Министерство образования Российской Федерации
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет математики и информатики

Эконометрика: Лабораторный практикум

для студентов очного отделения экономического факультета

Благовещенск
2004

Печатается по решению
редакционно-издательского совета
факультета математики и информатики
Амурского государственного
университета

Двоерядкина Н.Н., Киселева А.Н., Макарчук Т.А., (составители)

Эконометрика. Лабораторный практикум Благовещенск.: Амурский гос. ун – т, 2004.

Предназначено для студентов очного отделения экономического факультета, изучающих дисциплину «Эконометрика». Пособие составлено на основе учебных программ по эконометрике для специальностей: 0604, 0605, 0606. Приведены необходимые теоретические сведения, даны рекомендации по использованию компьютерного пакета и подобраны задания для самостоятельной работы студентов.

Рецензент: А.М. Емельянов, зав. кафедрой высшей математики ДальГАУ, д-р техн. наук, профессор

Введение.

В настоящее время широкое распространение получило использование моделирования и количественного анализа, на базе которых сформировалось одно из направлений экономических исследований – эконометрика.

Эконометрика – наука, исследующая количественные закономерности и взаимозависимости в экономике при помощи методов математической статистики. Основа этих методов - корреляционно – регрессионный анализ.

Эконометрика занимается построением, статистической оценкой и анализом экономических зависимостей и моделей на основе изучения эмпирических данных.

Но одним из важных направлений эконометрики является прогнозирование различных экономических показателей.

Предмет ее исследования – экономические явления, а изучение этих явлений в эконометрике осуществляется через эконометрические модели.

Эконометрические методы и модели широко применяются во всех производственных и коммерческих фирмах для принятия практических решений в прогнозировании, банковском деле, бизнесе, а так же для разработки вариантов перспективного развития предприятия.

Настоящий лабораторный практикум построен таким образом, чтобы конспективно изложить теоретические основы эконометрики и познакомить читателя с программой *Statistica*.

Практикум ориентирован на студентов экономических специальностей университетов. Также он будет полезен преподавателям и всем интересующимся статистическими методами анализа экономических процессов.

Предполагается, что студенты, изучающие эконометрику, уже прослушали базовые курсы по высшей математике, теории вероятностей и математической статистики, микро- и макро- экономике.

Лабораторная работа №1.

Основы программы *Statistica*.

Программа *Statistica* состоит из отдельных модулей, каждый из которых располагается в отдельном окне. Первым модулем является *Basic Statistics* – основные статистики.

Окно *Basic Statistics* – представлено в виде таблицы с данными и пунктов меню (рис.1).

Список пунктов меню можно закрыть или свернуть. При закрытии списка, пункты меню будут размещены в горизонтальном меню команды *Analysis.*

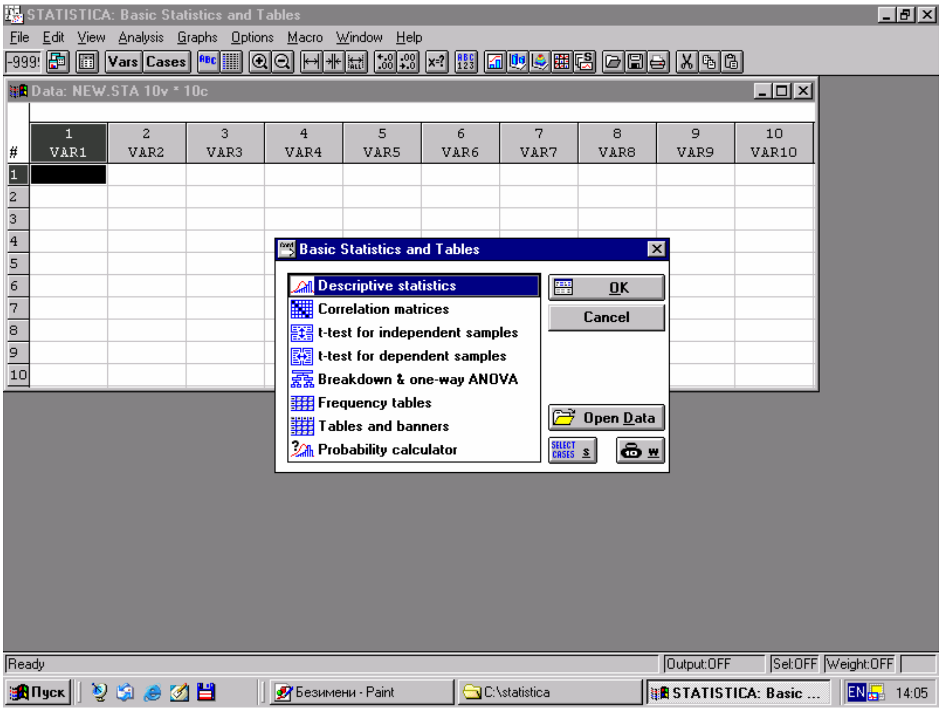


Рис.1

Таблица данных состоит из строк и столбцов. Столбцы используют для задания имен переменных (Variables), строки - для заполнения наблюдений (Cases). Строки и столбцы можно редактировать, выполнив двойной щелчок мыши на названии строки, столбца.

Диалоговое окно задания переменной позволяет:

а) name - задать имя переменной,

б) column width - ширину столбца в символах,

в) decimals - количество знаков после запятой, г) category - тип данных (например, number - числовой).

Строки, столбцы в таблице можно добавлять(Add), удалять(Delete), перемещать(Move) и др. Данные действия выполняются при помощи команд Edit/Variables (работа со столбцами), Edit/Cases (работа со строками).

Например, нужно добавить один столбец после столбца с переменной So. Окно диалога будет выглядеть, как на рис. 2.

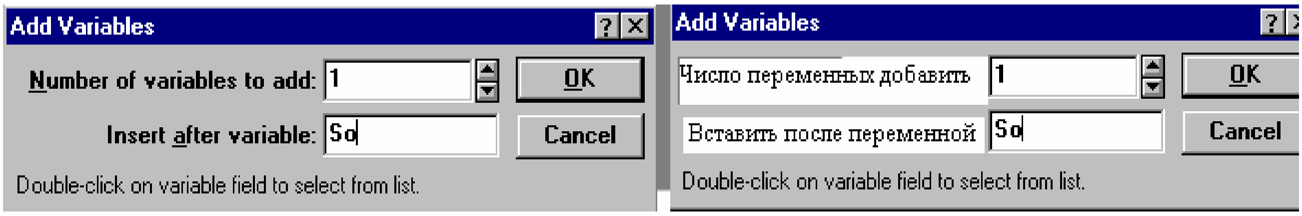


Рис.2

Задания:

1. Выберите пункты меню Пуск/Программы/ Statistica/ Basic Statistics. Откройте новый файл, выбрав пункт File/New Data. Сохраните файл с англоязычным именем.

2. Закройте окно пунктов меню, откройте в горизонтальном меню команду Analysis. Убедитесь, что пункты меню переместились в подменю Analysis.

