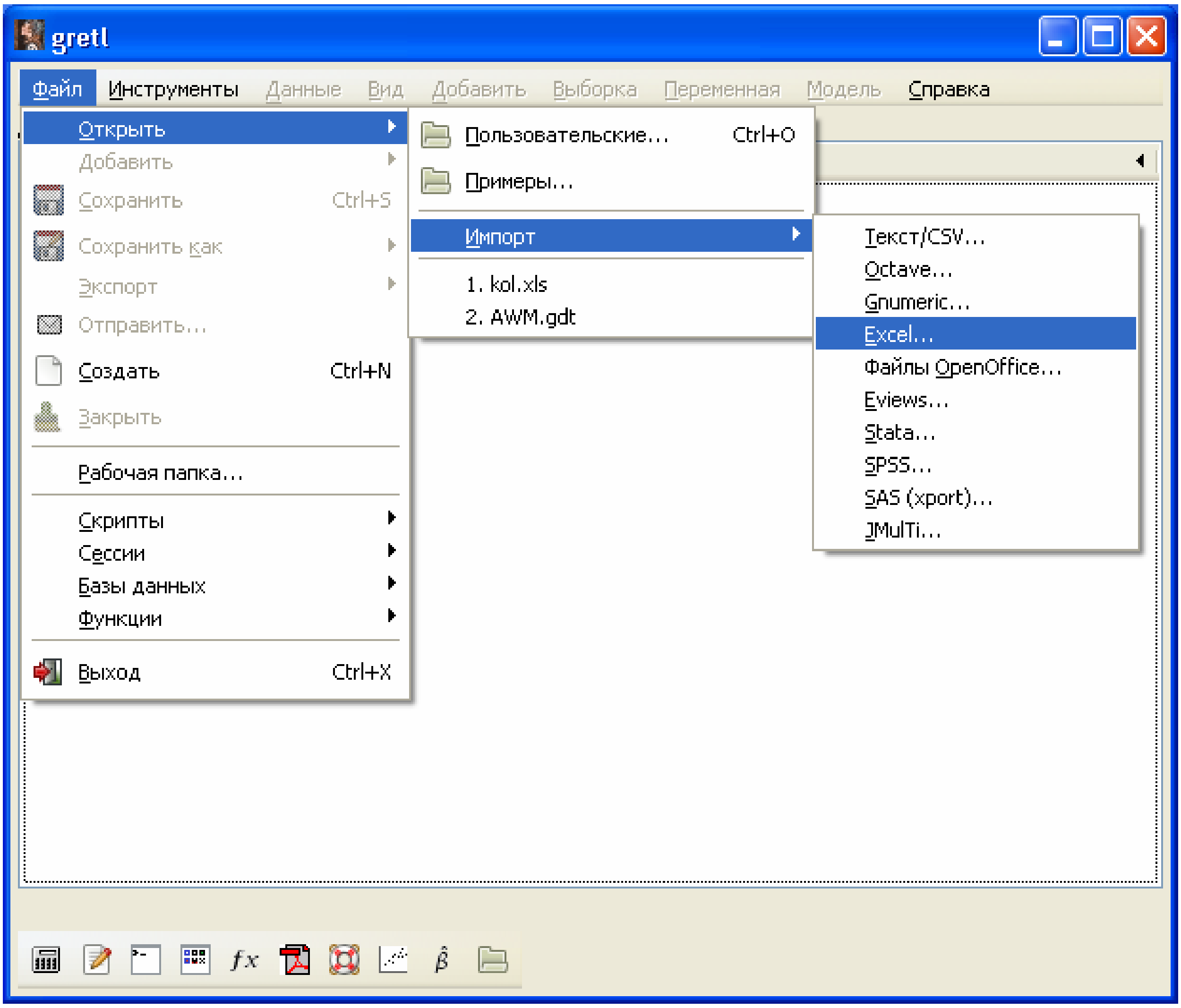


Рисунок 2 – Стартовое окно пакета GRETL

В появившемся окне укажем путь к файлу с данными, осуществляя навигацию с помощью списка слева (рисунок 3).

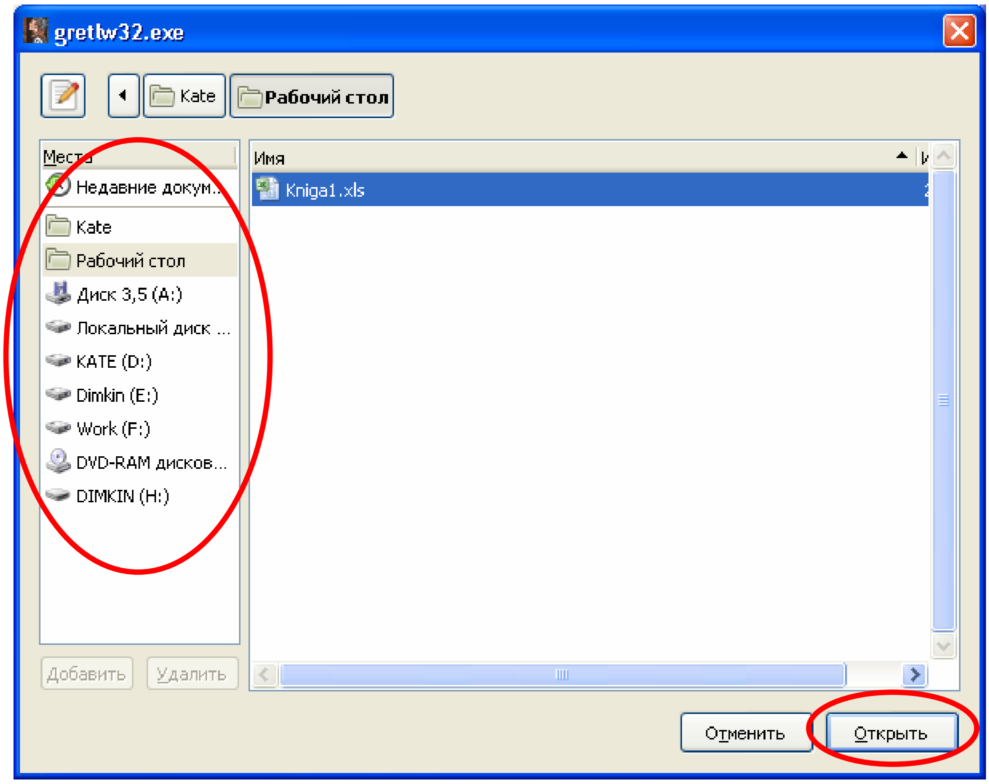


Рисунок 3 – Указание пути к файлу с данными

После нажатия на кнопку 

появится окно, в котором нужно указать номер первого столбца и первой строки массива с данными. В блоке **Лист для импорта** указывается, с какого листа открываемой книги будут считываться данные – как и требовалось, с листа list1 (рисунок 4).

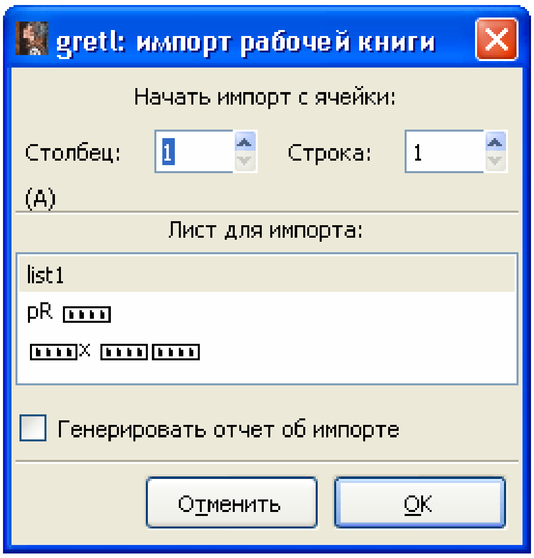


Рисунок 4 – Задание номера первого столбца и строки диапазона с данными

Поскольку все параметры диапазона данных заданы верно, нажмем кнопку 

. На экране появится окно с запросом о возможности задания структуры данных в виде временного ряда или панельных данных (рисунок 5). В нашем случае мы имеем данные пространственной структуры (множество объектов в один момент времени), поэтому нажмем кнопку 

.

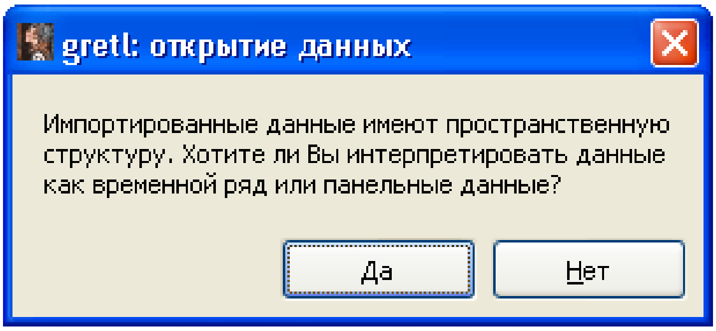


Рисунок 5 – Вид окна запроса структуры данных

На экране появится основная форма GRETL с переменными Y, X1, X2, X3, X4, X5 (в соответствии с заданными именами переменных в Excel).

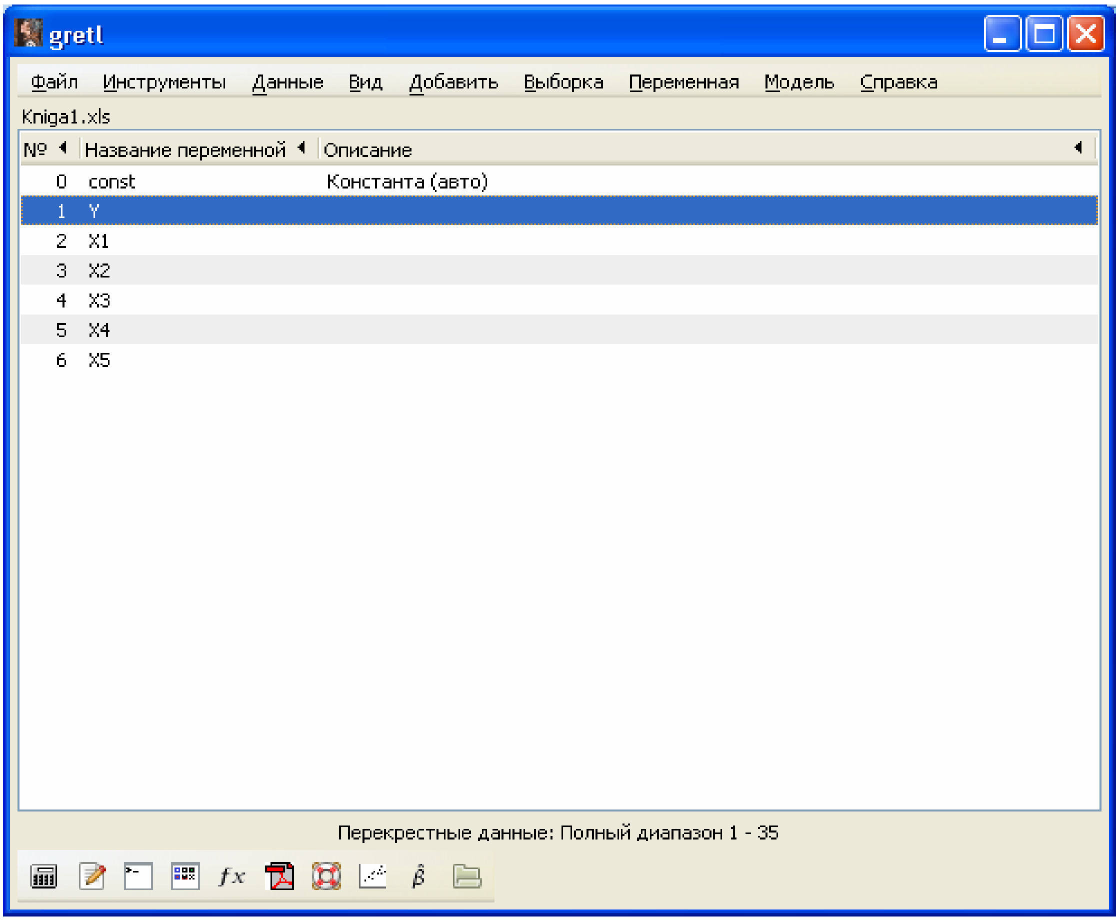
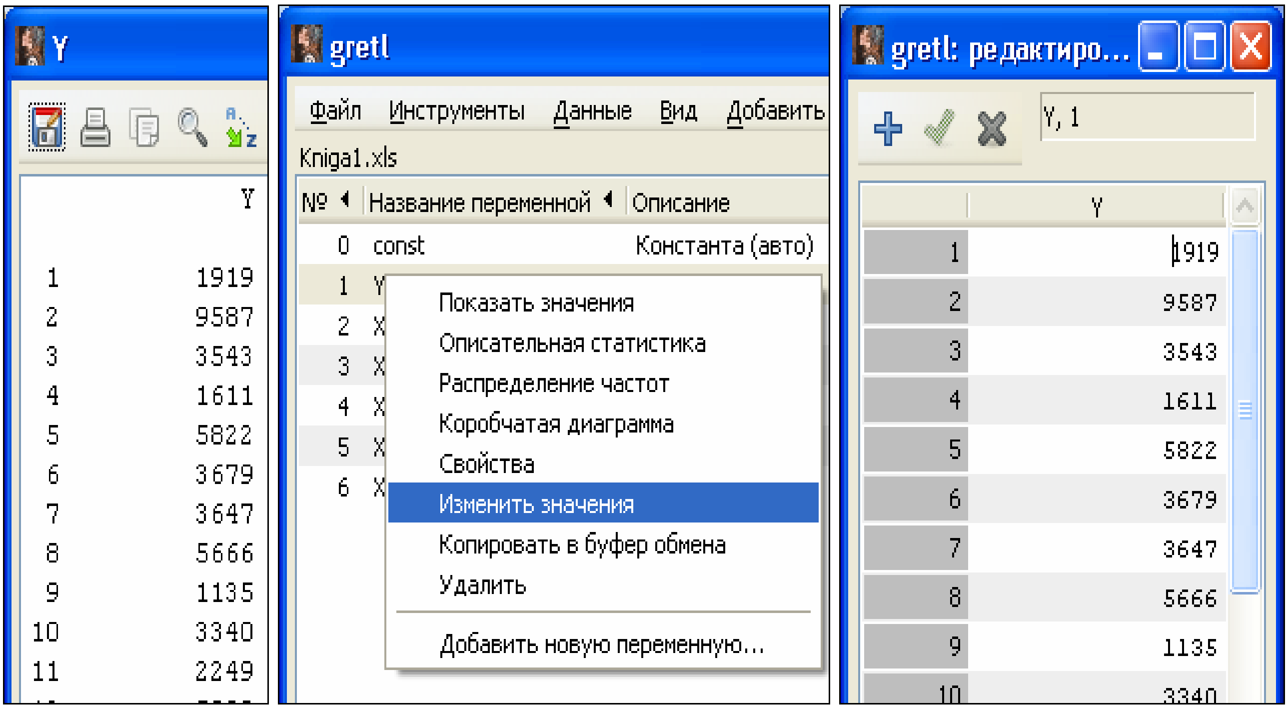


Рисунок 6 – Вид окна GRETL после импорта данных

Для просмотра значений любой переменной нужно выделить ее и сделать или двойной щелчок левой клавишей мыши, или щелчок правой клавишей мыши с выбором пункта контекстного меню **Показать значения**. Появится окно, как на рисунке 7а. При необходимости **изменения значений** нужно сделать щелчок правой клавишей мыши на переменной и выбрать соответствующий пункт контекстного меню (рисунок 7б), далее внести изменения в форму вида, представленного на рисунке 7в.



