МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

КУРГАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Кафедра «Экономическое моделирование и информатика»

Эконометрические методы исследования моделей

Методические указания

по дисциплине «Эконометрика»

для студентов очной формы обучения специальностей

«Финансы и кредит» - 080105(060400),

«Бухучет, анализ и аудит» - 080109(060500)

Курган 2005

Кафедра: «Экономическое моделирование и информатика»

Дисциплина: «Эконометрика» для студентов очной формы обучения специальностей «Финансы и кредит» - 080105(060400), «Бухучет, анализ и аудит» - 080109(060500)

Составили: доцент Аликас В.Э., ассистент Чеботина М.С.

Работа выполнена при равноценном участии авторов

Методические указания утверждены на заседании кафедры: 15 апреля 2005 г.

Рекомендованы методическим советом университета
«____»_______________2005 г.

1 Экономические модели и статистические методы

1.1 Общие положения

Закономерности в экономике выражаются в виде связей и зависимостей экономических показателей, математических моделей, их поведения. Такие зависимости и модели могут быть получены только путем обработки реальных статистических данных, с учетом внутренних механизмов связи и случайных факторов. Модель может быть получена и апробирована на основе анализа статистических данных и изменения в поведении последних, говорят о необходимости уточнения и развития модели. Особенно важен эконометрический анализ в макроэкономике, где взаимосвязи величин зачастую неочевидны и изменчивы. Нередко встречается ситуация, когда модель перестает «работать» в связи с появлением или активизацией какого-то фактора, и такие ситуации обусловливают развитие макроэкономической теории. Поэтому предлагаемый материал «привязан» к макроэкономическим проблемам и моделям. Эконометрический анализ дает возможность обосновать и уточнить форму зависимостей в рассматриваемых макроэкономических моделях, лучше понять механизмы взаимосвязи макроэкономических показателей.

Основным элементом экономического исследования является анализ и построение ***взаимосвязей экономических переменных***. Изучение таких взаимосвязей осложнено тем, что они, особенно в макроэкономике, не являются строгими, функциональными зависимостями. Во-первых, всегда очень трудно выявить все основные факторы, влияющие на данную переменную. Во-вторых, многие такие воздействия являются случайными, то есть содержат случайную составляющую. В-третьих, экономисты, как правило, располагают ограниченным набором данных статистических наблюдений, которые к тому же содержат различного рода ошибки. ***Математическая статистика*** (то есть теория обработки и анализа данных) и ее применение в экономике - ***эконометрика*** - позволяют строить экономические модели и оценивать их параметры, проверять гипотезы о свойствах экономических показателей и формах их связи. Это, в конечном счете, служит основой для экономического анализа и прогнозирования, создавая возможность для принятия обоснованных экономических решений.

Любое эконометрическое исследование всегда предполагает объединение теории (экономической модели) и практики (статистических данных). Мы используем теоретические модели для описания и объяснения наблюдаемых процессов и собираем статистические данные с целью эмпирического построения и обоснования моделей.

1.2 Введение случайного компонента в экономическую модель

Обычно предполагают, что все факторы, не учтенные явно в экономической модели, оказывают на объект некоторое результативное воздействие, величина которого неизвестна заранее и может быть описана как случайная функция. Для ее описания в модель добавляют (обычно аддитивным образом) случайный параметр ε, интегрирующий в себе влияние всех неучтенных явно факторов. Например, в модели спроса:

$$q = f(p, I) + \varepsilon,$$

где:

q - количество блага;

p – цена;

I - доход потребителя;

ε - переменная, которая учитывает влияние всех прочих факторов (цен на другие товары, изменений моды, погоды и т. д.), не учтенных явно в функции спроса.