3. Задайте переменные:

Price: ширина столбца - 6, количество знаков после запятой - 0;

So: ширина столбца - 5, количество знаков после запятой - 1;

R: ширина столбца - 4, количество знаков после запятой - 1.

4. В вашей таблице 10 строк. Добавьте две строки, чтобы у Вас получилось 12 наблюдений. Оставьте в таблице три столбца, остальные удалите. В итоге у Вас получится таблица размером 3x12.

Лабораторная работа №2

Модель парной регрессии.

Любая экономическая политика заключается в регулировании экономических переменных, и она должна базироваться на знании того, как эти переменные связаны с другими. Рассмотрим статистическую зависимость между двумя переменными x и y, т.е. такую зависимость при которой изменение одной из величин влечет за собой изменение закона распределения другой величины, в частности, такую зависимость, которая проявляется в том, что с изменением одной переменной изменяется среднее значение другой. Эту статистическую зависимость называют корреляционной.

Для нахождения значений основных статистик, в том числе и среднего значения, в программе *Statistica* необходимо выбрать пункт *Analysis/Descriptive Statistiscs*. В открывшемся окне диалога выбрать переменные, которые будут анализироваться (кнопка "*Variables*"). Клавиша "*Detailed Descriptive Statistisc*" вычисляет выборочное среднее значение (*Mean*), минимальное значение (*min*), максимальное значение *(max)* и др.

Можно указать два варианта рассмотрения взаимосвязей между двумя переменными. В первом случае обе переменные считаются равноценными, они не подразделяются на зависимую и независимую. Основным в этом случае является вопрос о наличии и силе взаимосвязи между переменными. При исследовании силы линейной связи обращаются к корреляционному анализу, основной мерой в котором является коэффициент корреляции.

Для нахождения коэффициента корреляции используется пункт *Analysis/Correlation matrices,* который позволяет просмотреть корреляционную матрицу, т.е. матрицу элементами которой являются коэффициенты корреляции, вычисленные для выбранных переменных по формуле:

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2} \cdot \sqrt{\overline{y^2} - \overline{y}^2}},$$

где $\overline{x}$, $\overline{y}$ - средние значения переменных.

Клавиша "*Two lists*" используется для выбора переменных, размещаемых в корреляционной матрице по строкам (first var) и столбцам (second var).

Коэффициент корреляции – это величина, позволяющая определить силу и тесноту линейной зависимости между двумя переменными.

Линейная связь между двумя переменными тем сильнее, чем ближе $|r_{xy}|$ к 1. Если $|r_{xy}|$=0, то линейной связи между x и y нет.

Положительное значение коэффициента корреляции говорит о прямой зависимости между переменными, а отрицательное – об обратной.

Другой вариант рассмотрения взаимосвязи между переменными выделяет одну из величин как независимую *x*, а другую как зависимую *y.* И изучает влияние этих переменных друг на друга.

Зависимость *y=f(x)* называется функцией регрессии *y* на *x*.

Если рассматривается зависимость двух величин, то регрессия называется парной.

Для определения вида парной регрессии в декартовой системе координат строят точки наблюдений и соединяют их отрезками. Полученную линию называют эмпирической линией регрессии.

По внешнему виду эмпирической линии регрессии определяют плавную кривую, около которой группируются все точки наблюдений. Эту кривую называют теоретической линией регрессии или регрессией.

Самой простой парной регрессией является линейная регрессия с уравнением: $y = \alpha + \beta \cdot x + \varepsilon$, где α и β − коэффициенты регрессии, которые находятся с помощью метода наименьших квадратов, ε – случайный член.

β − показывает на сколько изменяется *y* при изменении переменной *x* на 1 единицу;

α −это первоначальное значение *y* при *x=0.*

В программе *Statistica* пункт *Graphs/Stat 2D Graphs/ Scatterplot* позволяет отображать в декартовой системе координат диаграмму рассеяния точек наблюдения и применяет метод наименьших квадратов для построения модели. После выбора пункта меню открывается диалоговое окно, в котором задаются:

- переменные по осям OX (независимая переменная) и OY (зависимая переменная);
- выбирается модель: *Regular* - парная, *Multiple* - множественная и др.;
- вид регрессионной модели: *Linear* - линейная, *Logarifmic* - логарифмическая, *Exponention* - экспоненциальная и др.

Задания: Имеются данные о стоимости однокомнатных квартир в г. Благовещенске в период январь-март 2002 г.

Price - цены, тыс. руб.

So - общая площадь, м^2

R - расстояние до центра, км (центр - площадь Ленина)

	PRICE	SO	R
1	460	46	7
2	350	44	3,5
3	490	57,6	5
4	470	53,1	1,2
5	350	50	5
6	450	52,1	1,5
7	300	48	9
8	370	53	6,5
9	380	49	5
10	430	42	1,2
11	400	44	2
12	420	41	0

1. *занесите имеющиеся данные в таблицу;*
2. *найдите выборочное среднее значение переменных;*
3. *дайте определение коэффициента корреляции. Постройте корреляционную матрицу, в которой в строке находится переменная Price, в столбцах переменные So и R. Найдите коэффициенты корреляции r(Price, So) и r(Price,R), объясните их значения.*
4. *отобразить на графиках точки наблюдений Price(So), Price(R) зарисуйте схематично графики в тетрадь;*
5. *оцените регрессии:*

   $Price = \alpha + \beta * So$

   $Price = \alpha + \beta * R$
6. *дайте экономическое объяснение полученным регрессиям.*

Лабораторная работа №3.

Модель множественной регрессии.

На любой экономический показатель действует чаще всего не один, а несколько факторов, влиянием которых невозможно пренебречь. В этом случае следует все их включить в модель, т.е. построить уравнение множественной регрессии. Рассмотрим самую употребляемую и наиболее простую из моделей множественной регрессии - модель множественной линейной регрессии вида:

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_n \cdot x_n + \varepsilon ,$$

где y – зависимая переменная (*depended var*),

$x_1, x_2, \ldots, x_n$ – независимые переменные (*undepended var),*

ε –случайная величина.

После того как построено уравнение регрессии, необходимо оценить значимость как уравнения в целом, так и отдельных его параметров.

1. Статистическая значимость коэффициентов множественной линейной регрессии проверяется на основе *t*- статистики по основной схеме проверки статистических гипотез.

Формулируются нулевая и альтернативная гипотезы:

$H_0 : \beta_i = 0$ - коэффициент β_i статистически незначим*,* т.е. переменная x_i линейно не связана с зависимой переменной *y*. Её наличие среди объясняющих переменных не оправдано со статистической точки зрения. Не оказывая серьезного влияния на зависимую переменную она лишь искажает реальную картину взаимосвязи.

$H_1 : \beta_i \neq 0$ коэффициент β_i статистически значим.

С помощью компьютерного пакета определяются $t_{набл}$ для каждой переменной, включенной в модель, а по таблицам распределения Стьюдента при заданном уровне значимости α - $t_{крит}$ имеющее *n-k-1* степеней свободы, где *n* – объём выборки, *k* – количество независимых переменных.