(а) (б) (в)

Рисунок 7 – Просмотр и изменение значений переменной

Если нужно увеличить количество столбцов (переменных), то используем пункты меню **Добавить – Добавить новую переменную**, если нужно увеличить количество строк (объектов), то **Данные – Добавить наблюдения**.

При работе с реальными данными можно столкнуться с ситуацией, когда не для всех объектов есть значения всех признаков. Возникает так называемая проблема «пропущенных значений», для решения которой нужно привлекать специальные методы. Если число объектов и/или признаков невелико, то обнаружить пропущенные значения достаточно просто, однако с увеличением объема выборки или количества анализируемых переменных задача усложняется. Пакет GRETL имеет функцию автоматического обнаружения и подсчета числа пропусков: пункт меню **Данные** подпункт **Подсчитать пропуски**. В нашем случае пропусков нет (рисунок 8).

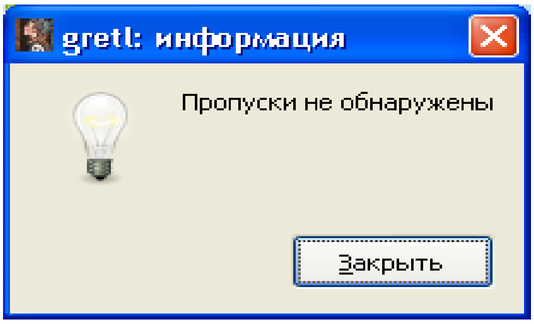


Рисунок 8 – Вид информационного окна команды подсчета пропусков

Сохраним импортированные данные в формате GRETL. Для этого выберем **Файл – Сохранить как – Файл gretl** (рисунок 9).

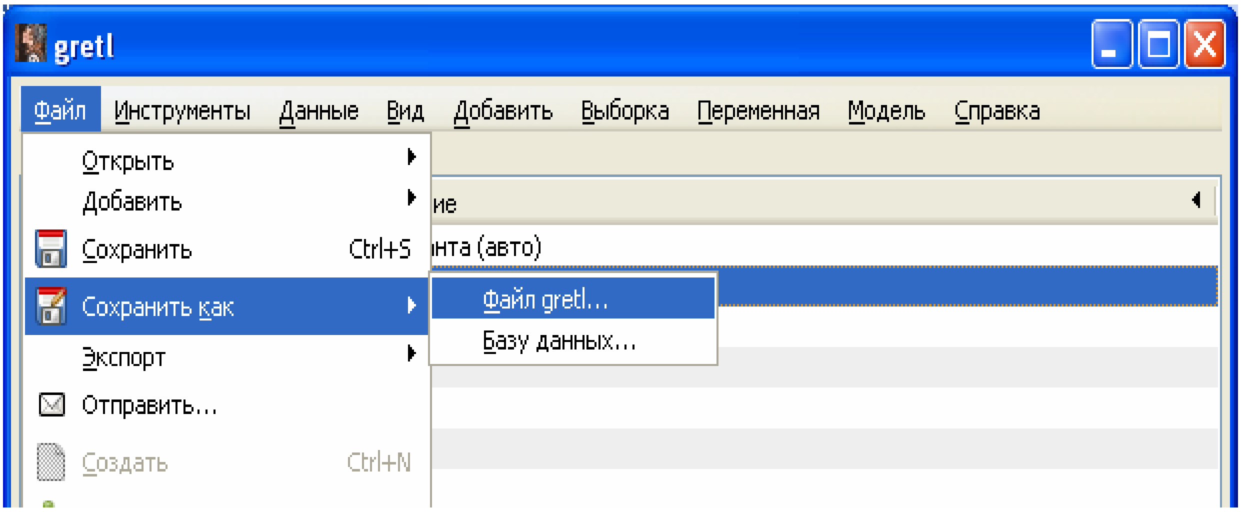
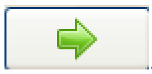


Рисунок 9 – Выбор пункта меню для сохранения данных в формате gretl

В появившемся окне следует указать переменные, которые нужно сохранить. Для этого они выделяются в списке **Доступные переменные** и переносятся в список **Выбранные переменные** с помощью кнопки с зеленой стрелкой . Для одновременного переноса всех переменных нажимается кнопка 

. Если решение о необходимости сохранения переменной изменилось, то соответствующая переменная удаляется из списка **Выбранные переменные** с помощью кнопки с красной стрелкой . Вид окна представлен на рисунке 10.

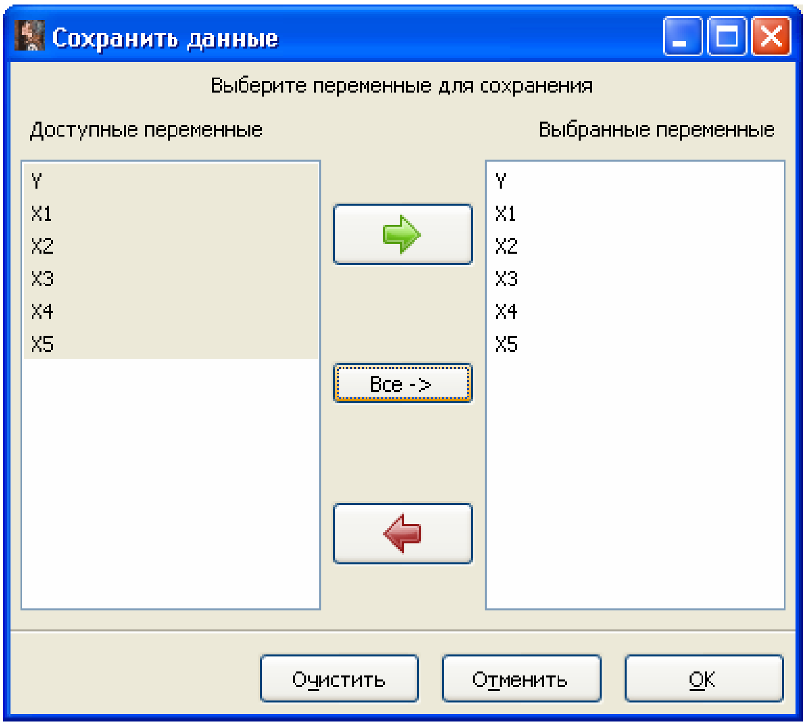


Рисунок 10 – Выбор переменных для сохранения

После нажатия  на экране появится окно, в котором задается имя и расположение будущего файла (рисунок 11).

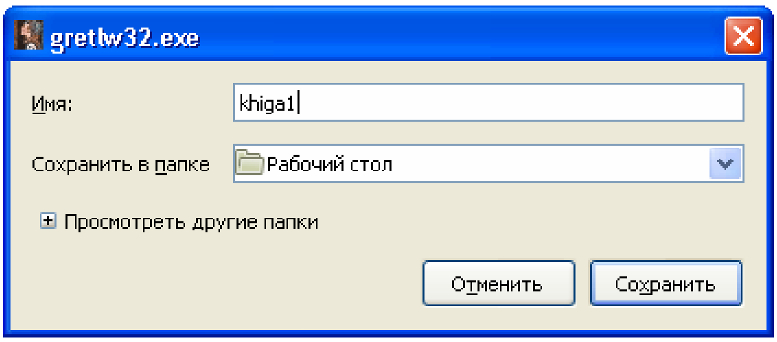
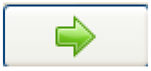


Рисунок 11 – Задание имени и расположения сохраняемого файла

Нажмем на кнопку .

Оценивание параметров линейной модели множественной регрессии

Для построения уравнения множественной регрессии в меню системы выберем **Модель - Метод наименьших квадратов.** На экране появится окно, в котором необходимо указать зависимую (результирующую, объясняемую) и независимые (объясняющие) переменные для анализа.

Для задания зависимой переменной выберем в списке слева переменную Y и нажатием на кнопку с СИНЕЙ стрелкой  перенесем ее в окно **Зависимая переменная**, аналогично, выделим переменные X1, X2, X3, X4, X5 и кнопкой с ЗЕЛЕНОЙ стрелкой  перенесем их в список **Независимые переменные** (рисунок 12). Выбор нескольких несмежных переменных производится при нажатой клавише **CTRL.**

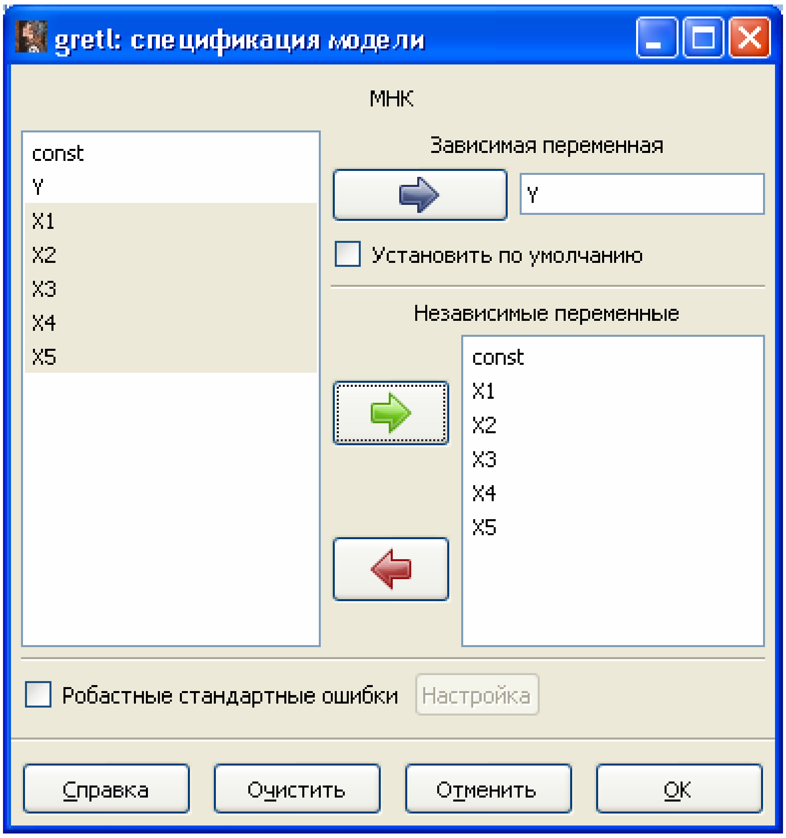


Рисунок 12 – Выбор переменных для проведения регрессионного анализа

После выбора переменных необходимо щелкнуть на кнопке . На экране появится окно с результатами (рисунок 13), структуру которого разберем подробно. В верхней информационной части окна выводится номер оцениваемой модели, количество наблюдений, использованных в анализе, указывается имя переменной, выступившей в качестве зависимой. Ниже выводятся оценки коэффициентов модели регрессии (первый столбец *Коэффициент*), стандартные ошибки коэффициентов (второй столбец *Ст. ошибка*), значение t-статистики (третий столбец) и достигаемый уровень значимости (четвертый столбец *P-значение*).

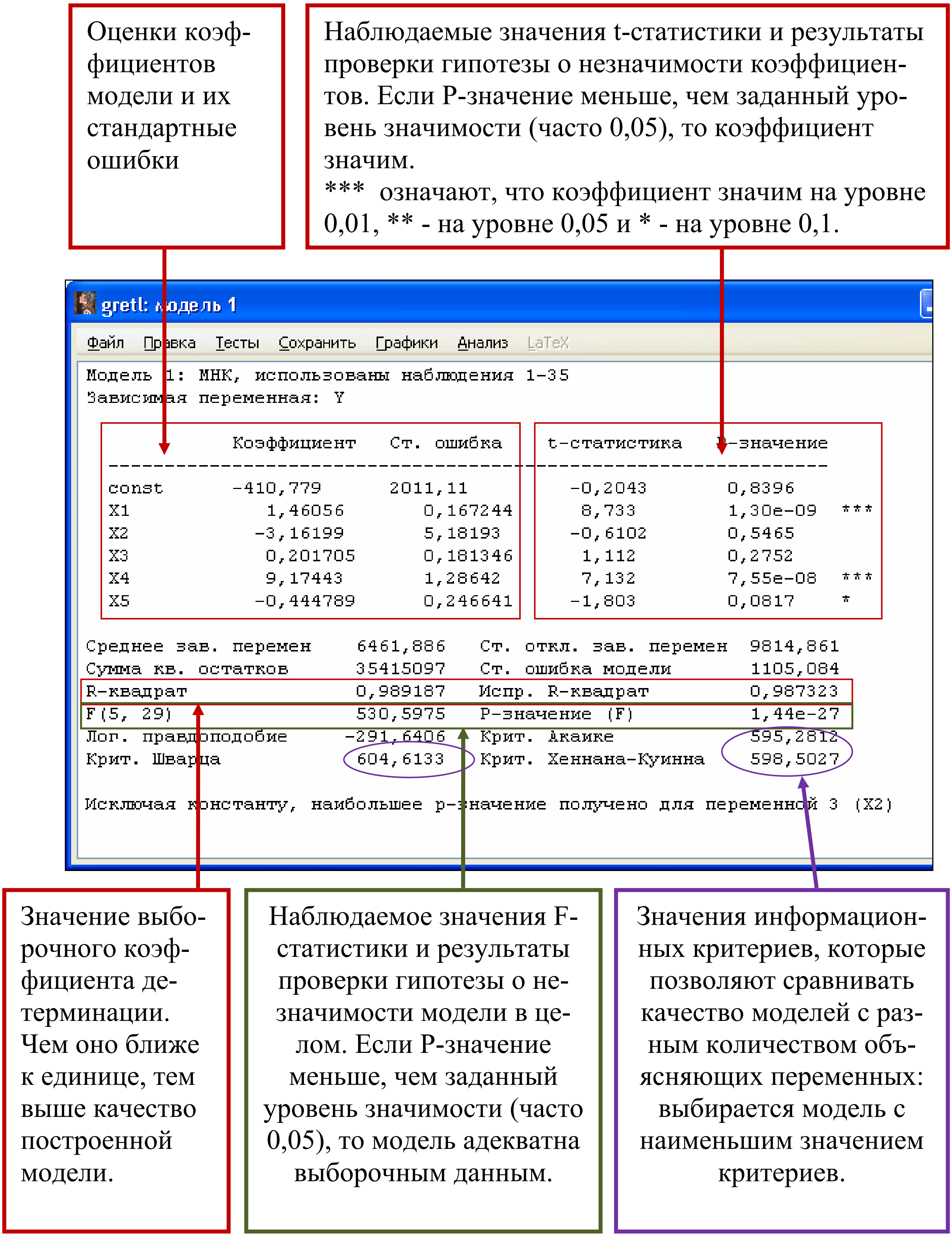


Рисунок 13 - Окно с результатами вычислений

Для копирования в отчет полученной таблицы результатов следует воспользоваться пунктом меню **Правка** формы **Модель 1** (рисунок 14).