1.3 Статистические данные и стохастическая модель

Введение случайного компонента в экономическую модель приводит к тому, что взаимосвязь остальных ее переменных перестает быть строго детерминированной и становится стохастической, что и наблюдается в реальной действительности. Это, отчасти, делает модель доступной для эмпирической (опытной) проверки на основе статистических данных о конкретном экономическом объекте. Если проверка показала адекватность модели, то иногда удается оценить параметры функционирования конкретного экономического объекта и сформулировать рекомендации для принятия практических решений. Работа с эконометрическими моделями требует использования инструментария оценивания и статистической проверки модели («наука» моделирования), а также решения проблем выбора типа модели, набора объясняющих переменных и вида связей между ними («искусство» моделирования).

1.4 Экономические данные

Статистические данные являются основой для выявления и обоснования эмпирических закономерностей. Без конкретных количественных данных, характеризующих функционирование исследуемого экономического объекта, не всегда возможно определить практическую значимость применяемой экономической модели, даже если целью является выявление преимущественно качественных закономерностей.

Экономические данные обычно делят на два вида: перекрестные данные (cross-section data) и временные ряды (time series). ***Перекрестные данные*** - это данные по какому-либо экономическому показателю, полученные для разных однотипных объектов (фирм, регионов). При этом либо все данные относятся к одному и тому же моменту времени, либо их временная принадлежность несущественна. ***Временные ряды*** - это данные, характеризующие один и тот же объект, но в различные моменты времени. К первому типу, например, относятся данные бюджетных обследований населения в определенный момент времени; ко второму - данные о динамике уровня инфляции за определенный период. Данные временных рядов характеризуются определенными зависимостями и закономерностями их последовательных значений, например, могут быть связаны между собой последовательные отклонения от общей тенденции развития. В этих связях экономических показателей могут присутствовать задержки (временные лаги) и так далее. Это обуславливает необходимость специальных методов их обработки и анализа по сравнению с данными перекрестных выборок.

1.5 Цели и методы сбора статистических данных

Целью сбора экономических данных является получение информационной базы для принятия решений. Естественно, что анализ данных и принятие решений проводится на основе какой-либо интуитивной (неявной) или количественной (явной) экономической модели. Поэтому собирают именно те данные, которые необходимы для соответствующей модели. Существуют различные методы сбора экономических данных: путем опроса, анкетирования и интервьюирования, получения официальной статистической отчетности и так далее. В большинстве стран существуют статистические органы, занимающиеся сбором, обработкой, распространением и публикацией важнейших данных. Этой деятельностью занимаются также многие специализированные государственные и частные агентства.

1.6 Подготовка статистических данных и использование их в модели

При подготовке статистических данных для работы с экономической моделью возникают две проблемы. Во-первых, могут отсутствовать необходимые для модели данные. Во-вторых (если все данные есть), нужно правильно отобрать их для конкретной модели так, чтобы они были согласованы и имели общую методическую базу оценки. При отсутствии нужных данных они нередко могут быть рассчитаны по имеющимся.

Например, если отсутствуют данные о темпе инфляции (INF), но имеются данные о дефляторе валового внутреннего продукта (DEF), то инфляция (по ВВП) может быть рассчитана (в %):

$$INF = \left(\frac{DEF}{DEF_{-1}} - 1 \right) \cdot 100 ,$$

где индекс «1» означает предшествующий год.

Если все нужные данные есть, то для модели необходимо их преобразовать в некоторый взаимно согласованный набор. Если это данные в денежном выражении, то это должны быть всюду одни и те же текущие, либо фиксированные (одного и того же года), денежные единицы. Реальным объемным показателям (то есть в фиксированных ценах) должны соответствовать реальные относительные показатели (например, процентные ставки нужно скорректировать на темп инфляции). В соответствии с решаемой задачей определяют и обобщающие показатели: валовой национальный продукт (ВНП или GNP), валовой внутренний продукт (ВВП, или GDP), валовые внутренние или национальные сбережения, дефлятор ВНП или ВВП и так далее. Например, если речь идет о внутреннем производстве и о влиянии на него внутренних инвестиций, то в качестве обобщающего показателя, на который влияют эти инвестиции, должен выступать ВВП, а не ВНП.