Строятся доверительные интервалы, и выделяется область принятия нулевой гипотезы:

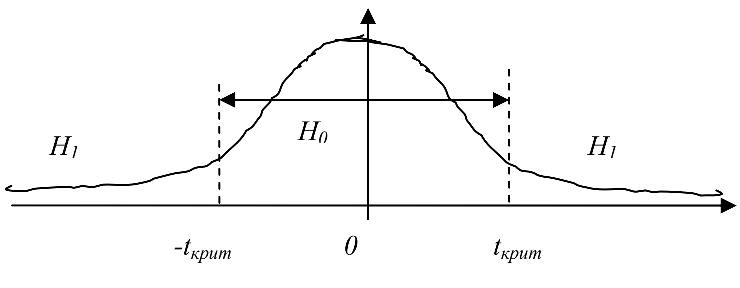


Рис.3

Если $t_{набл}$ попадает в область принятия нулевой гипотезы, то соответствующий параметр считаем статистически незначимым.

2. После проверки значимости каждого коэффициента регрессии проверяется общее качество уравнения. Для этой цели используется коэффициент детерминации R^2, который показывает долю дисперсии объясненной регрессией. Чем больше дисперсии объясняется регрессией, тем значимее уравнение регрессии, тем ближе значение R^2 к 1.

Будем считать, что регрессия имеет хорошее качество, если $R^2 \in [0,6;1]$.

3. Оценив индивидуально каждый коэффициент необходимо проанализировать их совокупную значимость.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_1 : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0$

Для проверки данных гипотез используется следующая F – статистика:

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

где R^2 - коэффициент детерминации,

n – объём выборки,

k – количество независимых переменных.

$F_{крит}$ рассчитывается по таблице распределения Фишера при заданном уровне значимости α и числе степеней свободы $v_1=n-k-1;\ v_2=k.$

Если нулевая гипотеза не отклоняется , то делается вывод о том, что совокупное влияние всех объясняющих переменных модели на зависимую переменную y можно считать несущественным, а качество модели невысоким.

Для оценки коэффициентов множественной регрессии в программе *Statistica* используется модуль *Multiple Regression*, который открывается пунктами *Analysis /Other Statistics / Multiple Regression.* Модуль открывается в новом окне, поэтому на панели задач располагаются два окна *Basic Statistics* и *Multiple Regression,* между которыми можно переключаться.

Первоначально в окне множественной регрессии появляется окно диалога, в котором кнопка *«Variables»* позволяет указать зависимые и независимые переменные.

После нажатия на клавишу ОК появляется окно результата множественной регрессии.

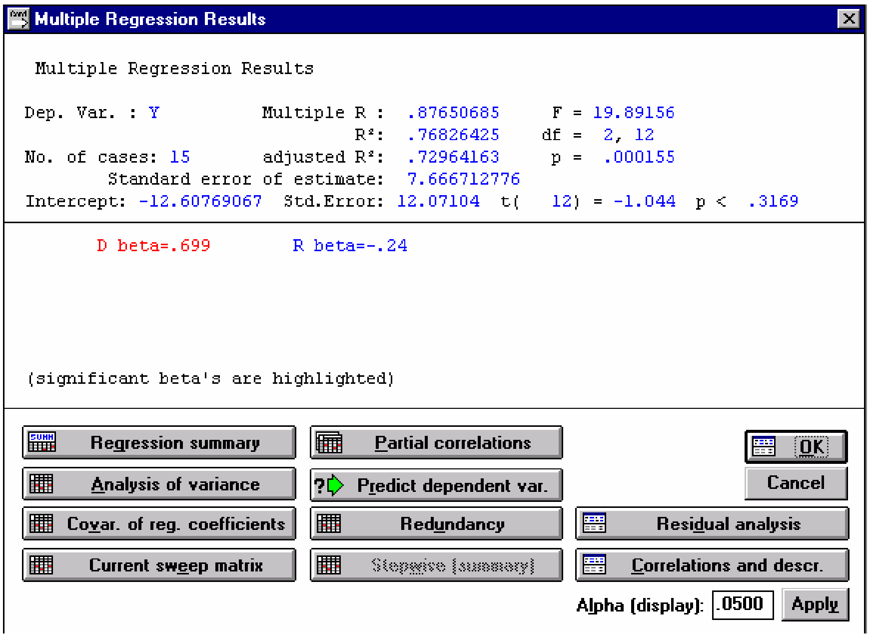


Рис.4

Dep var- зависимая переменная;

No of cases – количество наблюдений (объём выборки);

R^2 - коэффициент детерминации;

R^2_{adj} - скорректированный коэффициент детерминации;

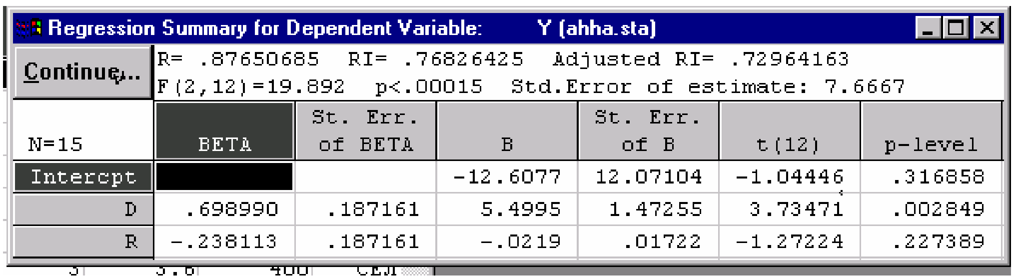
F – статистика;

dF(k, n-k-1) – число степеней свободы критерия Фишера;

Standart Error of estimate – стандартная ошибка регрессии;

p – уровень значимости (вероятность ошибки прогноза).

Оценки коэффициентов модели появляются при нажатии клавиши *«Regression Summary»:*



N=15	BETA	St. Err. of BETA	B	St. Err. of B	t(12)	p-level
Intercpt			-12.6077	12.07104	-1.04446	.316858
D	.698990	.187161	5.4995	1.47255	3.73471	.002849
R	-.238113	.187161	-.0219	.01722	-1.27224	.227389

Рис.5

Третий столбец *«B»* показывает оценку коэффициентов, столбец *«St.Err.of B»* указывает ошибку измерения коэффициентов, а пятый столбец – вычисленную t – статистику.

Согласно приведенным в таблице данным модель имеет вид:

$$Y = -12.6077 + 5.4995D - 0.0219R$$

Задания:

Имеется файл с данными о стоимости квартир в г. Благовещенске в период с января по март 2003г.