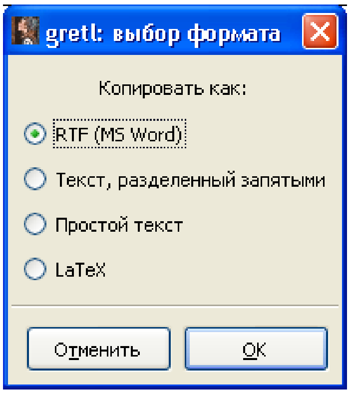


Рисунок 14 - Выбор формата копирования

Нажатие 

копирует таблицу в буфер обмена Windows, и после команды **Вставка**, например, в Word, результаты оценивания представляются в следующем виде:

Модель 1: МНК, использованы наблюдения 1-35
Зависимая переменная: Y

	Коэффициент	*Ст. ошибка*	*t-статистика*	*P-значение*	
const	-410,779	2011,11	-0,2043	0,83958	
X1	1,46056	0,167244	8,7331	<0,00001	***
X2	-3,16199	5,18193	-0,6102	0,54648	
X3	0,201705	0,181346	1,1123	0,27516	
X4	9,17443	1,28642	7,1317	<0,00001	***
X5	-0,444789	0,246641	-1,8034	0,08173	*

Среднее зав. перемен	6461,886	Ст. откл. зав. перемен	9814,861
Сумма кв. остатков	35415097	Ст. ошибка модели	1105,084
R-квадрат	0,989187	Испр. R-квадрат	0,987323
F(5, 29)	530,5975	Р-значение (F)	1,44e-27
Лог. правдоподобие	-291,6406	Крит. Акаике	595,2812
Крит. Шварца	604,6133	Крит. Хеннана-Куинна	598,5027

Оценка модели регрессии выглядит следующим образом:

$$\hat{y} = -410{,}779 + 1{,}46X_1 - 3{,}16X_2 + 0{,}20X_3 + 9{,}17X_4 - 0{,}44X_5$$

(2011,11) (0,17) (5,18) (0,18) (1,28) (0,25)

В круглых скобках записаны стандартные ошибки оценки коэффициентов $S_{b_j}, j = 0,1,..,.5$.

Проверка характера распределения регрессионных остатков

Для проверки значимости модели и значимости коэффициентов нужно убедиться, что остатки нормально распределены.

H_0: распределение регрессионных остатков не отличается от нормального.

H_1: распределение регрессионных остатков отличается от нормального.

Тестирование нормальности остатков модели, как и большинство тестов, которые нужно выполнить, работая с линейной моделью регрессии, доступны в меню **Тесты** формы **Модель 1** (рисунок 15).

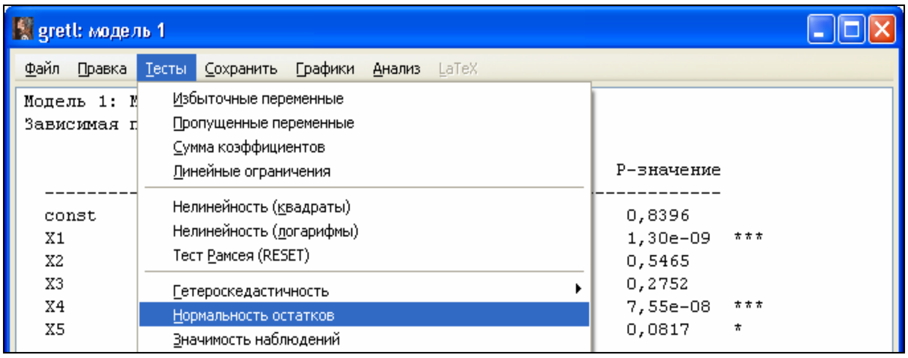


Рисунок 15 – Выбор пункта меню для проверки нормальности регрессионных остатков с помощью критерия хи-квадрат

Выбор этого пункта инициирует проверку нормальности распределения регрессионных остатков на основе критерия хи-квадрат, выводится гистограмма регрессионных остатков при автоматическом разбиении на интервалы (рисунок 16), а также распределение частот (рисунок 17).

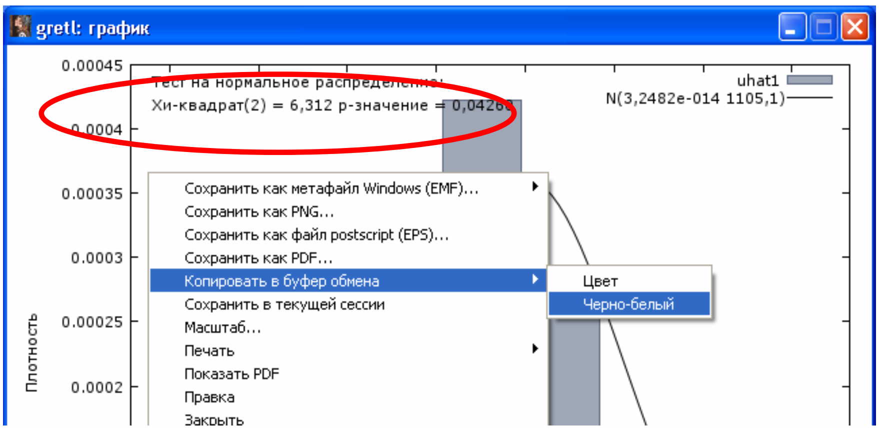
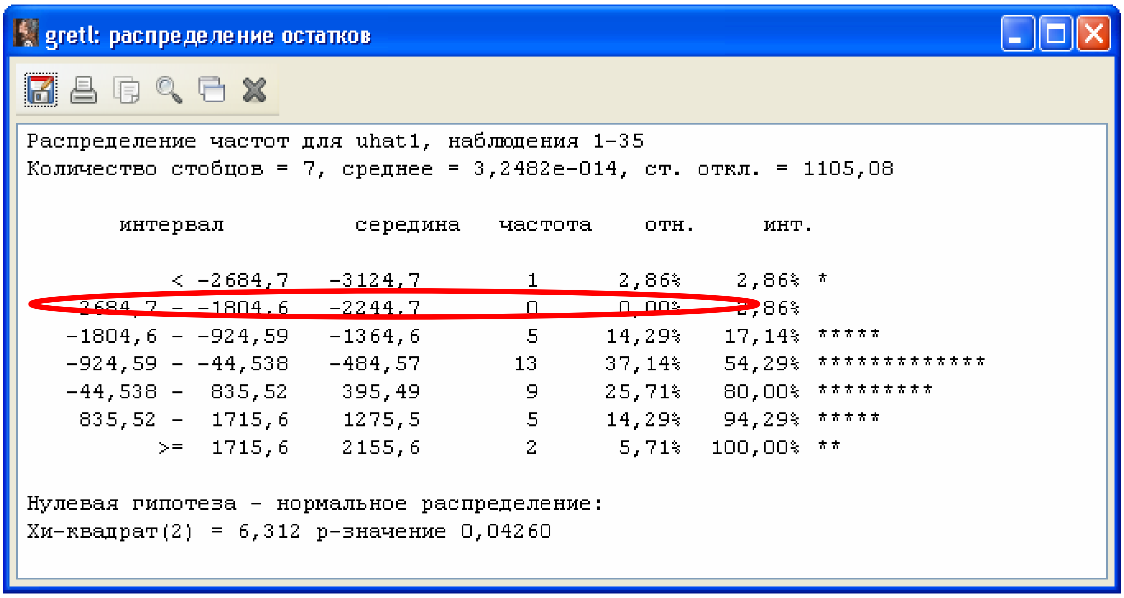


Рисунок 16 - Вид окна с гистограммой распределения регрессионных остатков и выбор пункта контекстного меню для копирования гистограммы в буфер обмена


gretl: распределение остатков

Распределение частот для uhat1, наблюдения 1-35
Количество стобцов = 7, среднее = 3,2482e-014, ст. откл. = 1105,08

интервал	середина	частота	отн.	инт.	
< -2684,7	-3124,7	1	2,86%	2,86%	*
-2684,7 - -1804,6	-2244,7	0	0,00%	2,86%	
-1804,6 - -924,59	-1364,6	5	14,29%	17,14%	*****
-924,59 - -44,538	-484,57	13	37,14%	54,29%	*************
-44,538 - 835,52	395,49	9	25,71%	80,00%	*********
835,52 - 1715,6	1275,5	5	14,29%	94,29%	*****
>= 1715,6	2155,6	2	5,71%	100,00%	**

Нулевая гипотеза - нормальное распределение:
Хи-квадрат(2) = 6,312 p-значение 0,04260

Рисунок 17 – Интервальный ряд частот и относительных частот

Для копирования гистограммы с результатами проверки гипотезы в буфер обмена нужно щелкнуть правой клавишей мыши на графике и выбрать в контекстном меню пункт **Копировать в буфер обмена – Черно-белый**. Вставка гистограммы из буфера осуществляется традиционно.

Таким образом, наблюдаемое значение статистики хи-квадрат составило 6,132 и вероятность того, что такое значение получилось случайно, если верна гипотеза H_0, составляет всего 0,04. Если принять уровень значимости $\alpha = 0{,}05$, то мы должны отвернуть нулевую гипотезу о нормальном распределении регрессионных остатков, так как p-значение 0,04<0,05. Между тем, в результате неудачной группировки интервал [-2684,7; -1804,6) содержит нулевую частоту. Кроме того, известно, что необходимыми условиями применимости критерия хи-квадрат является достаточно большой объем выборки, а в нашем случае выборка (N=35) невелика. Поэтому для проверки нормальности попробуем воспользоваться другими критериями.

В пакете GRETL реализованы такие критерии проверки согласия распределения с нормальным, как критерий Дурника-Хансена, Шапиро-Уилка, Лиллифорса и Жака-Бера. Выполним проверку нормальности распределения регрессионных остатков на их основе. Для этого сначала сохраним в новую переменную оценки регрессионных остатков $\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$ нашей модели, выбрав пункт меню **Сохранить – Остатки** в окне с результатами оценки модели **Модель 1**.

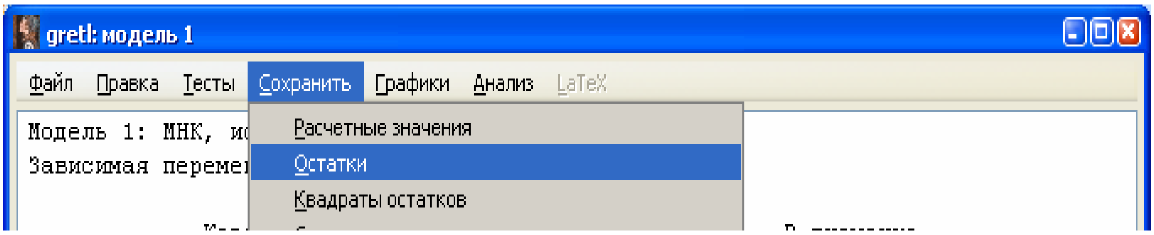


Рисунок 18 – Сохранение оценок регрессионных остатков в файле с основными данными

На экране появится окно, где можно задать имя переменной для сохранения оценок регрессионных остатков модели, и ее краткое описание (по умолчанию предлагается имя uhat[номер модели]) (рисунок 19).

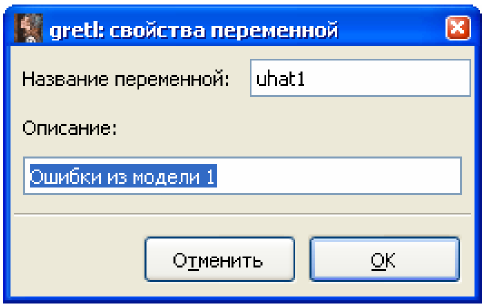


Рисунок 19 – Вид окна задания имени и описания переменной для сохранения значений оценок регрессионных остатков

В основном окне GRETL в списке переменных появится переменная с указанным именем. Выделим ее и выберем пункт основного окна GRETL **Переменная – Тест на нормальное распределение** (рисунок 20). Результаты проверки представлены на рисунке 21.