1.7 Различные способы представления экономических данных

Собранные данные могут быть представлены в различной форме: в виде таблиц, диаграмм, графиков. Сформулировав в явном виде экономическую модель, например, предположив, что совокупное потребление линейно растет с ростом совокупного дохода, мы должны собрать данные по тем экономическим показателям, которые входят в исследуемую модель, то есть, данные по совокупному потреблению и совокупному доходу. Это можно сделать, например, взяв годовые данные из национальных счетов какой-либо страны за некоторый промежуток времени. Эти данные могут быть представлены в виде таблицы. В таблице 1 приведены данные по ВВП ("доход") и объему личных потребительских расходов ("потребление") в США, в миллиардах долларов в ценах 1987 г.

Таблица 1

Год	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**
Доход	3860	4114	4243	4362	4497	4674	4801	4840	4784	4907
Потребление	2533	2657	2772	2878	2961	3075	3141	3178	3161	3243

Эти данные могут быть представлены также в виде точек на координатной плоскости (диаграмма рассеивания, рисунок 1):

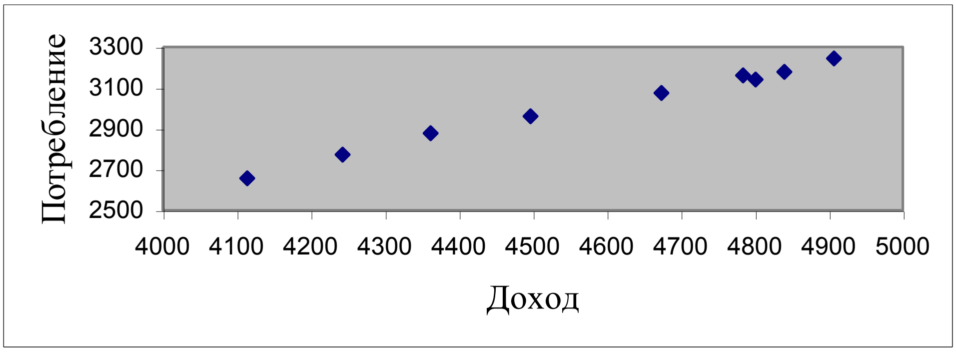

Рисунок 1

Подготовленные данные могут быть подставлены в теоретическую модель, представленную аналитически (в виде некоторой математической модели, например, уравнения $CONS = a + b \cdot GDP + \varepsilon$) или в графическом виде (например, прямой линии на плоскости $CONS, GDP$). Здесь $CONS$ - потребление, GDP - ВВП (доход). При этом возникает ряд проблем, важнейшими из которых являются проверка согласованности теоретической модели с данными, оценка параметров модели и проверка предположений (гипотез), лежащих в основе модели.

1.8 Проверка экономических моделей

Проверить экономическую модель - это значит, в первую очередь, определить, насколько она согласуется с реальными данными об изучаемом объекте. Для этого по эмпирическим данным вычисляются различные статистические характеристики, позволяющие оценить количественно параметры модели, проанализировать надежность этих оценок, проверить различные гипотезы, лежащие в основе исследуемой модели. Располагая экономическими данными, можно не только оценить параметры в уже имеющейся модели, но и выявить эмпирически неизвестные ранее закономерности. А затем уже, на основе выявленных закономерностей, построить соответствующую теоретическую модель.Задачей экономического исследования является уяснение природы экономического объекта, раскрытие механизма взаимосвязи между важнейшими его переменными. Такое понимание позволяет разработать и осуществить необходимые меры по управлению данным объектом или экономическую политику. Для этого нужны адекватные задаче методы, учитывающие природу и специфику экономических данных, служащих основой для качественных и количественных утверждений об изучаемом экономическом объекте или явлении.

2 Элементы математической статистики

Случайной величиной называется функция, заданная на множестве исходов (результатов) данного опыта.