Price - цена, тыс.руб.;

So - общая площадь, м^2;

Sg - жилая площадь, м^2;

H - этаж (0 – крайний этаж, 1 – средний этаж);

R - расстояние до центра, км;

T - тип дома (0 – дом панельный, 1 – дом кирпичный);

Bal - балкон (0 – нет балкона, 1 - иначе);

W - горячая вода (0 – нет горячей воды, 1 - иначе);

O - очаг (0 – газ,1 – электрическая плита) ;

Plan - планировка (0 – старая, 1 - новая);

Y - срок эксплуатации, год;

Kol - количество комнат.

1. *Откройте Basic Statistics. Выберите пункт File /Open Data для открытия файла.*
2. *Выпишите выборочное среднее значение переменных: Y, So, R, Price.*
3. *Постройте точечный график зависимости цены от срока эксплуатации.*
4. *Постройте корреляционную матрицу и выпишите значения коэффициентов корреляции:*
5. *Постройте модель зависимости цены от любых 4 переменных. Запишите полученную модель в тетрадь. Объясните значение R^2. Проведите F – статистику, укажите проверяемые гипотезы.*

6. *Просмотрите результаты оценки коэффициентов. Выпишите значения t – статистик, укажите проверяемые гипотезы, проверьте их для каждого коэффициента.*

7. *Спрогнозируйте цену квартиры, значения переменных задайте самостоятельно.*

Для прогнозирования используйте клавишу *«Predict depended var».*

Лабораторная работа № 4

Спецификация модели.

Под спецификацией модели подразумевается построение такой эконометрической модели, которая удобна для проведения анализа. Правильно специфицированное уравнение регрессии верно отражает соотношения между экономическими показателями, участвующими в модели.

Неправильный выбор функциональной формы уравнения или набора объясняющих переменных называется ошибками спецификации. Рассмотрим основные типы ошибок спецификации.

1. В модели отсутствуют значимые переменные.
2. В модель включены несущественные переменные.
3. Выбор неправильной функциональной формы.

Ошибки спецификации допускаются в основном из-за поверхностных знаний об исследуемом экономическом объекте либо из-за погрешностей сбора и обработки статистических данных при построении уравнения регрессии. Важно уметь обнаружить и исправить эти ошибки.

Правильный выбор функциональной формы модели осуществляется при согласовании эмпирических данных с теорией. Однако чем сложнее форма модели, тем менее интерпретируемы ее параметры.

Рассмотрим способ обнаружения несущественной переменной в модели на примере.

Пусть теоретическая модель имеет вид:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (1)$$

Исследователь заменяет ее более сложной моделью, добавляя новую переменную x_3:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \qquad (2)$$

Для ответа на вопрос является ли переменная x_3 существенной необходимо сравнить следующие характеристики моделей:

а) Скорректированный коэффициент детерминации R^2_{adj}.

Коэффициент детерминации R^2 показывает долю дисперсии объясненную регрессией. Чем больше факторов включается в модель, тем большая часть дисперсии объясняется, поэтому при включении даже несущественной переменной можно заметить увеличение коэффициента детерминации R^2. Для устранения этого эффекта при анализе переменных используют скорректированный коэффициент детерминации:

$R^2_{adj} = 1 - (1 - R^2)\dfrac{n-1}{n-k-1}$, где n – количество наблюдений, k – количество объясняющих переменных.

R^2_{adj} корректируется в сторону уменьшения с ростом числа объясняющих переменных и увеличивается при добавлении новой переменной только тогда, когда добавленная переменная является значимой.

б) Статистика Фишера (F – статистика).

F – статистика позволяет проверить гипотезу об обоснованности включения новых переменных в модель.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1 : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0$$

Нулевая гипотеза принимается в том случае если расчетное значение F – статистики меньше соответствующего критического значения статистики Фишера, найденного по таблице критических точек распределения. Чем больше наблюдаемое значение F - статистики, тем более значима вся совокупность параметров регрессии.

Итак, если при включении переменной x_3 F – статистика увеличивается, то эту переменную можно считать существенной.

в) Среднее квадратическое отклонение σ.

Среднее квадратическое отклонение показывает среднюю величину отклонения реальных значений от линии регрессии. Чем меньше отличаются эмпирические и теоретические значения, тем меньше значение σ и качественнее модель.

С добавлением значимой переменной в модель значение σ уменьшается.

Итак, если переменная x_3 значима, то в первой модели имеем неверную спецификацию. Оценки параметров β_1, β_2 будут смещены от истинного значения.

Если в уравнении регрессии имеется несущественная переменная, то она обнаружит себя по низкой t – статистике. В дальнейшем эту переменную следует исключить из рассмотрения. Однако добавлять и исключать переменные целесообразно по одной и сравнивать качество полученных уравнений регрессии. При сравнении качества двух моделей обязательным является требование, чтобы зависимая переменная была представлена в одной и той же форме и число наблюдений для обеих моделей было одинаковым.

Задания.

1. *Откройте файл Kvartir.sta$*
2. *постройте регрессионную модель зависимости цены от всех переменных;*
3. *запишите в тетради вид модели;*
4. *дайте экономическое объяснение коэффициентам при фиктивных переменных;*
5. *занесите полученные данные в таблицу 1;*
6. *закройте окна с результатами регрессии;*
7. *выберите Analysis / Startup panel и постройте вторую модель исключив одну несущественную переменную, занесите результат в таблицу1;*
8. *сравните качество моделей;*
9. *постройте третью, четвертую, и т.д. модель, результаты занесите в таблицу 1, выберите наилучшую модель.*

Таблица 1.

№	α	$\beta_1(SO)$	$\beta_2(SG)$	$\beta_3(H)$	…	$\beta_n(\ )$	$t_{0,95}(n-k-1)$	R^2	R^2_{adj}	F	σ
1.	*значение (ошибка) t-статистика*	*32 (3,1) t=2,7*									
2.											
3.											

Используйте для выбора нескольких несмежных переменных клавишу CTRL.

Лабораторная работа №5.

Тест Чоу.

При более детальном изучении модели на этапе спецификации требуется определить, совпадают ли уравнения регрессии для отдельных групп наблюдений. Распространенным тестом для поверки данной гипотезы является тест Чоу, суть которого состоит в следующем:

Пусть имеются две выборки объёма n_1 и n_2. Для каждой из этих выборок оценено уравнение вида:

$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n$ - для n_1

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$ - для n_2

Проверяется нулевая гипотеза о равенстве друг другу соответствующих коэффициентов регрессий, т.е. другими словами будет ли уравнение регрессии одним и тем же для обеих выборок.

$H_0 : \alpha_0 = \beta_0;\ \alpha_1 = \beta_1;\ \alpha_2 = \beta_2;\ \ldots;\ \alpha_n = \beta_n.$

$H_1 : \alpha_0 \neq \beta_0;\ \alpha_1 \neq \beta_1;\ \alpha_2 \neq \beta_2;\ \ldots;\ \alpha_n \neq \beta_n.$

Для проверки нулевой гипотезы используется F- статистика:

$$F_{набл.} = \frac{ESS_0 - ESS_1 - ESS_2}{ESS_1 + ESS_2} \cdot \frac{n_1 + n_2 - 2k - 2}{k+1},$$ где

ESS_1, ESS_2 - доли дисперсии необъясненные регрессиями , построенными для выборок объёмов n_1 и n_2 соответственно;

k – количество независимых переменных, входящих в уравнение.