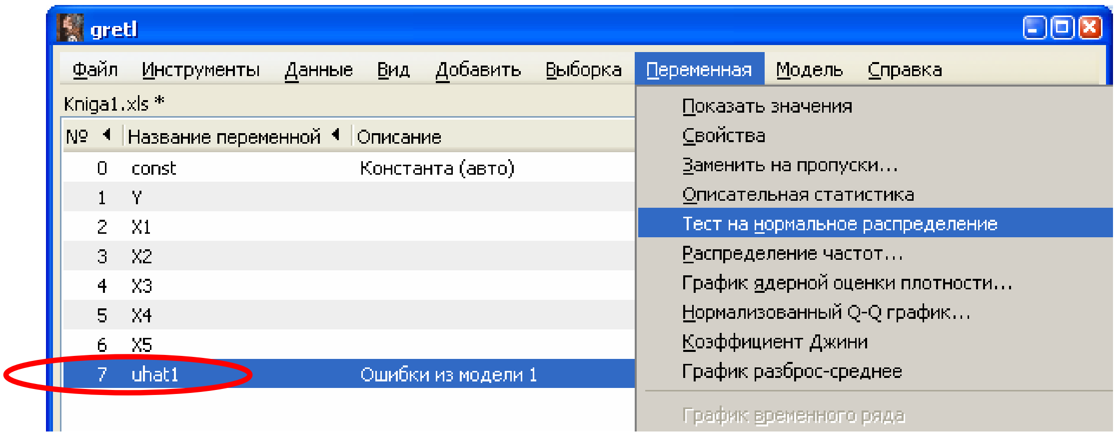


Рисунок 20 – Вид окна GRETL для выбора проверки нормальности

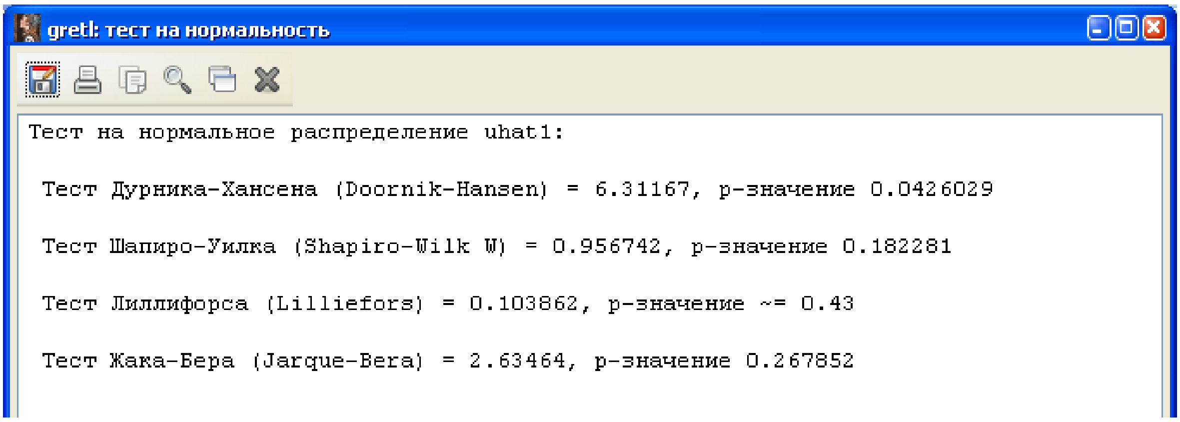


Рисунок 21 – Результаты проверки нормальности регрессионных остатков

По критериям Шапиро-Уилка, Лиллифорса и Жака-Бера на уровне значимости $\alpha = 0{,}05$ нулевая гипотеза о нормальности распределения регрессионных остатков не отвергается (достигаемые уровни значимости равны 0,18, 0,43 и 0,27 соответственно).

Исследование построенной регрессионной модели

Так как можно считать, что регрессионные остатки имеют нормальное распределение, то есть смысл проводить дальнейший анализ построенного уравнения множественной регрессии.

Проверка адекватности линейной модели множественной регрессии (ЛММР) выборочным данным

Общая вариация результативного признака складывается из вариации функции «регрессии», обусловленной варьированием значений объясняющих переменных $x_1,..,x_k$ (факторной дисперсии), и из вариации случайной величины ε относительно функции «регрессии» (остаточной дисперсии), то есть:

$$Q_{\text{общ}} = Q_{\text{факт}} + Q_{\text{ост}}$$

где $Q_{Общ} = \sum\limits_{i=1}^{n}(y_i - \bar{y})^2$ - общая сумма квадратов;

$Q_{факт} = \sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ - факторная сумма квадратов;

$Q_{ост} = \sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2 \equiv \sum\limits_{i=1}^{n}e_i^2$ - остаточная сумма квадратов.

Чем лучше построенное уравнение регрессии описывает исходные данные, тем больше будет факторная дисперсия $Q_{факт}$ и тем меньше будет остаточная дисперсия $Q_{ост}$. Этот очевидный факт положен в основу критерия проверки адекватности (значимости) построенного уравнения регрессии.

Выдвигается нулевая гипотеза о том, что ни один из признаков $x_1,..,x_k$ не оказывает значимого влияния на y:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (ЛММР неадекватна выборочным данным)

$H_1 : \exists j \in \{1,2,3\}: \quad \beta_j \neq 0$. (ЛММР адекватна выборочным данным)

Для проверки гипотезы H_0 используется статистика

$$F = \frac{Q_{факт}/k}{Q_{ост}/(n-k-1)} \quad \text{или} \quad F = \frac{\hat{R}^2/k}{\left(1-\hat{R}^2\right)/(n-k-1)}$$

которая при справедливости H_0 имеет распределение Фишера – Снедекора с числом степеней свободы $\nu_1 = k$ и $\nu_2 = n-k-1$.

Вернемся к рисунку 13:

R-квадрат	0,989187	Испр. R-квадрат	0,987323
F(5, 29)	530,5975	P-значение (F)	1,44e-27

Наблюдаемое значение статистики F составило $F_{набл} = 530{,}6$. Найдем критическое значение, выбрав пункт основного меню GRETL **Инструменты – Критические значения**, а в появившемся окне – вкладку *Фишера*. Зададим:

1) *Степени свободы 1* $\nu_1 = k$, то есть количество объясняющих переменных, в нашем случае $\nu_1 = k = 5$;

2) *Степени свободы 2* $\nu_2 = n-k-1$, то есть объем выборки минус количество объясняющих переменных, уменьшенное на единицу, в нашем случае $\nu_2 = n-k-1 = 35-5-1 = 29$;

3) *Правосторонняя вероятность,* или уровень значимости $\alpha = 0{,}05$ (рисунок 22).

После нажатия 

в новом окне будет выведена соответствующая критическая точка (рисунок 23). Таким образом, $F_{крит} = 2{,}55$ и $F_{крит} = 2{,}55 < F_{набл} = 530{,}6$, с вероятностью ошибиться 0,05 нулевая гипотеза отвергается, модель признается адекватной выборочным данным.

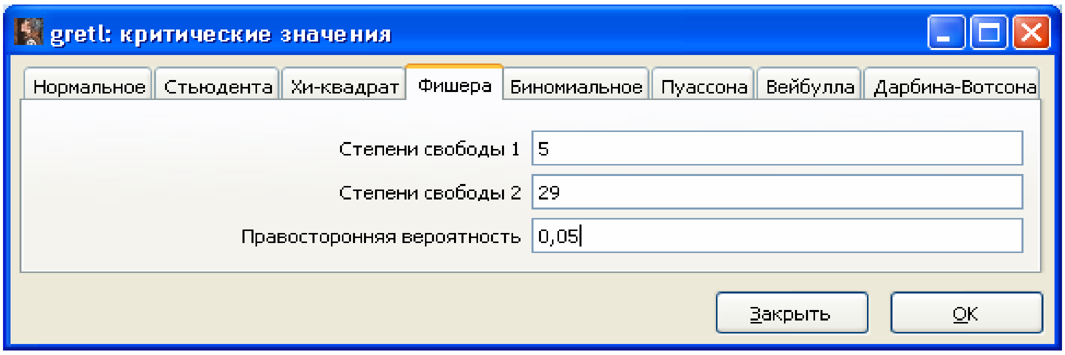


Рисунок 22 – Задание параметров для расчета критического значения распределения Фишера-Снедекора

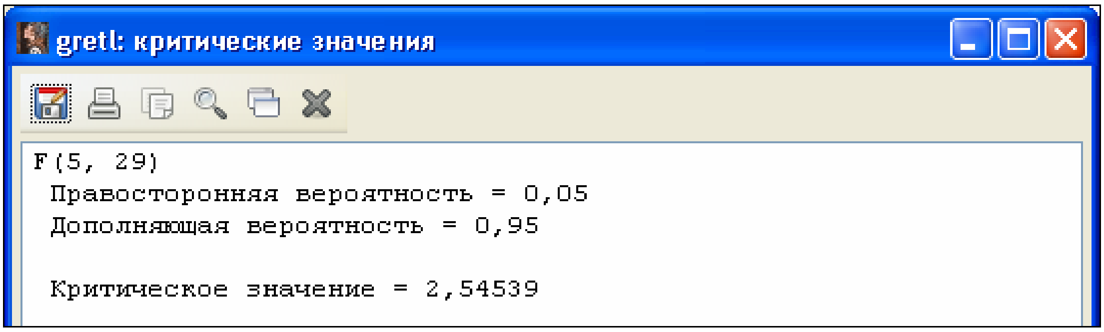


Рисунок 23 – Результат расчета критического значения распределения Фишера-Снедекора

Существует еще один вариант процедуры проверки статистической гипотезы, реализованной в большинстве статистических пакетов. Для наблюдаемого значения $F_{набл}$ рассчитывается вероятность того, что статистика примет значение больше него (так называемый «достигаемый уровень значимости»), которая сравнивается с заданным уровнем значимости. Если рассчитанная вероятность окажется меньше, что нулевая гипотеза отвергается.

Вернемся к рисунку 13:

R-квадрат	0,989187	Испр. R-квадрат	0,987323
F(5, 29)	530,5975	Р-значение (F)	1,44e-27

Достигаемый уровень значимости (p-значение) составил $1{,}44 \cdot 10^{-27}$, что намного меньше $\alpha = 0{,}05$, следовательно, H_0 отвергается, модель значима.

Поскольку нулевая гипотеза о незначимости уравнения регрессии была отвергнута, нужно проверить гипотезы о значимости каждого из коэффициентов уравнения регрессии.

Проверка значимости коэффициентов ЛММР

Для каждого коэффициента регрессии выдвигается гипотеза:

H_0: «коэффициент β_j незначимо отличен от нуля, признак x_j не оказывает влияния на y» (или формально $\beta_j=0$);

H_1: «коэффициент β_j значимо отличен от нуля, признак x_j оказывает влияния на y» (формально $\beta_j \neq 0$).

Для проверки H_0 используется статистика

$$t = \frac{b_j}{S_{b_j}}, \quad j = 1,2,...,k, \quad S_{b_j} = \sqrt{S_{ост} \cdot [(X^T X)^{-1}]_{jj}}, \; S_{ост} = \frac{1}{n-k-1} Q_{ост}$$

которая при справедливости H_0, имеет распределение Стьюдента с $\nu = n - k - 1$ степенями свободы. Далее сравниваем $|t_{набл}|$ с $t_{кр}(\alpha)$*- двухсторонним.*

Продемонстрируем проверку гипотезы для коэффициента β_1

H_0: $\beta_1=0$

H_1: $\beta_1 \neq 0$

Найдем наблюдаемое значение статистики, вернувшись к рисунку 13:

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-410,779	2011,11	-0,2043	0,8396	
X1	1,46056	0,167244	8,733	1,30e-09	***
X2	-3,16199	5,18193	-0,6102	0,5465	
X3	0,201705	0,181346	1,112	0,2752	
X4	9,17443	1,28642	7,132	7,55e-08	***
X5	-0,444789	0,246641	-1,803	0,0817	*

$$t_{набл} = \frac{1{,}461}{0{,}167} = 8{,}733$$

Найдем критическое значение $t_{кр}(\alpha)$, выбрав пункт основного меню GRETL **Инструменты – Критические значения**, а в появившемся окне – вкладку *Стьюдента*. Зададим:

1) Степени свободы $\nu = n - k - 1$, то есть объем выборки минус количество объясняющих переменных, уменьшенное на единицу, в нашем случае $\nu = n - k - 1 = 35 - 5 - 1 = 29$;

2) Правосторонняя вероятность, или уровень значимости $\alpha = 0{,}025$ (поскольку используется двухсторонняя критическая область) (рисунок 24).

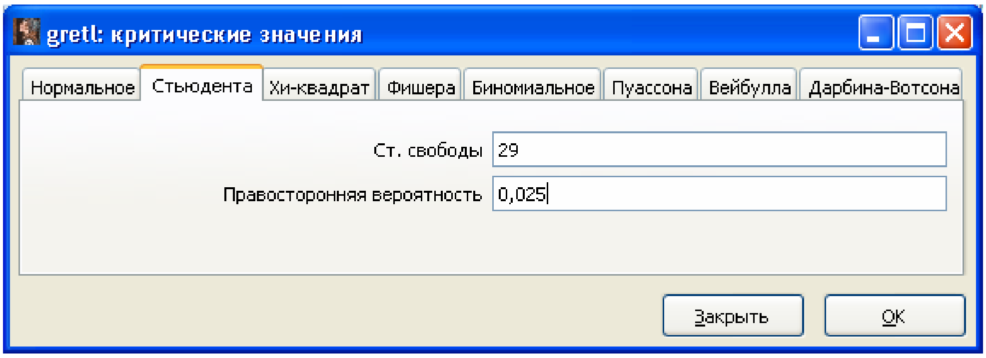


Рисунок 24 – Задание параметров для расчета критического значения распределения Стьюдента

После нажатия 

в новом окне будет выведена соответствующая критическая точка (рисунок 25).

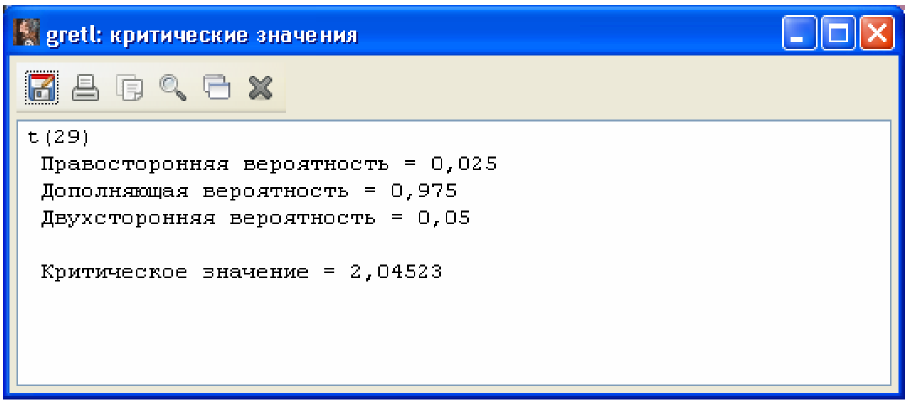


Рисунок 25 – Результат расчета критического значения распределения Фишера-Снедекора

Таким образом, $t_{крит} = 2{,}05$ и $t_{крит} = 2{,}05 > |t_{набл}| = 8{,}733$, нулевая гипотеза отвергается, коэффициент β_1 значим, признак x_1 оказывает влияние на ввод в действие жилых домов.

Аналогично проводятся расчеты для остальных коэффициентов.