Это значит, что каждому исходу опыта E_K поставлено в соответствие единственное число x_K, которое называется значением случайной величины ξ на исходе опыта E_K. Пишут: $x_K = \xi(E_K)$. При этом некоторые из чисел x_K могут совпадать. Если же все значения случайной величины совпадают ($x_1 = x_2 = ... = x_N = a$), то говорят, что рассматриваемая случайная величина есть постоянная ($\xi = a$). Если множество исходов конечно, случайная величина называется ***дискретной***, если нет, то - ***непрерывной***. Чаще всего дискретная случайная величина задается таблицей (таблица 2).

Таблица 2

Исходы	E_1	E_2	$\dots$	E_K	$\dots$	E_N
ξ	x_1	x_2	$\dots$	x_K	$\dots$	x_N

Со случайными величинами, рассматриваемыми в одном и том же опыте, обращаются как с обыкновенными числовыми функциями. Так, если дискретная случайная величина ξ задана таблицей 2, а дискретная случайная величина η задана таблицей 3, то сумма, разность, произведение и частное этих случайных величин определяются так, как показано в таблице 4.

Таблица 3

Исходы	E_1	E_2	$\dots$	E_K	$\dots$	E_N
η	y_1	y_2	$\dots$	y_K	$\dots$	y_N

Таблица 4

Исходы	E_1	E_2	$\dots$	E_K	$\dots$	E_N
$\xi+\eta$	x_1+y_1	x_2+y_2	$\dots$	x_K+y_K	$\dots$	x_N+y_N
$\xi-\eta$	x_1-y_1	x_2-y_2	$\dots$	x_K-y_K	$\dots$	x_N-y_N
$\xi\cdot\eta$	$x_1\cdot y_1$	$x_2\cdot y_2$	$\dots$	$x_K\cdot y_K$	$\dots$	$x_N\cdot y_N$
$\dfrac{\xi}{\eta}$	$\dfrac{x_1}{y_1}, y_1\neq 0$	$\dfrac{x_2}{y_2}, y_2\neq 0$	$\dots$	$\dfrac{x_K}{y_K}, y_K\neq 0$	$\dots$	$\dfrac{x_N}{y_N}, y_N\neq 0$

Во многих вопросах для изучения случайной величины бывает достаточно ее простейших числовых характеристик – математического ожидания и дисперсии.

Математическим ожиданием случайной величины ξ, задаваемой таблицей 2, в опыте с n равновероятными исходами называется число

$$M(\xi) = \frac{1}{n}\sum_{k=1}^{n} x_k .$$

Математическое ожидание случайной величины есть ее «среднее значение».

Пример 1 Для случайной величины ξ, заданной таблицей 5, в опыте с равновероятными исходами E_k вычислить $M(\xi)$.

Таблица 5

Исходы	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
ξ	-3	10	0	5	-3	-1	-3	5	1	-3

Решение

По определению имеем: $M(\xi) = \frac{1}{10}\left(-3+10+0+5-3-1-3+5+1-3\right) = 0{,}8$.

Для математического ожидания дискретной случайной величины справедливы следующие свойства:

a) если a и b постоянные, то $M(a\xi + b) = aM(\xi) + b$;

b) если a и b постоянные, то $M(a\xi) = aM(\xi),\ M(b) = b$;

c) для случайных величин $\xi_1, \xi_2, ..., \xi_r$ справедливо равенство: $M\left(\sum_{i=1}^{r}\xi_i\right) = \sum_{i=1}^{r} M(\xi_i)$.

Для вычисления математического ожидания случайной величины на самом деле не обязательно знать всю таблицу 4. Об этом говорит следующая теорема о вычислении математического ожидания.

Пусть события $A_1, A_2, ..., A_r$ ***попарно несовместны (то есть произойти может только одно событие из пары),*** $A_1 \cup A_2 \cup ... \cup A_r = E$ ***(множество исходов),*** $\xi(E_i) = a_k$ ***при*** $E_i \subset A_k$***. Тогда*** $M(\xi) = \sum_{k=1}^{r} a_k p_k$ ***, где*** $p_k = P(A_k)$ ***(вероятность того, что в ходе опыта наступит событие*** A_k***).***

Из определения чисел p_k и попарной несовместности событий A_k следует, что $\sum_{k=1}^{r} p_k = 1,\ 0 < p_k \le 1$.