ESS_0 - доля дисперсии необъясненная регрессией, построенной для объединенной выборки объёма (n_1 + n_2).

По таблице Фишера определяется $F_{крит.}$, которое имеет следующие степени свободы $v_1=k+1$ $v_2=n_1+n_2-2k-2$.

Строим доверительные интервалы и выделяем область принятия нулевой гипотезы при заданном уровне значимости α.

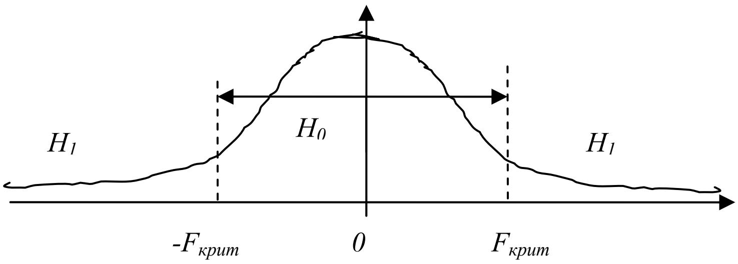


Рис.6

Если $F_{набл.}$ близка к нулю, то это означает, что коэффициенты регрессии совпадают и нецелесообразно рассматривать две различные регрессии для двух групп наблюдений. В противном случае, говорят о наличии структурного сдвига в модели и используют для прогнозирования модели для подвыборок.

Для отбора наблюдений в порграмме *Statistica* используется клавиша «*Select Cases*» , которая появляется при открытии стартовой панели для построения множественной регрессии. Например, нужно построить регрессию для 1 – комнатных квартир в этом случае отбор переменных будет происходить при условии *Kol=1* (см рис.).

Для определения значений ESS_i в диалоговом окне *Multuple Regression Result* нажмите клавишу *Analisis of Variance* , появится таблица первый столбец которой состоит из сумм квадратов отклонений:

Analysis of Variance (ahha.sta)

Continue...	Sums of Squares	df	Mean Square	F	p-level
Regress.	2338.392	2	1169.196	19.89156	.000155
Residual	705.342	12	58.778		
Total	3043.733				

Рис.7

- *Regression Sums of Squares (RSS)* – объясненная регрессией часть дисперсии;
- *Residual Sums of Squares (ESS)* – необъясненная регрессией часть дисперсии;
- *Total Sums of Squares (TSS)* – общая дисперсия.

Задания:

1. имеется предположение о наличии структурного сдвига ценообразования для 1-комнатных, 2-комнатных, 3-комнатных и 4- комнатных квартир. Запишите в тетради сущность теста Чоу для проверки этого предположения.

2. постройте отдельно регрессии зависимости цены (PRICE) от переменных SO, R, Y для 1-комнатных. 2 – комнатных, 3,4 – комнатных квартир. Осуществите тест Чоу.

Лабораторная работа №6.

Мультиколлинеарность.

Мультиколлинеарность- это наличие линейной связи между двумя или несколькими объясняющими переменными. Мультиколлинеарность бывает полная и частная.

Полная мультиколлинеарность наблюдается в случае, когда объясняющие переменные связаны строгой функциональной зависимостью ($r_{xy}=1$). Полная мультиколлинеарность является скорее теоретическим примером, на практике встречается редко.

Частичная мультиколлинеарность наблюдается в случае, когда объясняющие переменные связаны сильной корреляционной зависимостью, но эта зависимость не является функциональной ($0,5<r_{xy}<1$).

Если в модели присутствует полная мультиколлинеарность, то оценки параметров регрессии определить невозможно.

В случае частичной мультиколлинеарности оценки параметров регрессии существуют, но обладают «плохими» свойствами, т.е. имеют большие ошибки и низкие *t*- статистики.

Признаки мультиколлинеарности:

1. Коэффициент детерминации достаточно высок, но некоторые из коэффициентов регрессии статистически незначимы, т.е. они имеют низкие t – статистики.

2. Небольшое изменение исходных данных приводит к существенному изменению оценок коэффициентов регрессии.

3. Оценки коэффициентов имеют неправильные с точки зрения теории знаки или неоправданно большие значения.

4. Частные коэффициенты корреляции имеют большие значения.

Частные коэффициенты корреляции определяют силу линейной зависимости между двумя переменными без учета влияния на них других переменных.

Для нахождения частного коэффициента корреляции предположим, что имеется регрессионная модель:

$$y = \alpha + \beta_1 \ x_1 + \beta_2 \ x_2 + \varepsilon$$

Наша цель – определить корреляцию между y и переменной x_1, исключив влияние переменной x_2.

1) Осуществим регрессию y на x_2 и получим прогнозные значения

$$\hat{y} = \alpha_1 + \alpha_2 \ x_2$$

2) Осуществим регрессию x_1 на x_2 и получим прогнозные значения $\hat{x}_1 = \beta_1 + \beta_2 \ x_2$.

3) Удалим влияние x_2 взяв остатки $\varepsilon_y = y - \hat{y}$ и $\varepsilon_{x_1} = x_1 - \hat{x}_1$.

4) Определим частный коэффициент корреляции между переменными y и x_1, после исключения влияния переменной x_2, как обычный коэффициент корреляции между остатками ε_y и ε_{x_1}.

Частный коэффициент корреляции обозначают $r(yx_1,x_2)$ после запятой в скобках указываются те переменные, влияние которых исключается.

В случае двух независимых переменных справедлива формула:

$$r(yx_1,x_2)=\frac{r(y,x_1)-r(y,x_2)\cdot r(x_1,x_2)}{\sqrt{1-r^2(y,x_2)}\cdot\sqrt{1-r^2(x_1,x_2)}}$$

Значение $r(yx_1,x_2)$ лежит в интервале [-1,1] как и у обычного коэффициента. Равенство $r(yx_1,x_2)$ нулю означает отсутствие линейного влияния переменной x_1 на y.

В случае, когда исключается влияние не одной, а нескольких переменных достаточно переменную x_2 заменить набором переменных.

5.Сильная вспомогательная регрессия.

Для анализа строятся уравнения регрессии каждой из независимых переменных на оставшиеся независимые переменные. Вычисляются соответствующие коэффициенты детерминации R^2 и рассчитывается их статистическая значимость на основе F- статистики:

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

n – число наблюдений,

k – число независимых переменных в первоначальном уравнении регрессии.

Статистика $F_{крит}$ имеет распределение Фишера с $к-1$ и $n-к$ степенями свободы (определяется по таблице).

Если $-F_{крит}<F<F_{крит}$, то соответствующий коэффициент детерминации статистически не значим, значит переменная x не является линейной комбинацией других переменных её можно оставить в уравнении регрессии. В противном случае можно считать, что x существенно зависит от других переменных и имеет место мультиколлинеарность.