Можно ориентироваться на другой подход, сравнивая достигаемый уровень значимости (столбец р-значение) с заданным, при этом для удобства значимые коэффициенты отмечены звездочками: * - значимые на уровне 0,1, ** - значимые на уровне 0,05 и *** - на уровне 0,01.

Вернемся к рисунку 13:

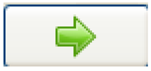
	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-410,779	2011,11	-0,2043	0,8396	
X1	1,46056	0,167244	8,733	1,30e-09	***
X2	-3,16199	5,18193	-0,6102	0,5465	
X3	0,201705	0,181346	1,112	0,2752	
X4	9,17443	1,28642	7,132	7,55e-08	***
X5	-0,444789	0,246641	-1,803	0,0817	*

В нашем случае на уровне значимости 0,05 значимыми являются только коэффициенты β_1 и β_4.

Исследование модели на мультиколлинеарность

Итак, в целом модель значима, но из пяти коэффициентов при объясняющих переменных значимы только два – при переменных X_1 и X_4. Стандартные ошибки остальных коэффициентов превышают или сравнимы по абсолютной величине с оценками коэффициентов, что свидетельствует о возможности включения точки 0 в соответствующие доверительные интервалы. Кроме того, вызывает сомнение отрицательный знак при переменной X_5: согласно модели, при увеличении среднемесячной начисленной заработной плате работников ввод в действие жилых домов, построенных населением за свой счет и с помощью кредитов, уменьшается. Одной из возможных причин перечисленных проблем может быть мультиколлинеарность – наличие тесных статистических связей между объясняющими переменными. Проверим это.

1. В первую очередь анализируют оценку матрицы парных коэффициентов корреляции между объясняющими переменными. Считается, что наличие значимых коэффициентов корреляции, по абсолютной величине превосходящих 07-0,8, свидетельствуют о присутствии мультиколлинеарности [1].

Для вычисления оценки матрицы парных коэффициентов корреляции выберем пункт главного меню GRETL **Вид – Корреляционная матрица**. В появившемся окне нужно выделить в списке *Доступные переменные* объясняющие признаки X1, X2, X3, X4, X5 и кнопкой с ЗЕЛЕНОЙ стрелкой  перенести их в список *Выбранные переменные* (рисунок 26).

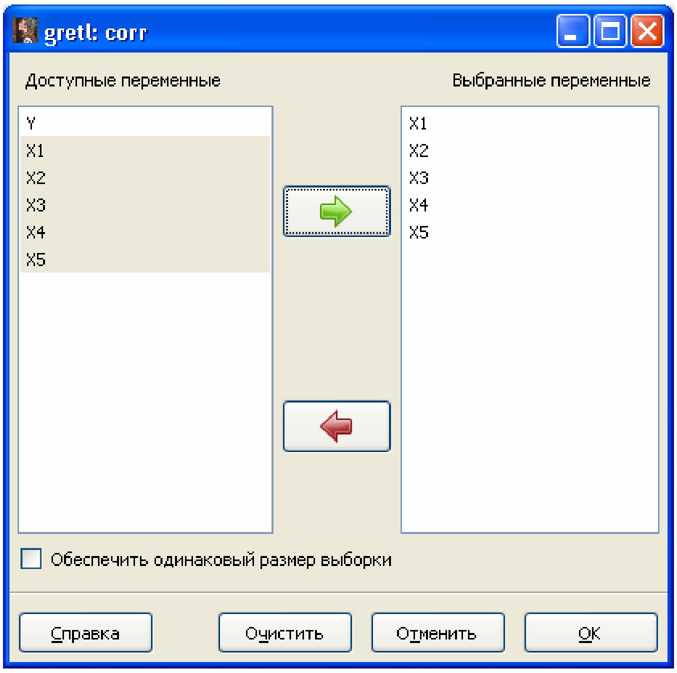


Рисунок 26 – Окно для вычисления оценки матрицы парных коэффициентов корреляции

После нажатия  появится окно с результатами (рисунок 27).

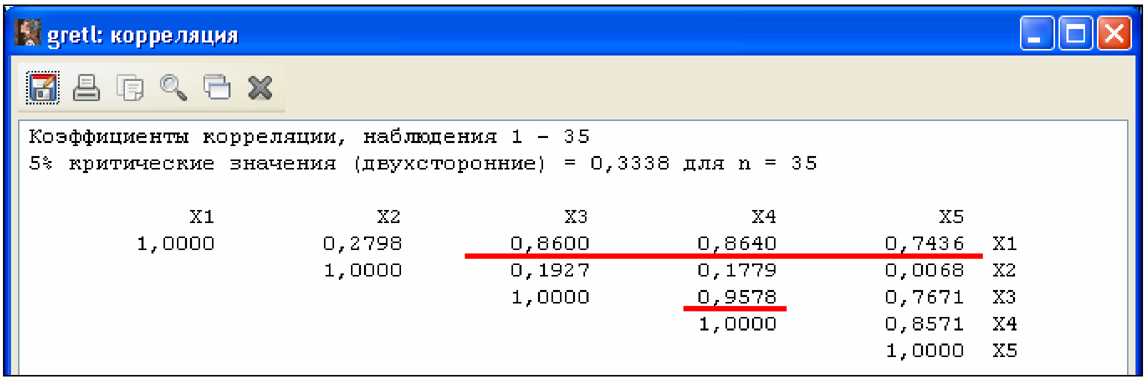


Рисунок 27 – Результаты оценки корреляционной матрицы между объясняющими переменными

Как видно из рисунка, между объясняющими переменными X1 и X3 ($r_{X1,X3} = 0{,}86$), X1 и X4 ($r_{X1,X4} = 0{,}86$), X1 и X5 ($r_{X1,X5} = 0{,}74$), X3 и X4 ($r_{X3,X4} = 0{,}96$) наблюдается тесная связь. Это один из признаков мультиколлинеарности.

Внимание!

Чтобы скопировать оценку корреляционной матрицы в отчет, вместо PrintScreen используйте пункт меню **Копировать.**

2. Более подробное изучение вопроса наличия взаимосвязи между объясняющими переменными достигается с помощью расчета значений коэффициентов детерминации $\hat{R}^2_{x^{(j)}.X(j)}$ каждой из объясняющих переменных $x^{(j)}$ по всем остальным переменным $X(j) = \left(x^{(1)},...,x^{(j-1)}, x^{(j+1)},...x^{(k)}\right)$.

Для этого нужно оценить модели регрессии, где в качестве зависимой переменной выбрать $x^{(j)}$, все остальные объясняющие переменные в качестве независимых. Например, для нахождения $\hat{R}^2_{x_1/x_2x_3x_4x_5}$ необходимо выбрать пункт главного меню **Модель – Метод наименьших квадратов** и в появившемся окне перенести переменную X1 в окно *Зависимая переменная*, а переменные X2, X3, X4, X5 - в список *Независимые переменные* (рисунок 28).

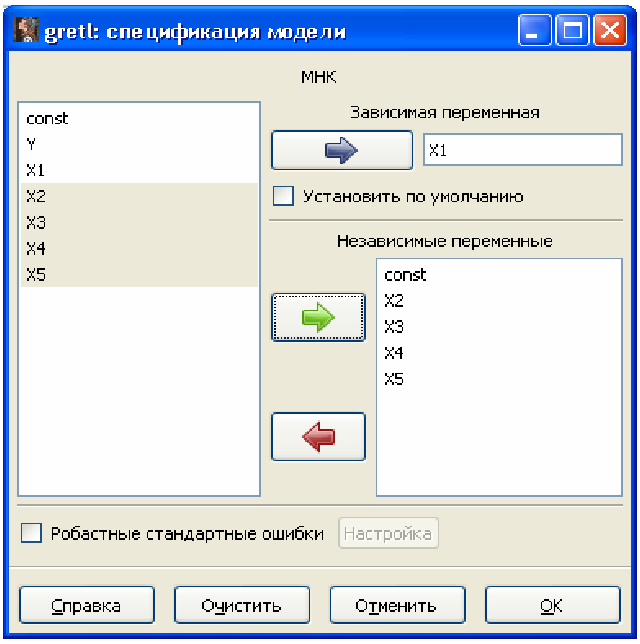
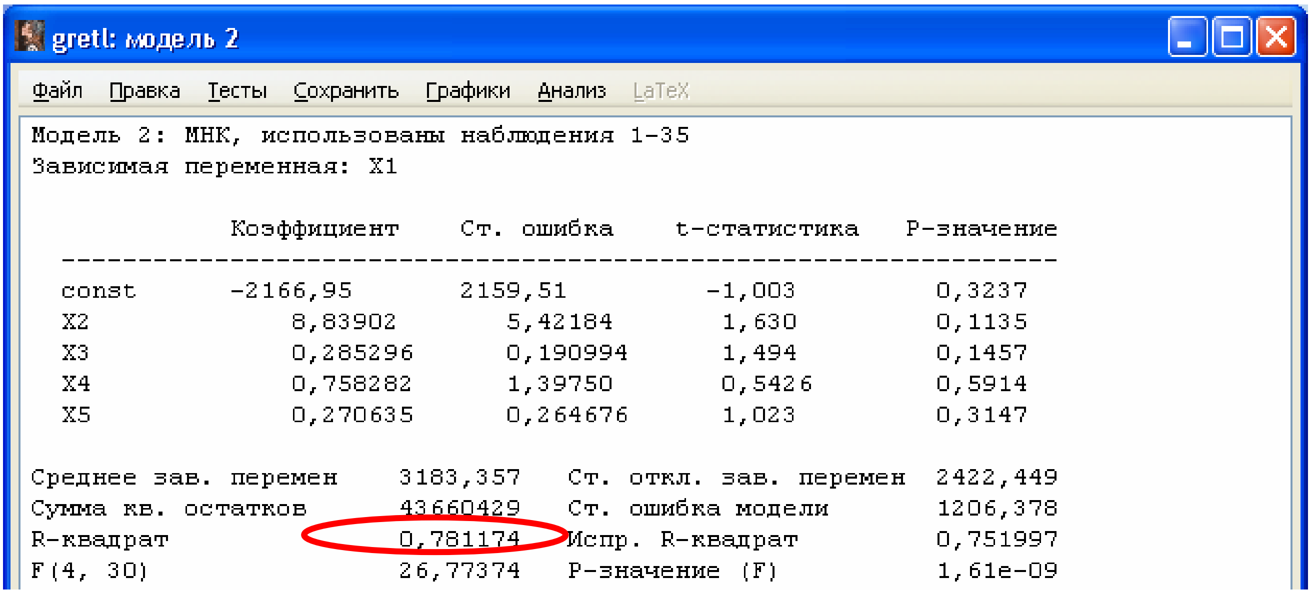


Рисунок 28 – Выбор переменных для нахождения оценки коэффициента детерминации $\hat{R}^2_{x_1/x_2x_3x_4x_5}$

В появившемся окне с результатами нас интересует R-квадрат (рисунок 29).


gretl: модель 2

Файл Правка Тесты Сохранить Графики Анализ LaTeX

```
Модель 2: МНК, использованы наблюдения 1-35
Зависимая переменная: X1

             Коэффициент    Ст. ошибка    t-статистика   P-значение
  ---------------------------------------------------------------
  const       -2166,95       2159,51         -1,003        0,3237
  X2              8,83902       5,42184        1,630        0,1135
  X3              0,285296      0,190994       1,494        0,1457
  X4              0,758282      1,39750        0,5426       0,5914
  X5              0,270635      0,264676       1,023        0,3147

Среднее зав. перемен     3183,357   Ст. откл. зав. перемен  2422,449
Сумма кв. остатков       43660429   Ст. ошибка модели       1206,378
R-квадрат                0,781174   Испр. R-квадрат         0,751997
F(4, 30)                 26,77374   Р-значение (F)          1,61e-09
```

Рисунок 29 – Оценка коэффициента детерминации переменной x_1

Таким образом, $\hat{R}^2_{x_1/x_2x_3x_4x_5} = 0{,}78$.

По той же схеме были найдены

$\hat{R}^2_{x_2/x_1x_3x_4x_5} = 0{,}18$

$\hat{R}^2_{x_3/x_1x_2x_4x_5} = 0{,}93$

$\hat{R}^2_{x_4/x_1x_2x_3x_5} = 0{,}96$

$\hat{R}^2_{x_5/x_1x_2x_3x_4} = 0{,}79$.

Для четырех из пяти переменных оценки коэффициентов детерминации высоки, превышают 0,7, что может говорить о наличии мультиколлинеарности.

Аналогом данного критерия является так называемый *метод инфляционных факторов*. Суть метода заключается в анализе величины

$$VIF_j = \frac{1}{1 - \hat{R}^2_{x^{(j)}.X(j)}}, \quad j = 1,\ldots,k.$$

Считается, что значения $VIF_j > 10$ могут свидетельствовать о наличии мультиколлинеарности.

Выберем пункт меню **Тесты – Мультиколлинеарность** окна *Модель 1*. Результаты представлены на рисунке 30.

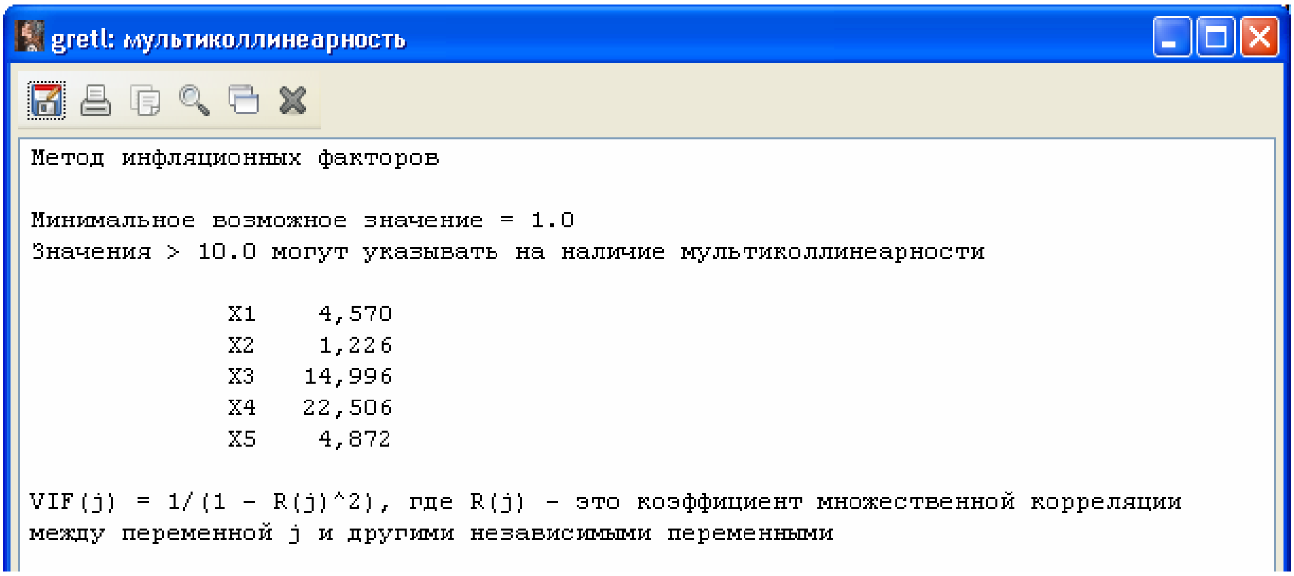


Рисунок 30 – Результаты проверки мультиколлинеарности по методу инфляционных факторов

Как видно из рисунка 30, инфляционные факторы переменных X1 и X5 превышают значение 4, а для переменных X3 и X4 – намного превышают 10 (равны 15 и 22,5 соответственно).