Таким образом, для вычисления математического ожидания случайной величины достаточно знать только числа a_k и p_k. Условия можно записать в виде таблицы 6.

Таблица 6

События	A_1	A_2	$\dots$	A_r
ξ	a_1	a_2	$\dots$	a_r
p	p_1	p_2	$\dots$	p_r

Таблица 6 называется ***законом распределения*** случайной величины ξ.

Фундаментальную роль в теории вероятностей и ее приложениях играет понятие независимых случайных величин.

Две случайные величины ξ и η с законами распределения

ξ	$\dots$	a_i	$\dots$
p	$\dots$	p_i	$\dots$

η	$\dots$	b_j	$\dots$
p	$\dots$	p'_j	$\dots$

называются ***независимыми***, если при любых i и j выполнено равенство:

$$p((\xi = a_i) \cap (\eta = b_j)) = p(\xi = a_i) \cdot p(\eta = b_j).$$

Пример 2 Для случайной величины примера 1 записать закон распределения и выписать события A_k, фигурирующие в теореме.

Решение состоит в том, что в таблице, задающей случайную величину, надо «собрать вместе» столбцы, в которых указаны одинаковые значения этой случайной величины – получим соответствующее событие A_k. Его вероятность рассчитывается как отношение числа исходов опыта, благоприятствующих наступлению этого события, к общему числу исходов опыта. Так, для примера 1 получим:

ξ	-3	-1	0	1	5	10
p	0,4	0,1	0,1	0,1	0,2	0,1

Здесь $A_1 = E_1 \cup E_5 \cup E_7 \cup E_{10}$, $A_2 = E_6$, $A_3 = E_3$, $A_4 = E_9$, $A_5 = E_4 \cup E_8$, $A_6 = E_2$.

Если математическое ожидание случайной величины дает нам «ее среднее значение» или точку на координатной прямой, «вокруг которой разбросаны» значения рассматриваемой случайной величины, то дисперсия характеризует «степень разброса» значений случайной величины около ее среднего значения. На рисунке 2 изображены дискретные случайные величины ξ и η. Они имеют равные математические ожидания, но случайная величина ξ имеет «больший разброс значений», чем случайная величина η.

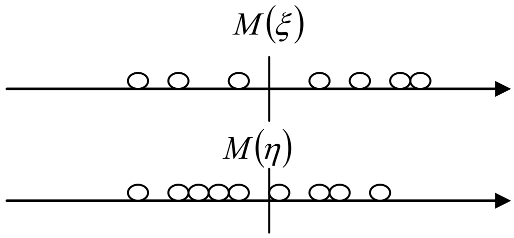

Рисунок 2

Дисперсией случайной величины ξ называется число $D(\xi)=M(\xi-M(\xi))^2=M(\xi^2)-(M(\xi))^2$.

Для дисперсии случайной величины справедливы следующие свойства:

a) если a - постоянная, то $D(a\xi)=a^2D(\xi)$;

b) для попарно независимых случайных величин $D\left(\sum_{i=1}^{r}\xi_i\right)=\sum_{i=1}^{r}D(\xi_i)$.

Число $\sigma=\sqrt{D(\xi)}$ называется ***средним квадратическим отклонением*** или ***стандартным отклонением***.

Ранее мы ввели понятие независимых случайных величин. Наряду с этим вводят ***числовую характеристику, показывающую, насколько данные случайные величины независимы. Эту характеристику называют коэффициентом корреляции.***

Коэффициентом корреляции случайных величин ξ и η называют число $r(\xi;\eta)=\dfrac{M(\xi\eta)-M(\xi)\cdot M(\eta)}{\sqrt{D(\xi)\cdot D(\eta)}}$.