Если основная задача модели – прогноз будущих значений, то при достаточно большом R^2 (>0,9) наличие мультиколлинеарности не сказывается на прогнозных качествах модели.

Если же целью исследования является выявление степени влияния каждой из независимых переменных на зависимую, то наличие мультиколлинеарности, приводящее к увеличению стандартных ошибок исказит истинные зависимости между переменными. В этой ситуации мультиколлинеарность является серьезной проблемой.

Методы устранения мультиколлинеарности.

1. Исключение переменных из моделей, однако в этой ситуации возможны ошибки спецификации.

2. Попробовать добавить новую переменную, которая возможно была упущена, однако, если она тоже будет иметь сильную зависимость от других переменных, то её введение может ещё больше усугубить проблему мультиколлинеарности.

3. Увеличить число наблюдений, однако получение новой выборки не всегда возможно или связано с серьезными издержками.

Задания:

1. Открыть файл Kvartir. sta

2. Найти частные коэффициенты корреляции переменной PRICE со всеми независимыми переменными, сравнить их с обычными коэффициентами корреляции;

3. Выбрать переменные, которые сильнее всего влияют на PRICE, и построить уравнение регрессии, включив в него эти переменные;

4. Найти частные коэффициенты корреляции независимых переменных, включенных в модель. Сделать предположение о наличии мультиколлинеарности в модели;

5. Проверить предположение о наличии мультиколлинеарности с помощью вспомогательной регрессии;

6. Избавиться от мультиколлинеарности путем удаления коррелированных переменных из модели. Полученное уравнение регрессии записать в тетрадь и дать экономическую интерпретацию коэффициентам регрессии.

Лабораторная работа № 7.

Гетероскедастичность.

Рассмотрим модель множественной регрессии: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$.

переменная y зависит не только от объясняющих переменных x_i , коэффициентов регрессии α, β_i, но и от случайной величины ε. Для того, чтобы регрессия имела хорошее качество необходимо, чтобы разброс значений случайного члена был стабильным, т.е. дисперсия ошибки была независимой от переменных модели ($\sigma^2=const$).

Условие независимости дисперсии ошибки называется гомоскедастичностью.

Если разброс остатков подчиняется некоторому закону, то наблюдается Гетероскедастичность в модели ($\sigma^2 \neq const$).

Для обнаружения гетероскедастичности существует несколько тестов и критериев, но все они проверяют справедливость гипотез:

$H_0 : \sigma^2 = const$ - модель гомоскедастична;

$H_1 : \sigma^2 \neq const$ - модель гетероскедастична.

1.Графический анализ остатков.

По оси абсцисс откладывают значения объясняющей переменной, а по оси ординат либо отклонения ε_i, либо их квадраты ε_i^2.

Если все отклонения находятся внутри полосы постоянной ширины, параллельной оси абсцисс, то модель гомоскедастична.

Для анализа остатков в окне *Multiple Regression* нажмите клавишу *Results Residual analysis*, а затем в открывшемся окне *Display residuals $ pred.* Третий столбец таблицы *Residual* показывает остатки ε_i.

Predicted & Residual Values (ahha.sta)

Continue... Dependent variable: Y

Case No.	Observed Value	Predictd Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val
1	10.00000	15.99879	-5.99879	-.06715	-.78245	3.286880
2	50.00000	29.74325	20.25675	.99634	2.64217	2.936944
3	0.00000	6.65839	-6.65839	-.78987	-.86848	3.559447
4	40.00000	39.74978	.25022	1.77060	.03264	4.205288
5	5.00000	13.82545	-8.82545	-.23532	-1.15114	3.954526
6	20.00000	20.96604	-.96604	.31719	-.12600	3.466517
7	3.00000	-1.57328	4.57328	-1.42681	.59651	3.531309
8	2.00000	-2.12765	4.12765	-1.46970	.53839	3.777624
9	8.00000	7.10541	.89459	-.75529	.11668	2.595940
10	12.00000	8.88469	3.11531	-.61761	.40634	5.061101
11	20.00000	22.05271	-2.05271	.40127	-.26774	2.156269
12	6.00000	5.01735	.98265	-.91685	.12817	2.810307
13	30.00000	27.76253	2.23747	.84308	.29184	2.966625
14	20.00000	24.24811	-4.24811	.57115	-.55410	2.448325
15	27.00000	34.68843	-7.68843	1.37897	-1.00283	3.529928
Minimum	0.00000	-2.12765	-8.82545	-1.46970	-1.15114	2.156269
Maximum	50.00000	39.74978	20.25675	1.77060	2.64217	5.061101

Рис.8

2.Тест Уайта (White, 1980г.).

К обычной модели применяют обычный метод наименьших квадратов и находят остатки ε_i . Осуществляется регрессия ε_i^2 на все независимые переменные x_i. Если построенная регрессия имеет хорошее (коэффициент детерминации близок к 1, F – статистика имеет большое значение), то нулевая гипотеза не принимается и говорят о наличии гетероскедастичности в модели.

3. Тест Голдфелда – Квандта (Goldfeld - Quandt).

Суть этого теста состоит в следующем:

1) Все n наблюдений упорядочиваются по независимой переменной x_i.

2) Упорядоченная выборка визуально разбивается на три части размерностей m, $n\text{-}2m$, m соответственно.

3) Оцениваются отдельно регрессии для первой подвыборки (m первых наблюдений) и для третьей (m последних наблюдений). Если предположение о наличии гетероскедастичности верно, то дисперсия регрессии по первой подвыборке ESS_1 будет значительно меньше дисперсии регрессии по третьей подвыборке ESS_3 .

4) Для сравнения соответствующих дисперсий строится следующая F – статистика:

$$F_{набл} = \frac{ESS_3}{ESS_1},$$

по таблице распределения Фишера при заданном уровне значимости α находится $F_{крит}$ с числами степеней свободы $v_1=v_2=m\text{-}k\text{-}1,$ где m – объем подвыборки, k – количество независимых переменных в уравнении регрессии.

5) Если $-F_{крит} < F_{набл} < F_{крит}$, то принимаем нулевую гипотезу, т.е. модель гомоскедастична, в противном случае говорим о наличии гетероскедастичности.

Для упорядочения наблюдений по переменной в *Basic Statistics* выберите *Analysis / Other Statistics/ Data Management/ Sort Cases*. Укажите переменную, по которой хотите упорядочить наблюдения.

Задания:

1. Откройте файл GDP. sta:

EE- государственные расходы на образование;

GDP – валовой внутренний продукт;

P – численность населения.

2. Постройте модель линейной множественной регрессии EE(GDP, P). Запишите полученную модель в тетрадь. Просмотрите остатки ε_i в модели. В *рабочем документе (файл GDP) добавьте переменную RES, которая рассчитывает остатки как разность между реальным значением EE и ее прогнозным значением, полученным в модели.*

Сравните ε_i и RES.

3. Постройте графики зависимостей $(RES)^2$ от GDP и $(RES)^2$ от P. Сделайте предположение о наличии гетероскедастичности в модели для каждой переменной.