3. Достаточным условием плохой обусловленности матрицы $X^T X$ системы нормальных уравнений (наличия мультиколинеарности) является большое значение числа обусловленности:

$$M = \frac{\lambda_{max}}{\lambda_{min}},$$

где $\lambda_{max} = \max\limits_{1 \le i \le k+1} |\lambda_i|$,

$\lambda_{min} = \min\limits_{1 \le i \le k+1} |\lambda_i|$,

λ_i - собственные числа матрицы $X^T X$.

В пакете GRETL есть встроенные функции для работы с матрицами, в том числе для вычисления собственных чисел. Нам понадобятся следующие:

eigensym (A) – возвращает собственные числа симметричной матрицы A;

maxc(A) – возвращает строку, содержащую максимальные элементы столбцов матрицы А;

minc(A) – возвращает строку, содержащую минимальные элементы столбцов матрицы А;

A' – транспонирование матрицы А.

Для расчета числа обусловленности напишем небольшую последовательность команд - *скрипт*. Выберем пункт основного меню **Файл – Скрипты – Новый скрипт – Скрипт для gretl (*.inp)** (рисунок 31).

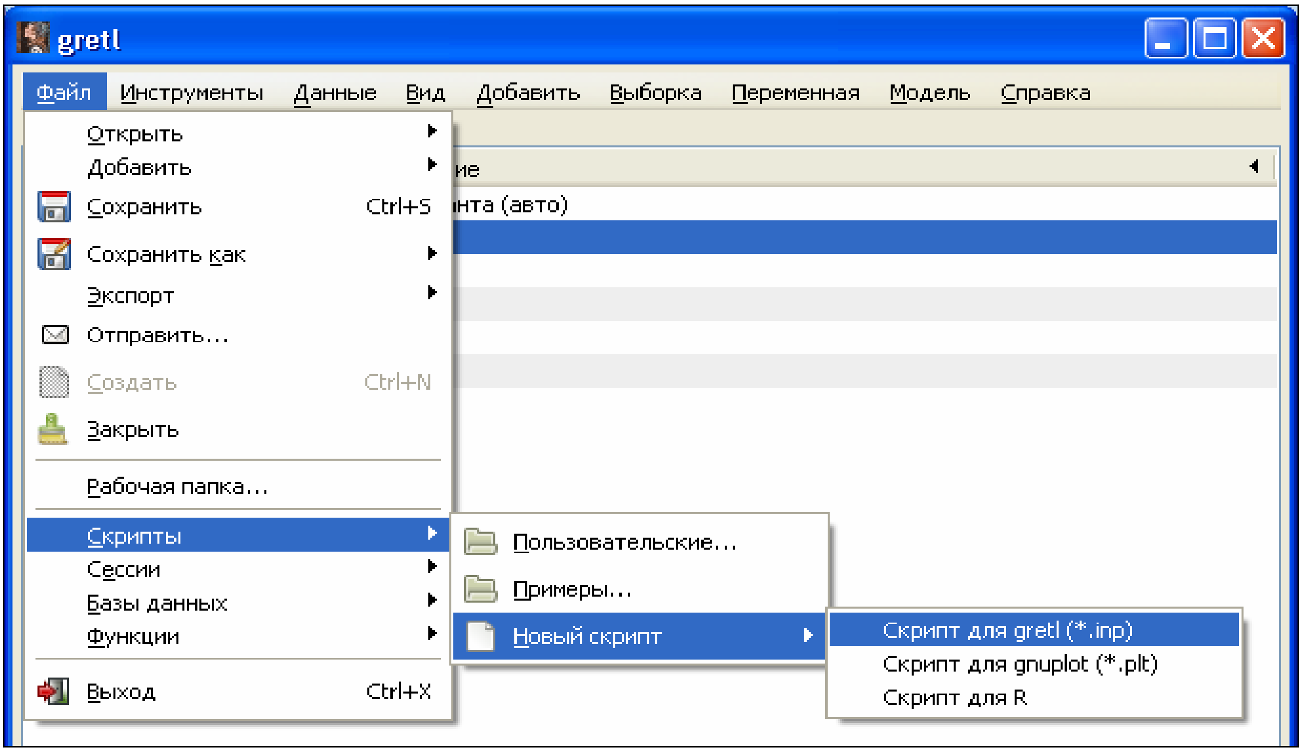
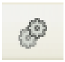


Рисунок 31 – Создание нового скрипта

В появившемся окне введем с клавиатуры следующие команды (рисунок 32):

matrix X={X1, X2, X3, X4, X5} – *создается матрица с именем X, столбцами которой будут значения переменных X1, X2, X3, X4, X5*.

matrix L=eigensym(X'X) – *создается матрица с именем L, содержащая собственные числа матрицы* X^TX

scalar M=maxc(L)/minc(L) – *создается скаляр с именем M, содержащий отношение максимального собственного числа матрицы* X^TX *к минимальному* Переход на новую строку осуществляется клавишей Enter.Запустим скрипт на выполнение кнопкой .

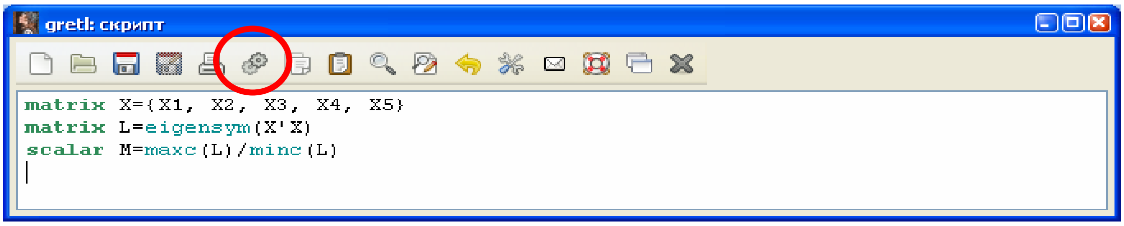


Рисунок 32 – Скрипт для вычисления числа обусловленности матрицы X^TX

Появится окно **Вывода скриптов** и окно **Значки** (рисунок 33).

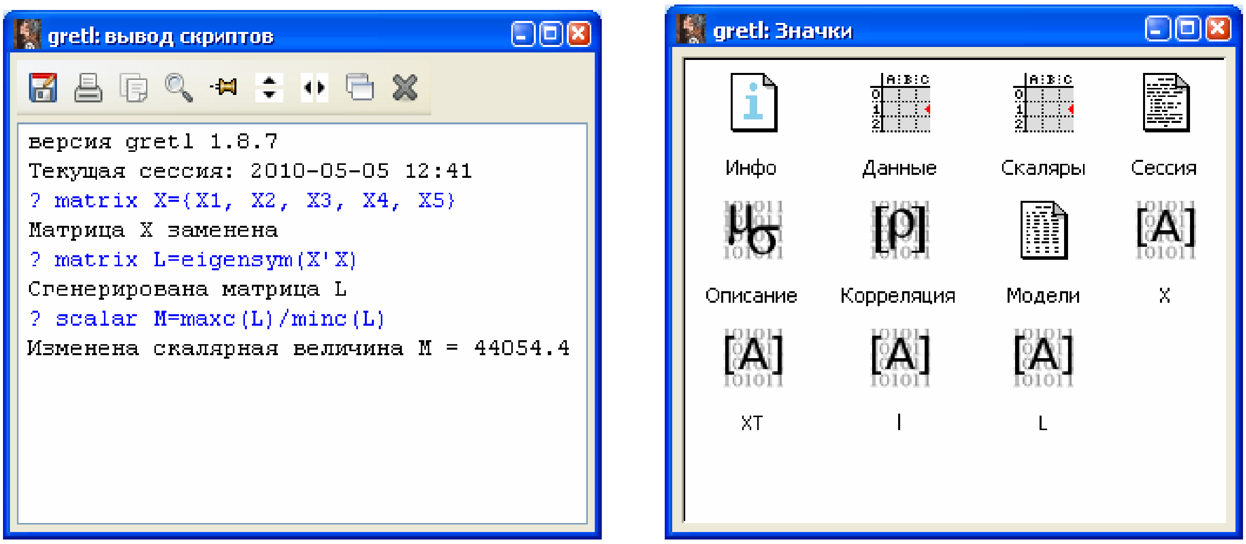
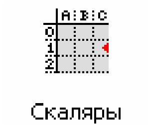


Рисунок 33 – Вид окон **Вывод скриптов** и **Значки** после выполнения скрипта для нахождения числа обусловленности матрицы

В окне вывода скриптов выведены результаты расчета – число обусловленности M=44054,4, что очень велико и говорит о наличии мультиколлинеарности. Число обусловленности сохранено в скалярах, его можно просмотреть в любой момент, кликнув на 

окна **Значки** (рисунок 34).

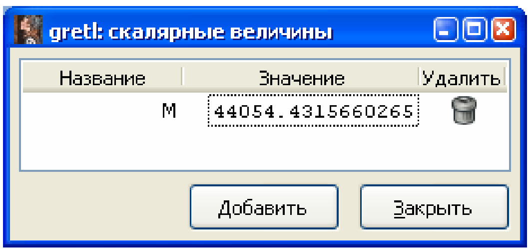


Рисунок 34 – Просмотр скаляров

Таким образом, можно сделать вывод, что перечисленные выше проблемы, связанные с незначимыми коэффициентами, неверными знаками коэффициентов являются следствием мультиколлинеарности. Плохая обусловленность матрицы системы привела к большим погрешностям в МНК-оценках коэффициентов и их стандартных ошибок. Эти оценки неустойчивы, незначительное изменение состава выборки или состава объясняющих переменных может вызвать кардинальное изменение модели, что делает модель непригодной для практических целей.

Для оценивания линейной модели множественной регрессии в условиях мультиколлинеарности используются методы пошаговой регрессии, использование гребневой регрессии (ридж-регрессии), переход от первоначальных переменных к их главным компонентам и др. [1,2].

Будем устранять мультиколлинеарность методом пошаговой регрессии с включением, суть которого заключается в переходе от исходного количества объясняющих переменных X1,…, Xk к меньшему числу X1,…,Xp, отобрав наиболее существенные с точки зрения их влияния на результативный признак.

На первом шаге $(l=1)$ определяется первая объясняющая переменная $x^{(i_1(1))}$, которую можно назвать наиболее информативной, при условии, что в регрессионную модель Y по X мы можем включить только одну из набора объясняющих переменных. Для этого нужно оценить k моделей регрессии: Y на X1, …, Y на Xk.

Создадим новый скрипт **Файл – Скрипты – Новый скрипт – Скрипт для gretl (*.inp)** и запишем в него команды, как показано на рисунке 35.

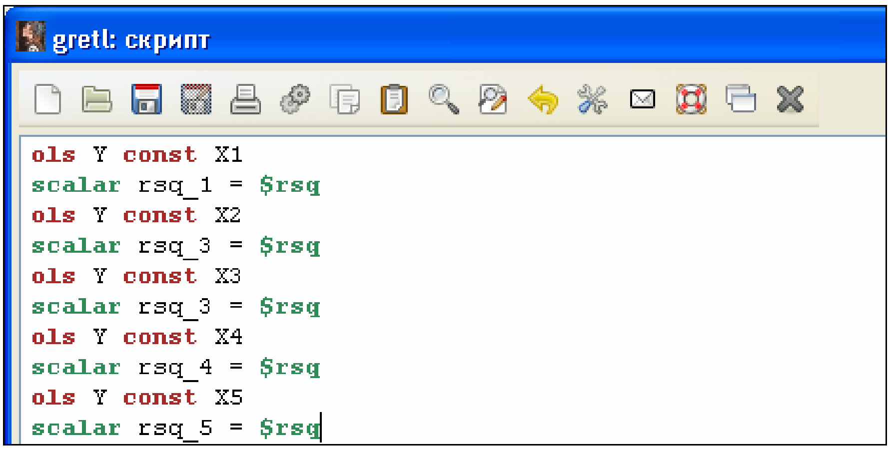
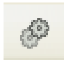


Рисунок 35 – Скрипт для оценивания регрессии Y на X1, …, Y на Xk.

Команда **ols Y const X1** – *оценить МНК линейную модель регрессии Y на X1 (включая const – свободный член).* Команда **scalar rsq_1 = $rsq** – *создать скаляр с именем rsq_1 и записать в него коэффициент детерминации оцененной выше модели*. После запуска скрипта на выполнение кнопкой  окно **Скаляры** примет вид, как на рисунке 36.

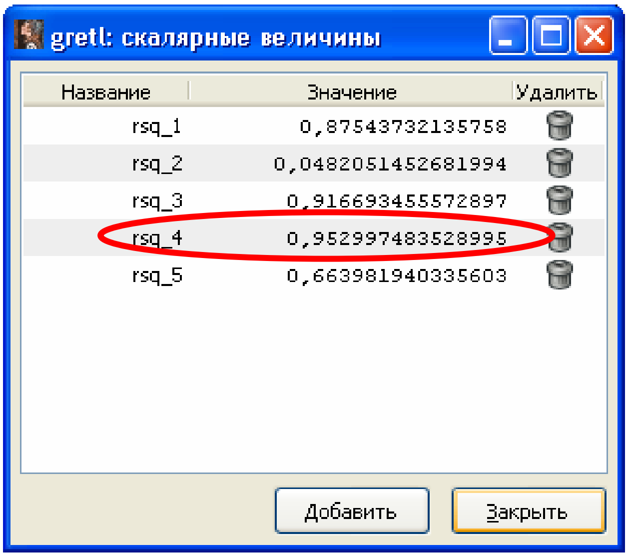


Рисунок 36 – Вид окна **Скалярные величины** после выполнения скрипта оценивания регрессий Y на X1, …, Y на Xk

Внимание!