Корреляционным моментом (ковариацией) случайных величин ξ и η называют число $K(\xi;\eta)=Cov(\xi;\eta)=M(\xi\cdot\eta)-M(\xi)\cdot M(\eta)$.

Случайные величины ξ и η называются ***некоррелированными (зависимыми)***, если $Cov(\xi;\eta)=0$.

Пример 3 Подсчитать коэффициент корреляции для случайных величин ξ и η, заданных таблицей 7.

Таблица 7

Исходы	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
ξ	7	-2	1	-5	3	-2	1	-2	0	1
η	5	-1	-7	0	10	4	-1	0	5	3

Решение

Найдем $M(\xi)=\dfrac{1}{10}(7-2+1-5+3-2+1-2+0+1)=0{,}2$,

$$M(\eta)=\frac{1}{10}(5-1-7+0+10+4-1+0+5+3)=1{,}8.$$

Вычислим $M(\xi^2), M(\eta^2), M(\xi \cdot \eta)$. Для этого заполним таблицу 8.

Таблица 8

Исходы	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
ξ^2	49	4	1	25	9	4	1	4	0	1
η^2	25	1	49	0	100	16	1	0	25	9
$\xi \cdot \eta$	35	4	-7	0	30	-8	-1	0	0	3

$M(\xi^2) = \frac{1}{10}(49+4+1+25+9+4+1+4+0+1) = 9{,}7$;

$M(\eta^2) = \frac{1}{10}(25+1+49+0+100+16+1+0+25+9) = 22{,}6$;

$M(\xi \cdot \eta) = \frac{1}{10}(35+2-7+0+30-8-1+0+0+3) = 5{,}4$;

$D(\xi) = M(\xi^2) - (M(\xi))^2 = 9{,}76$, $\quad D(\eta) = M(\eta^2) - (M(\eta))^2 = 19{,}36$;

$Cov(\xi;\eta) = M(\xi \cdot \eta) - M(\xi) \cdot M(\eta) = 5{,}04$, $\quad r(\xi;\eta) = \frac{Cov(\xi;\eta)}{\sqrt{D(\xi) \cdot D(\eta)}} = 0{,}37$.

Для вычислений с дискретными случайными величинами мы пользовались их законом распределения, то есть выписывали вероятности $P(\xi = a_i)$ для ***всех*** возможных значений случайной величины ξ. Для непрерывной же случайной величины выписать ее закон распределения невозможно. Вместо него вводится функция распределения случайной величины. Для любой случайной величины ξ можно рассматривать событие, состоящее в том, что в результате опыта она (случайная величина) приняла значение, меньшее числа x. Можно рассматривать вероятность этого случайного события - $P(\xi < x)$. На множестве всех действительных чисел определена функция $F(x) = P(\xi < x)$. Она называется ***функцией распределения случайной величины*** ξ.

Пример 4 Найти функцию распределения случайной величины с законом распределения:

ξ	-1	0	2	2,5
p	0,2	0,3	0,4	0,1

Решение Значения случайной величины ξ разбивают все действительные числа на пять промежутков. На каждом из них будем определять функцию распределения $F(x)$ случайной величины ξ.

Пусть $x \le -1$; тогда $(\xi < x)$ - ложное событие, так как случайная величина ξ не имеет значений, меньших $x \le -1$. Поэтому $F(x) = P(\xi < x) = 0$.

Пусть $-1 < x \le 0$, тогда $(\xi < x) = (\xi = -1)$, поэтому $F(x) = P(\xi = -1) = 0{,}2$.

Пусть $0 < x \le 2$, тогда $(\xi < x) = (\xi = -1) \cup (\xi = 0)$, так как $(\xi = -1)$ и $(\xi = 0)$ - несовместны, то $F(x) = P(\xi = -1) + P(\xi = 0) = 0{,}2 + 0{,}3 = 0{,}5$.