4. Проверьте наличие гетероскедастичности в модели по каждой переменной используя тесты Уайта и Голдфелда – Квандта.

Лабораторная работа № 8.

Взвешенный метод наименьших квадратов.

Допустим, что модель $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$ гетероскедастична по переменной x_1. Значит дисперсии отклонений зависят от переменной x_1. Предположим, что σ пропорционально x_1.

Для устранения гетероскедастичности воспользуемся взвешенным методом наименьших квадратов (ВМНК). Он применяется к преобразованным данным – взвешенным по гетероскедастичному регрессору x_1.

Разделим обе части равенства $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$ на переменную x_1 получим:

$$\frac{y}{x_1} = \frac{\alpha}{x_1} + \beta_1 + \beta_2 \frac{x_2}{x_1} + \ldots + \beta_n \frac{x_n}{x_1} + \frac{\varepsilon}{x_1}$$

Переобозначим $\frac{y}{x_1} = y^*, \quad \frac{1}{x_1} = x_1^*, \quad \frac{x_2}{x_1} = x_2^*, \ \ldots, \quad \frac{x_n}{x_1} = x_n^*, \quad \frac{\varepsilon}{x_1} = \varepsilon^*,$

применим к преобразованным данным обычный метод наименьших квадратов и получим модель вида: $y^* = \beta_1 + \alpha \cdot x_1^* + \beta_2 \cdot x_2^* + \ldots + \beta_n \cdot x_n^* + \varepsilon^*$.

В ней остатки не зависят от x_1, а значит имеют постоянную дисперсию, т.е. модель гомоскедастична. В этой модели β_1 играет роль постоянного числа. Оценки коэффициентов $\alpha, \beta_1, \beta_2, \ldots \beta_n$ становятся точнее и, как правило, ниже. Для того чтобы использовать эту модель для дальнейшего анализа необходимо вернуться к прежним переменным.

Если в уравнении регрессии присутствуют несколько гетероскедастичных переменных, то в качестве «весов» можно использовать либо их линейную комбинацию, либо наиболее подходящую, исходя из графического представления.

Для применения взвешенного метода наименьших квадратов в программе *Statistica* достаточно на стартовой панели множественной линейной регрессии нажать на клавишу и в открывшемся окне задать переменную, по которой необходимо взвесить наблюдения.

Задание:

1. *Откройте файл GDP. sta;*
2. *Постройте модель множественной регрессии EE(GDP,P).*
3. *Устраните гетероскедастичность в модели по переменным GDP и P, используя взвешенный метод наименьших квадратов.*

Лабораторная работа № 9.

Автокорреляция остатков.

Временной ряд – это совокупность значений какого – либо показателя за несколько последовательных промежутков времени. Для временных рядов гетероскедастичность проявляется в виде автокорреляции остатков.

Автокорреляция может быть положительной и отрицательной. Чаще всего положительная автокорреляция вызывается направленным воздействием некоторых неучтенных в модели факторов.

Например, пусть исследуется спрос y на прохладительные напитки в зависимости от дохода x по ежемесячным данным. Зависимость, отражающая увеличение спроса с ростом дохода, может быть представлена линейной функцией. $y = \alpha + \beta \cdot x$

Однако фактические точки наблюдений обычно будут превышать линию графика в летние периоды и будут ниже ее в зимние.

Отрицательная автокорреляция означает, что за положительным отклонением следует отрицательное и наоборот.

На практике отрицательная автокорреляция встречается редко.

Наиболее известным критерием обнаружения автокорреляции является критерий Дарбина – Уотсона, общая схема применения которого следующая:

Для построенного уравнения регрессии определяют значение статистики Дарбина – Уотсона DW. По таблице критических точек распределения для заданного уровня значимости α, числа наблюдений n и количества независимых переменных k определяются два значения d_l - нижняя граница, d_u – верхняя граница.

Далее осуществляем выводы по правилу:

$0<DW<d_l$ – существует положительная автокорреляция;

$d_l<DW<d_u$ – зона неопределенности;

$d_u<DW<4-d_u$ – автокорреляция отсутствует;

$4-d_u<DW<4-d_l$ – зона неопределенности;

$4-d_l<DW<4$ – существует отрицательная автокорреляция.

Графически выводы можно представить следующим образом:

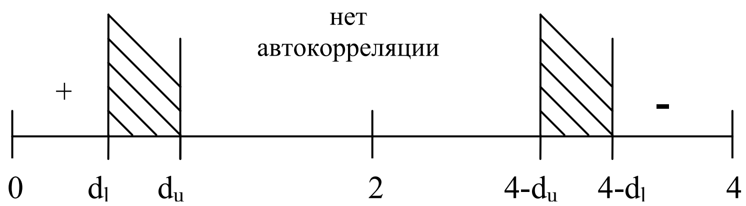


Рис.9

Для устранения автокорреляции необходимо, прежде всего, скорректировать саму модель. Возможно, автокорреляция вызвана отсутствием в модели некоторой важной объясняющей переменной. И добавление этой переменной поможет устранить автокорреляцию. Чаще всего приходится добавлять фиктивные переменные, отвечающие за сезонность.

Например, при изучении спроса на прохладительные напитки добавим переменную $s = \begin{cases} 0, & \text{если холодное время года}, \\ 1, & \text{если теплое время года}. \end{cases}$

Тогда y может быть представлено в виде: $y = \alpha + \beta_1 \cdot x + \beta_2 \cdot s$.

При s=0 y покажет объем продаж прохладительных напитков в холодное время года, а при s=1 – в теплое. Коэффициент β_2 показывает на сколько изменяется объем продаж в теплое время года по сравнению с холодным.

Иногда приходится добавлять несколько фиктивных переменных.

Задания:

1. *Откройте файл Avto.sta;*
2. *Постройте модель зависимости*
3. *Проверьте наличие автокорреляции в модели;*
4. *Добавьте 4 фиктивных переменных, отвечающих за времена года (зима, весна, лето, осень), постройте новую модель, включив в нее фиктивные переменные, проверьте наличие автокорреляции в этой модели.*
5. *Дайте экономическую интерпретацию всем коэффициентам модели.*

Литература.

1. Доугерти К. Введение в эконометрику. М., 1997г.
2. Магнус Я., Катышев П., Пересецкий А. Эконометрика. Начальный курс. М., 1997г.
3. Груббер Й. Эконометрия. К., 1996 г.
4. Бородич С.А. Эконометрика. Минск, 2001г.

Приложения.

1. Данные о стоимости квартир в г. Благовещенске в 2003г *(файл kvartir.sta)*.