Для вывода окна Значки можно воспользоваться пунктом главного меню **Вид- Сессия**

Таким образом,

$\hat{R}^2_{Y/X1} = 0{,}875 \qquad \hat{R}^2_{Y/X2} = 0{,}048 \qquad \hat{R}^2_{Y/X3} = 0{,}917$

$\hat{R}^2_{Y/X4} = 0{,}953 \qquad \hat{R}^2_{Y/X1} = 0{,}664$

Следовательно, на первом шаге в модель включаем переменную X4, так как

$\hat{R}^2_{Y/X4} = \max\limits_{1 \le j \le 5} \hat{R}^2_{Y/X_j} = 0{,}95$

Рассчитаем также несмещенную оценку коэффициента детерминации:

$$\hat{R}^{*2}(1) \cong 1 - \left(1 - \hat{R}^2(1)\right)\frac{N-1}{N-1-1} = 1 - (1 - 0{,}95)\frac{35-1}{35-1-1} = 0{,}948$$

и величину нижней доверительной границы $\hat{R}^2_{\min}(l)$:

$$\hat{R}^2_{\min}(l) = \hat{R}^{*2}(l) - 2\sqrt{\frac{2 \cdot l \cdot (N - l - 1)}{(N-1)(N^2-1)}}\left(1 - \hat{R}^2(l)\right) = 0{,}948 - 2\sqrt{\frac{2 \cdot 1 \cdot (35-1-1)}{(35-1)(35^2-1)}}(1 - 0{,}95) = 0{,}946$$

На втором шаге $(l = 2)$ нужно найти уже наиболее информативную пару объясняющих переменных $x^{(i_1(1))}, x^{(i_2(2))}$, при чем одна их них та, которую отобрали на предыдущем шаге – X4. Для этого нужно оценить k-1=5-1=4 модели регрессии: Y на X4 и X1, Y на X4 и X2, Y на X4 и X3, Y на X4 и X5. Модифицируем скрипт, как показано на рисунке 37 и запустим его на выполнение. Результаты представлены на рисунке 38.

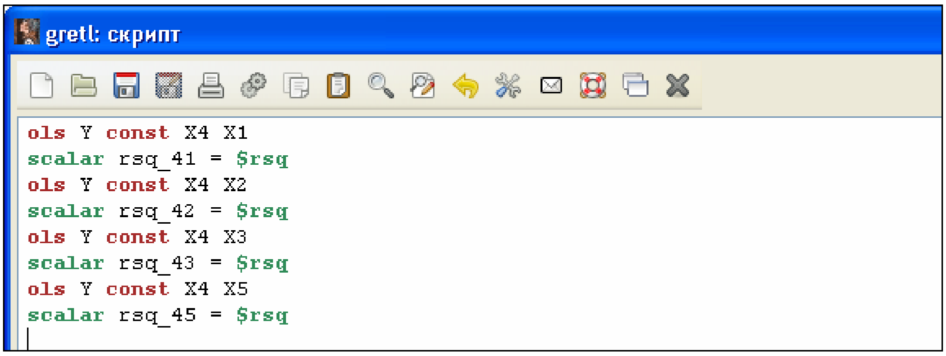


Рисунок 37 – Скрипт для оценивания регрессии Y на пару объясняющих переменных

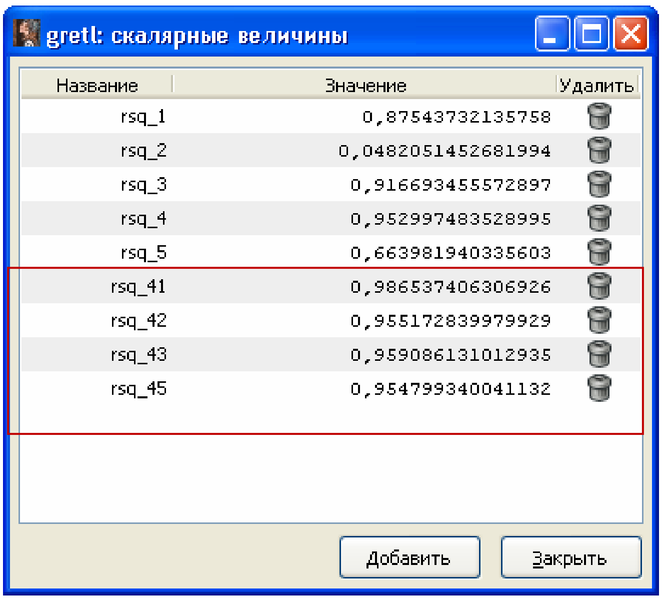


Рисунок 38 - Вид окна **Скалярные величины** после выполнения скрипта оценивания регрессий Y на пару объясняющих переменных

Как видно из рисунка 38,

$\hat{R}^2_{Y/X4,X1} = 0{,}986 \qquad \hat{R}^2_{Y/X4,X2} = 0{,}955$

$\hat{R}^2_{Y/X4,X31} = 0{,}959 \qquad \hat{R}^2_{Y/X4,X5} = 0{,}955$

То есть на втором шаге в модель добавляем переменную X1, так как

$$\hat{R}^2_{Y/X4,X1} = \max_{j \in \{1,2,3,5\}} \hat{R}_{Y/X4Xj}{}^2 = 0{,}99$$

$$\hat{R}^{*2}(2) \cong 1 - (1 - 0{,}99)\frac{35 - 1}{35 - 2 - 1} = 0{,}989$$

$$\hat{R}^2_{\min}(2) = 0{,}989 - 2\sqrt{\frac{2 \cdot 2 \cdot (35 - 2 - 1)}{(35 - 1)(35^2 - 1)}}(1 - 0{,}99) = 0{,}989.$$

$\hat{R}^2_{\min}(2)$ на шаге 2 больше, чем на шаге 1, поэтому продолжаем процедуру – будем строить регрессии на тройки переменных (рисунки 39 и 40).

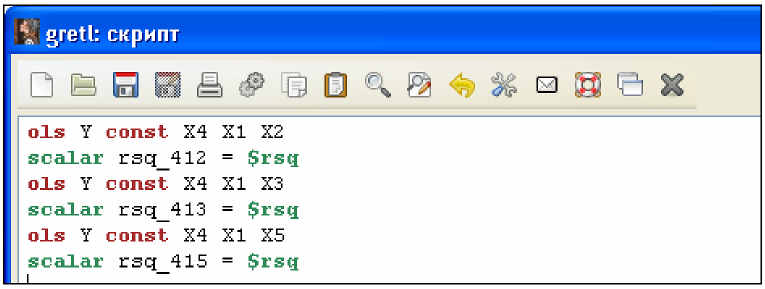


Рисунок 39 - Скрипт для оценивания регрессии Y на три объясняющие переменные

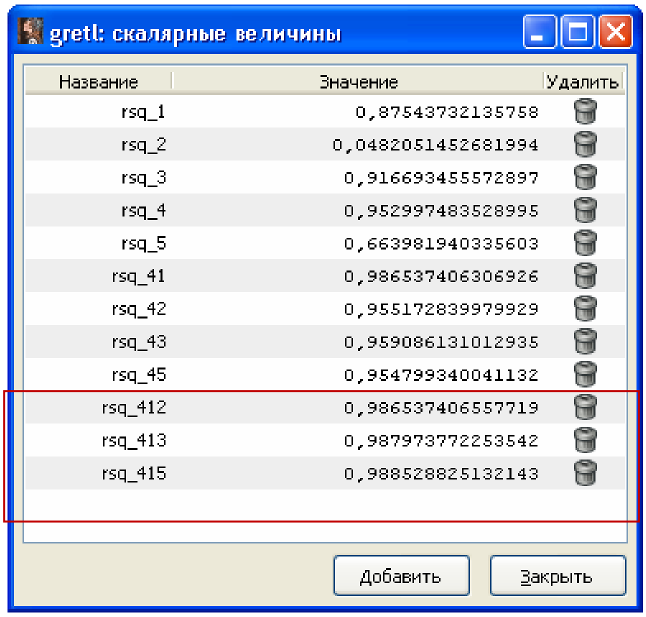


Рисунок 40 - Вид окна **Скалярные величины** после выполнения скрипта оценивания регрессий Y на три объясняющие переменные

Как видно из рисунка 40,

$$\hat{R}^2_{Y/X4,X1,X2} = 0{,}987 \qquad \hat{R}^2_{Y/X4,X1,X3} = 0{,}988 \qquad \hat{R}^2_{Y/X4,X1,X5} = 0{,}989$$

То есть на третьем шаге в модель добавляем переменную X5, так как $\hat{R}^2_{Y/X4,X1,X5} = \max\limits_{j\in\{2,3,5\}} \hat{R}_{Y/X4Xj}{}^2 = 0{,}99$

$$\hat{R}^{*2}(3) \cong 1-(1-0{,}99)\frac{35-1}{35-3-1} = 0{,}989$$

$$\hat{R}^2_{\min}(3) = 0{,}989 - 2\sqrt{\frac{2\cdot 3\cdot(35-3-1)}{(35-1)(35^2-1)}}(1-0{,}99) = 0{,}988 .$$

Построим в Excel график, на оси абсцисс которого будем откладывать номер шага (=количество переменных в модели), а по оси ординат – $\hat{R}^2(l)$ и $\hat{R}^2_{\min}(l)$ (рисунок 41).

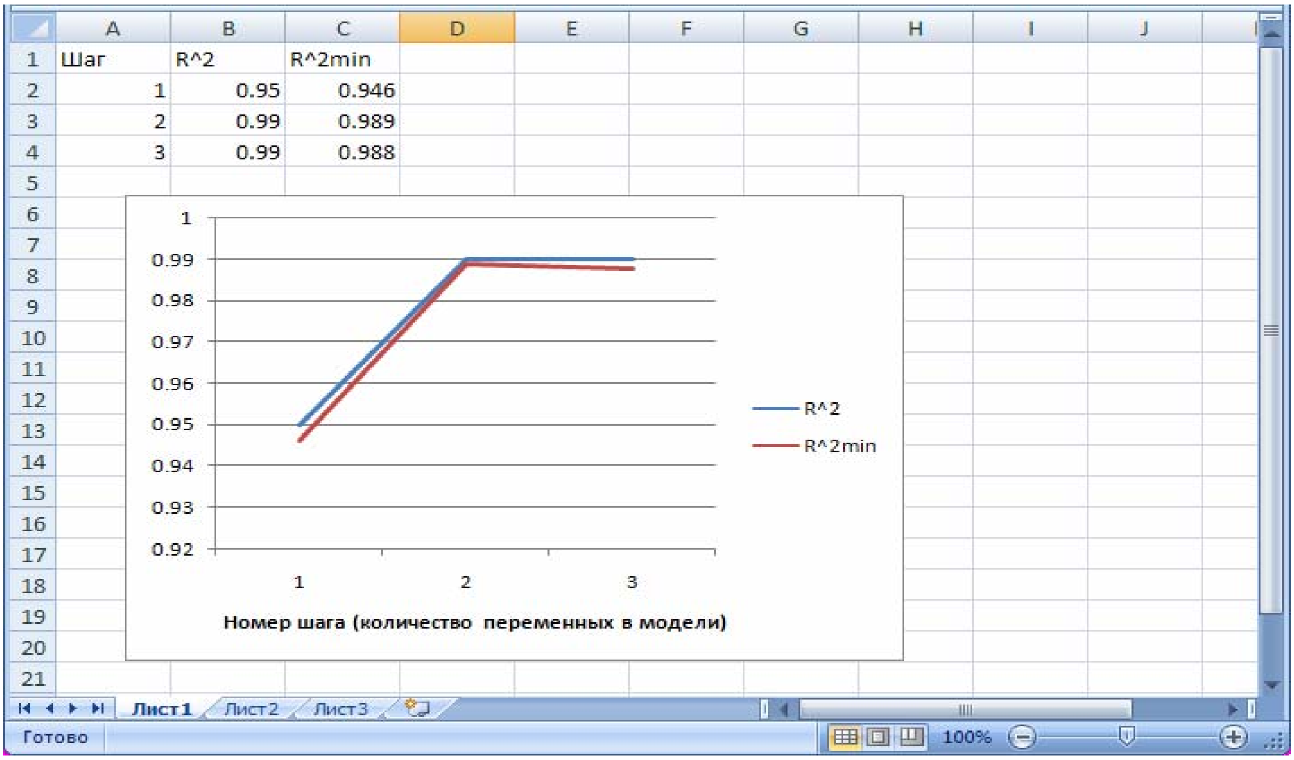


Рисунок 41- График изменения $\hat{R}^2(l)$ и $\hat{R}^2_{\min}(l)$ в зависимости от номера шага

Как видно из графика, нижняя граница доверительного интервала скорректированного коэффициента детерминации принимает максимальное значение на шаге 2, значит, процедура останавливается и окончательный состав переменных в модели: X1 и X4.

Внимание!

Если после включения в модель l-ой переменной (например, как в нашем примере, 3-ей переменной) $\hat{R}^{*2}(l)$ и $\hat{R}^2_{min}(l)$ возрастают, то процедура должна быть продолжена: оцениваются регрессии на четыре переменные, пять переменных и т.д. до тех пор, пока $\hat{R}^2_{min}(l)$ не начнет убывать.

Таким образом, в результате проведения пошаговой регрессии получили следующую оценку модели регрессии:

$$\hat{y} = \underset{(374{,}94)}{-2734{,}23} + \underset{(0{,}17)}{1{,}47}x_1 + \underset{(0{,}57)}{9{,}30}x_4, \hat{R}^2 = 0{,}99 \; ,$$

Проверка подтвердила нормальный характер распределения регрессионных остатков модели (таблица 1).

Таблица 1 – Результаты проверки гипотезы о нормальности регрессионных остатков модели, полученной методом пошаговой регрессии

№	Критерий	p-значение
1	Хи-квадрат	0,15
2	Дурника-Хансена	0,15
3	Шапиро-Уилка	0,50
4	Лиллифорса	0,14
5	Жака-Бера	0,57

Таблица 2 – Результаты оценки модели и проверки значимости коэффициентов

	Коэффициент	*Ст. ошибка*	*t-статистика*	*P-значение*	
const	-2734.23	374.943	-7.2924	<0.00001	***
X1	1.47364	0.165044	8.9288	<0.00001	***
X4	9.29612	0.57205	16.2505	<0.00001	***

Для значимых коэффициентов модели можно построить доверительные интервалы, знание которых позволит получить больше информации о диапазоне влияния исследуемых факторов на результативный показатель – ввод в действие жилых домов. Доверительный интервал надежности γ для коэффициента β_j имеет вид:

$$b_j - t(\alpha, n-k-1)\cdot S_{b_j} \le \beta_j \le b_j + t(\alpha, n-k-1)\cdot S_{b_j},$$

где b_j - оценка коэффициента β_j;

S_{b_j} - стандартная ошибка оценки коэффициента;

$t(\alpha, n-k-1)$ - $100\,\alpha\,\%$ -ная точка или квантиль уровня γ распределения Стьюдента.