Пусть $2 < x \le 2{,}5$, тогда $(\xi < x) = (\xi = -1) \cup (\xi = 0) \cup (\xi = 2)$, так как $(\xi = -1)$, $(\xi = 0), (\xi = 2)$ - попарно несовместны,

то $F(x) = P(\xi = -1) + P(\xi = 0) + P(\xi = 2) = 0{,}2 + 0{,}3 + 0{,}4 = 0{,}9$.

При $x > 2{,}5$ $(\xi < x)$ - истинное событие, так как любое значение случайной величины ξ меньше такого x, то $F(x) = 1$.

Тогда:

$$F(x) = \begin{cases} 0; x \le -1, \\ 0{,}2; -1 < x \le 0, \\ 0{,}5; 0 < x \le 2, \\ 0{,}9; 2 < x \le 2{,}5, \\ 1; x \ge 2{,}5. \end{cases}$$

Если функция распределения F случайной величины ξ дифференцируема, то ее ***плотностью вероятности*** называется функция $p(x) = F'(x)$.

Перечислим основные свойства функции распределения $F(x)$ и плотности вероятности $p(x)$ случайной величины ξ:

a) $0 \le F(x) \le 1$ для любого x;
b) функция распределения не убывает и $P(x_1 \le \xi < x_2) = F(x_2) - F(x_1)$;
c) $\lim\limits_{x \to \infty} F(x) = 1$ и $\lim\limits_{x \to -\infty} F(x) = 0$;
d) для любого x плотность вероятности неотрицательна и $P(x_1 < \xi \le x_2) = \int\limits_{x_1}^{x_2} p(x) dx$;
e) $\int\limits_{-\infty}^{+\infty} p(x) dx = 1$;
f) $F(x) = \int\limits_{-\infty}^{x} p(t) dt$.

Математическим ожиданием непрерывной случайной величины ξ, возможные значения которой находятся на отрезке $[a,b]$ (a и b могут быть равными $\pm\infty$), и имеющей плотность вероятности $p(x)$, называется число $M(\xi) = \int\limits_a^b x \cdot p(x) dx$,

а ее ***дисперсией*** называется число

$$D(\xi) = M\left((\xi - M(\xi))^2\right) = \int\limits_a^b \left(x - M(\xi)^2\right) \cdot p(x) dx = \int\limits_a^b x^2 p(x) dx - (M(\xi))^2.$$

Пример 5 Найти математическое ожидание, дисперсию, среднее квадратическое отклонение случайной величины ξ, заданной плотностью вероятности $p(x)=1$ на отрезке $[0,1]$.

Решение Согласно определениям, получаем:

$$M(\xi)=\int_0^1 x\cdot p(x)dx=\int_0^1 xdx=\frac{1}{2}x^2\Big|_0^1=\frac{1}{2},$$

$$D(\xi)=\int_0^1 x^2p(x)dx-(M(\xi))^2=\int_0^1 x^2dx-\left(\frac{1}{2}\right)^2=\frac{1}{3}-\frac{1}{4}=\frac{1}{12},\quad \sigma(\xi)=\sqrt{D(\xi)}\approx 0{,}289 .$$

3 Модель линейной регрессии

3.1 Проблема оценивания линейной связи экономических переменных

Проблема изучения взаимосвязей экономических показателей является одной из важнейших проблем экономического анализа. Любая экономическая политика заключается в регулировании экономических переменных, и она должна основываться на знании того, как эти переменные влияют на другие переменные, являющиеся ключевыми для принимающего решения политика. Так, в рыночной экономике нельзя непосредственно регулировать темп инфляции, но на него можно воздействовать средствами бюджетно-налоговой и кредитно-денежной политики. Поэтому, в частности, должна быть изучена зависимость между предложением денег и уровнем цен. Невозможно строить, проверять или улучшать экономические модели без статистического анализа их переменных с использованием реальных статистических данных. Вся сфера экономических исследований может быть в определенном смысле охарактеризована как изучение взаимосвязей экономических переменных, и инструментарием их базового анализа являются методы статистики и эконометрики.Изучение зависимостей экономических переменных начнем со случая двух переменных (обозначим их x и y). Этот случай наиболее прост и может быть рассмотрен графически.