№	PRICE	SO	SG	H	R	T	BAL	W	O	PLAN	Y	KOL
1	450	35	18	1	3,5	0	1	1	1	1	17	1
2	450	29	18	0	0,5	1	0	1	0	0	36	1
3	360	31	18	0	1	0	0	1	0	0	30	1
4	510	41	22	1	1	0	1	1	1	1	7	1
5	500	32	18	1	1,5	1	1	1	1	1	9	1
6	370	28	13	0	1,7	1	0	1	1	0	40	1
7	470	36	18	0	3,2	0	1	1	1	1	15	1
8	320	32	18	1	6	0	1	0	0	0	33	1
9	539	34	18	0	1,5	1	1	1	1	1	6	1
10	360	27	17	1	0	1	0	0	0	0	32	1
11	700	50	30	0	3	1	1	1	0	1	22	2
12	570	45	31	0	0,8	0	0	1	0	0	30	2
13	530	54	37	1	0	1	0	1	1	0	27	2
14	650	54	32	1	1,2	0	1	1	1	1	9	2
15	550	53	31	0	3	0	0	1	1	1	15	2
16	700	48	28	1	0,6	0	1	1	0	0	25	2
17	650	55	32	0	0,9	0	0	1	0	0	25	2
18	530	42	27	0	0	1	1	1	0	0	30	2
19	570	46	32	1	0,6	0	1	1	0	0	30	2
20	550	42	29	0	0	1	0	1	0	0	30	2
21	550	45	27	0	0	0	0	1	0	0	25	2
22	550	52	30	0	3	0	0	1	1	0	25	2
23	999	74	32	0	3	1	1	1	1	1	0	2
24	600	62	32	0	8	1	1	1	1	1	2	2
25	600	47	27	1	3	1	1	1	1	1	9	2
26	650	53	32	0	1,3	0	1	1	0	0	25	2
27	200	41	27	1	16	1	1	1	0	0	25	2
28	750	54	41	1	0	0	1	1	1	1	14	2
29	550	47	30	1	0,6	0	1	1	0	0	25	2
30	1380	102	50	1	0	1	1	1	1	1	9	2
31	620	60	42	1	3	1	1	1	0	0	40	3
32	930	65	40	1	0	0	1	1	0	1	12	3
33	960	71	51	0	0,7	1	1	1	1	1	14	3
34	1560	130	70	0	0	1	1	1	1	1	0	3
35	650	60	36	1	3,5	0	1	1	1	1	9	3
36	650	50	38	1	3	1	1	1	0	0	30	3
37	780	64	40	0	0	0	1	1	1	1	14	3
38	600	62	42	1	0,8	1	0	1	0	0	40	3
39	700	58	36	1	0	1	0	1	1	1	16	3
40	750	71	51	0	0,6	0	1	1	0	0	25	3
41	820	86	61	0	3	0	1	1	1	1	23	4
42	700	72	48	0	3,2	0	0	1	1	1	18	4
43	620	86	68	1	8	0	1	1	1	1	8	4
44	950	87	62	0	0,4	0	1	1	1	1	19	4
45	910	80	52	0	0	1	1	1	1	1	22	4
46	800	71	48	1	3	0	1	1	1	1	16	4
47	670	61	45	1	0,8	0	1	1	0	0	28	4
48	750	85	61	1	2,8	0	1	1	1	1	17	4
49	775	70	47	0	2,3	0	1	1	1	1	20	4
50	2460	205	150	1	0	1	1	1	1	1	0	4

2.Государственные расходы на образование (EE), валовой внутренний продукт (GDP) и численность населения (P) в выборке стран *(файл GDP. sta)*.

страна	EE	GDP	P
Люксембург	0,34	5,67	0,36
Уругвай	0,22	10,13	2,9
Сингапур	0,32	11,34	2,39
Ирландия	1,23	18,88	3,44
Израиль	1,81	20,94	3,87
Венгрия	1,02	22,16	10,71
Нов.Зеландия	1,27	23,83	3,1
Португалия	1,07	24,67	9,93
Гонконг	0,67	27,56	5,07
Чили	1,25	27,57	11,1
Греция	0,75	40,15	9,6
Финляндия	2,8	51,62	4,78
Норвегия	4,9	57,71	4,09
Югославия	3,5	63,03	22,34
Дания	4,45	66,32	5,12
Турция	1,6	66,97	44,92
Австрия	4,26	76,88	7,51
Швейцария	5,31	101,65	6,37
Саудавская Аравия	6,4	115,97	8,37
Бельгия	7,15	119,49	9,86
Швеция	11,22	124,15	8,31
Австралия	8,66	140,98	14,62
Аргентина	5,56	153,85	27,06
Нидерланды	13,41	169,38	14,14
Мексика	5,46	186,33	67,4
Испания	4,79	211,78	37,43
Бразилия	8,92	249,72	123,03
Канада	18,9	261,41	23,94
Италия	15,95	395,52	57,04
Великобритания	29,9	534,97	55,95
Франция	33,59	655,29	53,71
Германия	38,62	815	61,56
Япония	61,61	1040,45	116,78
США	181,3	2586,4	227,64

3.Ежемесячные данные зависимости между экспортом (EX) и импортом (IM) (*файл AVTO. sta)*.

	EX	IM
январь 2001	12,47	11,07
февраль 2001	12,65	11,5
март 2001	12,89	12,01
апрель2001	12,97	12,28
май 2001	10	13,16
июнь 2001	13,31	13,43
июль 2001	13,25	13,28
август 2001	12,65	13,5
сентябрь 2001	14,49	15,32
октябрь 2001	14,47	15,62
ноябрь 2001	14,74	17,44
декабрь 2001	14,62	16,14
январь 2002	17,6	16,13
февраль 2002	17,7	16,08
март 2002	16,6	16,55
апрель2002	15,26	15
май 2002	19,49	18,72
июнь 2002	19,08	17,8
июль 2002	18,69	16,64
август 2002	18,65	17,39
сентябрь 2002	19,33	18,7
октябрь 2002	19,11	18,02
ноябрь 2002	18,62	17,46
декабрь 2002	18,4	16,96
январь 2003	16,15	15,06
февраль 2003	16,58	16,01
март 2003	17,6	16,63
апрель 2003	18,48	17,86
май 2003	15,36	14,56
июнь 2003	15,25	15,64
июль 2003	15,61	16,45
август 2003	15,93	17,42
сентябрь 2003	14,38	14,3
октябрь 2003	14,3	14,59
ноябрь 2003	14,75	14,66
декабрь 2003	15,58	14,95

Оглавление

Введение. 1

Основы программы *Statistica*. 4

Модель парной регрессии. 5

Модель множественной регрессии. 8

Спецификация модели. 13

Тест Чоу. 16

Мультиколлинеарность. 18

Гетероскедастичность. 21

Взвешенный метод наименьших квадратов. 24

Автокорреляция остатков. 25

Литература. 27

Приложения. 28

Наталья Николаевна Двоерядкина
Ст. преподаватель кафедры ОМиИ АмГУ;
Татьяна Анатольевна Макарчук
Кандидат педагогических наук,
ст. преподаватель кафедры ОМиИ АмГУ;
Алена Николаевна Киселева
Ст. преподаватель кафедры ОМиИ АмГУ.

Лабораторный практикум по эконометрике для экономических специальностей