Для построения 95%-ного доверительных интервалов ($\gamma = 0{,}95 \Rightarrow \alpha = 0{,}05$) нам понадобится $t_{крит}$, найденное для числа степеней свободы $\nu = n - k - 1 = 35 - 2 - 1 = 32$: $t_{крит} = 2{,}037$ (найдено с использованием **Инструменты – Критические значения**, а в появившемся окне – вкладку *Стьюдента*, правосторонняя доверительная вероятность задается равной $0{,}5\alpha = 0{,}025$).

Получаем доверительный интервал для коэффициента при x1:

$$1{,}473 - 2{,}037\cdot 0{,}165 \le \beta_1 \le 1{,}473 + 2{,}037\cdot 0{,}165$$

$$1{,}137 \le \beta_1 \le 1{,}809$$

Получаем доверительный интервал для коэффициента при x4:

$$9{,}296 - 2{,}037\cdot 0{,}872 \le \beta_1 \le 9{,}296 + 2{,}037\cdot 0{,}872$$

$$8{,}131 \le \beta_1 \le 10{,}461$$

Интерпретация результатов

Модель регрессии значима ($F_{набл} = 1172{,}48, p - значение = 1\cdot 10^{-30} < 0{,}05$); коэффициенты при всех переменных также значимы. Коэффициент детерминации составил 0,99, т.е. 99% вариации ввода в действие жилых домов можно объяснить вариацией инвестиций, направленные в жилищное хозяйство (на душу населения) и вариацией фонда оплаты труда работников, а 1% вариации, вероятно, объясняется неучтенными в модели факторами.

Согласно полученной модели, увеличение инвестиций, направленных в жилищное хозяйство, на 1 рубль на душу населения приводит к увеличению ввода в действие жилых домов, построенных населением за свой счет и с помощью кредитов в среднем на 1,5 кв. м (а с вероятностью 0,95 не меньше, чем на 1,137 кв.м, но не больше, чем 1,809 кв.м), а рост фонда оплаты труда работников на 1 млн. руб. – к увеличению ввода в действие жилых домов в среднем на 9,3 кв.м. (а с вероятностью 0,95 не меньше, чем на 8,131 кв.м, но не больше, чем 10,461 кв.м).

4 Содержание письменного отчета

Отчет должен быть выполнен на листах формата А4 с титульным листом, оформленным соответствующим образом и содержать следующее:

1) постановку задачи с вариантом выборок;
2) краткое изложение теории по исследованию зависимости между количественными переменными методом регрессионного анализа;
3) результаты компьютерной обработки данных;
4) анализ полученных результатов;
5) выводы по полученным результатам.

Отчет должен содержать описание и результаты основных этапов исследования (результаты оценивания линейной модели регрессии, проверки остатков на нормальность, исследование модели на мультиколлинеарность и т.д. с формулировкой всех необходимых гипотез), при этом обязательно четко обосновывать необходимость проведения каждого этапа. Технические аспекты и подробности реализации этапов в конкретном статистическом пакете должны быть опущены. Рекомендуемая схема описания каждого этапа: 1) постановка задачи этапа, 2) указание на используемый статистический метод, 3) полученные результаты решения задачи, 4) окончательные выводы по этапу.

5 Вопросы к защите

Группа А – базовые вопросы по лекционному материалу

1. Запишите линейную модель множественной регрессии и опишите исходные данные, которые необходимы для ее построения.
2. Что такое регрессионный остаток? Чем обусловлено его наличие в модели? Проиллюстрируйте графически.
3. Запишите условия Гаусса-Маркова.
4. Каким методом оцениваются коэффициенты модели регрессии? Запишите итоговую формулу.

5. Что характеризует общая дисперсия (определение и оценка)? Факторная дисперсия? Остаточная дисперсия?

6. Что характеризует коэффициент детерминации? В каких пределах он изменяется?

7. Проверка адекватности модели регрессии выборочным данным.

8. Проверка значимости коэффициентов модели регрессии.

9. Перечислите свойства МНК-оценок (при выполнении условий Гаусса-Маркова).

10. Как интерпретируются коэффициенты линейной модели регрессии?

11. Раскройте понятие полной и частичной мультиколлинеарности.

12. К чему приводит наличие мультиколлинеарности в модели?

13. Как выявить полную мультиколлинеарность? Частичную?

14. Опишите алгоритм устранения мультиколлинеарности методом пошаговой регрессии.

Группа В – вопросы, связанные с выводом формул, доказательством теорем и свойств

1. Выведите формулу для нахождения оценок линейной модели множественной регрессии методом наименьших квадратов.

2. Докажите свойство несмещенности МНК-оценки коэффициентов линейной модели множественной регрессии

3. Получите формулу для ковариационной матрицы вектора МНК-оценок линейной модели множественной регрессии.

Группа С – дополнительные вопросы

1. Опишите алгоритм устранения мультиколлинеарности методом ридж-регрессии.

2. Проиллюстрируйте графически потенциальное преимущество смещенных оценок, полученных по методу ридж-регрессии, перед несмещенными МНК-оценками в условиях мультиколлинеарности.

3. В чем суть метода главных компонент как средства устранения мультиколлинеарности?

4. При анализе линейной модели регрессии на мультиколлинеарность в матрице парных коэффициентов корреляции между объясняющими переменными не оказалось элементов, превышающих 0,7 по модулю. Можно ли в этом случае говорить об отсутствии мультиколлинеарности?

5. Зачем при проведении регрессионного анализа проверяется нормальный характер распределения остатков модели?

6. Почему при устранении мультиколлинеарности методом пошаговой регрессии критерием остановки выбрано достижение максимума нижней границы доверительного интервала для скорректированного коэффициента детерминации, а не самим коэффициентом детерминации?

Список использованных источников

1. Айвазян, С. А. Прикладная статистика и основы эконометрики: учебник для вузов / С. А. Айвазян, В. С. Мхитарян. – М.: ЮНИТИ, 1998. – 1022 с.

2. Тихомиров, Н.П. Эконометрика: учебник / Н. П. Тихомиров, Е. Ю. Дорохина. – М.: Издательство «Экзамен», 2003. – 512 с.

Приложение А
Исходные данные
(обязательное)

Таблица А.1 – Значения социально-экономических показателей, характеризующих города и районы Оренбургской области, за 2007 год

Наименование района/города	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
Абдулинский	109189	8877	38908	3163,3	1919	156,0	1919	24,0	235,6	2260	119,6	4408	18,2	3	1	5	8378
Адамовский	646948	21565	130112	4337,1	9848	328,3	9587	19,6	200,1	8695	638,5	6119	5,6	16	7	36	11051
Акбулакский	229355	7722	58886	1982,7	3543	119,3	3543	17,5	169,4	5525	334,6	5046	26,7	20	14	34	7194
Александровский	73185	3792	21560	1117,1	1611	83,5	1611	18,8	154,8	3947	264,1	5576	26,7	14	4	32	11919
Асекеевский	259212	11270	77916	3387,7	5822	253,1	5822	19,6	209,4	4956	301,1	5063	29,4	12	4	24	8043
Беляевский	129238	6662	49744	2564,1	3679	189,6	3679	19,2	180,5	4048	280,2	5768	37,5	9	5	23	8629
Бугурусланский	288477	12936	56792	2546,7	3647	163,5	3647	20,4	250,1	4366	271,5	5182	50,0	3	12	25	13728
Бузулукский	176376	5281	79807	2389,4	5666	169,6	5666	20,0	199,0	6969	545,5	6523	36,8	24	10	26	6121
Гайский	124545	11220	15191	1368,6	1135	102,3	1135	20,8	245,2	2147	128,4	4984	28,6	7	1	9	12275
Грачевский	178909	11927	44703	2980,2	3340	222,7	3340	22,7	214,3	3606	297,6	6877	18,2	10	3	29	9073
Домбаровский	113887	6090	40844	2184,2	2456	131,3	2249	20,4	115,2	3479	258,1	6182	37,5	14	10	31	8897
Илекский	201464	7069	79755	2798,4	7316	256,7	5800	18,8	167,6	5619	353,2	5238	25,3	6	7	20	8187
Кваркенский	164722	7661	39571	1840,5	2995	139,3	2767	19,4	184,4	5227	363,9	5802	42,1	3	3	20	8495
Красногвардейский	205538	8821	70022	3005,2	5194	222,9	5194	21,0	231,3	4616	394,7	7125	16,7	15	8	22	9672
Кувандыкский	200475	8754	114663	5007,1	8090	353,3	8090	18,2	244,8	4327	227,48	4379	18,8	6	3	7	4494
Курманаевский	119348	5938	53821	2677,7	3641	181,1	2791	21,7	198,2	4251	327,4	7808	40,0	13	8	31	8203
Матвеевский	80125	5451	38967	2650,8	3034	206,4	2781	20,9	208,8	3179	209,8	5500	33,3	2	5	14	13456
Новоорский	1113783	34806	186969	5842,8	14196	443,6	13282	21,3	153,3	6820	749,0	9153	25,0	29	24	49	14059
Новосергиевский	451312	12231	172317	4669,8	12700	344,2	12700	21,3	232,5	9898	806,7	6792	27,8	22	12	57	22151
Октябрьский	357604	16108	106976	4818,7	9272	417,7	5296	21,9	166,8	5695	439,7	6434	17,6	20	9	23	12860
Оренбургский	4542065	62220	1056362	14470,7	68415	937,2	59234	20,5	234,1	26176	4389,9	13976	20,4	165	290	586	25230
Первомайский	251218	8753	105431	3673,6	6431	224,1	6431	18,2	192,9	6384	534,4	6976	18,8	20	34	31	11456
Переволоцкий	174074	5881	40391	1364,6	3122	105,5	2974	19,8	198,0	5826	419,0	5994	9,1	18	11	47	11327
Пономаревский	95263	5670	64800	3857,1	4842	288,2	4842	22,9	161,6	2827	223,3	6584	83,3	8	10	16	16004
Сакмарский	432616	14325	135436	4484,6	10120	335,1	10120	17,9	173,9	5684	463,0	6788	52,4	18	24	62	10789

Продолжение таблицы А.1

Наименование района/города	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
Саракташский	482734	11200	189860	4405,1	14219	329,9	12330	18,3	208,4	9824	654,5	5552	10,7	31	20	66	12486
Светлинский	91578	5324	15524	902,6	1598	92,9	1027	19,9	113,5	4764	425,8	7448	7,7	16	5	25	9309
Северный	148732	8548	47460	2727,6	3546	203,8	3546	21,2	162,7	3900	306,1	6540	36,4	24	5	20	16139
Соль-Илецкий	93681	3419	9351	341,3	1199	43,8	478	16,6	191,3	4973	274,7	4602	41,7	9	2	14	16105
Сорочинский	113378	7268	17266	1106,8	1409	90,3	1230	21,4	238,8	3804	243,1	5326	13,3	5	10	1	5818
Ташлинский	436789	16359	136321	5105,7	9162	343,1	9010	21,0	150,6	7831	439,0	4672	17,6	13	3	34	13005
Тоцкий	170168	4212	74639	1847,5	7188	177,9	5539	18,1	148,9	5243	398,3	6331	40,0	20	12	48	12484
Тюльганский	150351	6398	47043	2001,8	3510	149,4	3510	19,8	167,0	5051	360,8	5953	40,0	20	5	28	11844
Шарлыкский	238024	11499	76059	3674,3	4934	238,4	4934	21,0	258,0	4717	331,7	5860	9,1	8	11	21	16074
Ясненский	43901	6456	830	122,1	62	9,1	62	17,5	98,4	1912	115,8	5047	25,7	2	10	3	6326

X1 – инвестиции в основной капитал, тыс.руб.

X2 – инвестиции в основной капитал на душу населения, руб.

X3 – инвестиции, направленные в жилищное хозяйство, тыс. руб.

X4 – инвестиции, направленные в жилищное хозяйство, на душу населения, руб.

X5 – ввод в действие жилых домов, кв.м

X6 – ввод в действие жилых домов на 1000 человек населения, кв.м

X7 – ввод в действие жилых домов, построенных населением за свой счет и с помощью кредитов, кв.м

X8 – площадь жилищ, приходящаяся в среднем на одного жителя, кв.м

X9 – обеспеченность населения собственными легковыми автомобилями в расчете на 1000 населения, штук

X10 – среденесписочная численность работников, человек

X11 – фонд оплаты труда работников, млн. руб.

X12 – среднемесячная начисленная заработная плата работников, руб.

X13 – удельный вес убыточных организаций, в % от общего числа организаций

X14 – число предприятий и организаций обрабатывающих производств

X15 – число предприятий и организаций строительства

X16 – число предприятий и организаций торговли

X17 – оборот розничной торговли на душу населения, руб.

Таблица А.2 – Варианты заданий

№ варианта	Результативный показатель (обозначить Y)	Номера факторных признаков X
1	X5	1, 8, 12, 13, 14
2		1, 8, 10, 12, 13
3		1, 8, 11, 13, 14
4		1, 8, 10, 13, 14
5		1, 8, 10,11, 13
6		2, 8, 12, 13, 14
7		2, 9, 10, 12, 13
8		2, 9, 11, 13, 14
9		2, 9, 10, 13, 14
10		2, 8, 10, 11, 13
11	X6	3, 4, 8, 9, 15
12		3, 4, 13, 9, 15
13		2, 3, 4, 16, 17
14		1, 8, 10,12, 17
15		1, 4, 3, 14, 15
16		1, 4, 14, 8, 15
17		2, 3, 5, 8, 14
18		4, 10, 8, 15, 16
19		8, 12, 14, 16, 17
20		9, 12, 14, 16, 17
21	X7	1, 10, 14, 13, 15
22		2, 11, 14, 13, 15
23		3, 12, 14, 13, 15
24		8, 9, 10, 11, 17
25		1, 9, 10, 12, 17
26		1, 9, 10, 12, 16
27		4, 9, 10, 12, 17
28		3, 14, 15, 16 ,17
29		4, 14, 15, 16, 17
30		3, 13, 15,16, 17