Предположим, что имеются ряды значений переменных, соответствующие им точки нанесены на график и соединены линией. Если это реальные статистические данные, то мы никогда не получим простую линию - линейную, квадратичную, экспоненциальную и т.д. Всегда будут присутствовать отклонения зависимой переменной, вызванные ошибками измерения, влиянием неучтенных величин или случайных факторов. Но если мы не получили, например, точную прямую линию, это еще не значит, что в основе рассматриваемой зависимости лежит нелинейная функция. Возможно, зависимость переменных линейна, и лишь случайные факторы приводят к некоторым отклонениям от нее. То же самое можно сказать и про любой другой вид функции. Связь переменных, на которую накладываются воздействия случайных факторов, называется статистической связью. Наличие такой связи заключается в том, что изменение одной переменной приводят к изменению математического ожидания другой переменной. Можно указать два типа взаимосвязей между переменными x и y. В одном случае может быть неизвестно, какая из двух переменных является независимой, и какая - зависимой. В этом случае переменные равноправны, и имеет смысл говорить о статистической взаимосвязи корреляционного типа. Другая ситуация возникает, если две исследуемые переменные не равноправны, но одна из них рассматривается как объясняющая (или независимая), а другая - как объясняемая (или зависящая от первой). Если это так, то изменение одной из переменных служит причиной для изменения другой. Например, рост дохода ведет к увеличению потребления; снижение процентной ставки увеличивает инвестиции; увеличение валютного курса сокращает чистый экспорт. Это тот случай, когда должно быть оценено уравнение регрессии $y = f(x)$. ***Уравнение регрессии - это формула статистической связи между переменными***. Если эта формула линейна, то речь идет о линейной регрессии. Формула статистической связи двух переменных называется ***парной регрессией***, зависимость от нескольких переменных - ***множественной регрессией***. Например, Кейнсом была предложена линейная формула зависимости частного потребления C от располагаемого дохода Y: $C = C_0 + b \cdot Y$, где $C_0 > 0$ - величина автономного потребления, $1 > b > 0$ - предельная склонность к потреблению.

Выбор формулы связи переменных называется ***спецификацией уравнения регрессии***; в данном случае выбрана линейная формула. Однако до тех пор, пока не оценены количественные значения параметров C_0 и b, не проверена надежность сделанных оценок, эта формула остается лишь гипотезой. Оценка значений параметров выбранной формулы статистической связи переменных называется ***параметризацией уравнения регрессии***. Как же оценить значения параметров и проверить надежность оценок? Рассмотрим вначале рисунок 3.

На рисунке 3 изображены три ситуации:

- на графике (а) взаимосвязь x и y близка к линейной; прямая линия (1) здесь близка к точкам наблюдений, которые отклоняются от нее лишь в результате сравнительно небольших случайных воздействий;

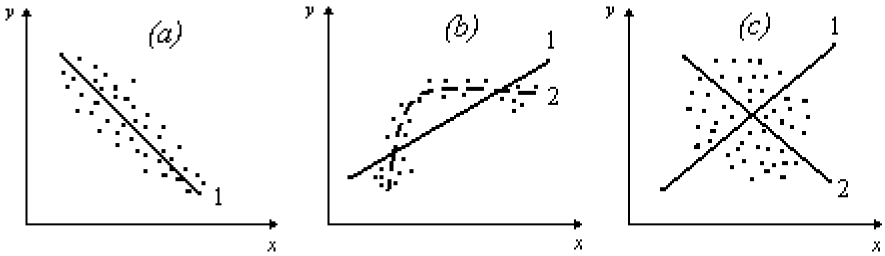

Рисунок 3

- на графике (b) реальная взаимосвязь величин x и y дописывается нелинейной функцией (2). И какую бы мы ни провели прямую линию (например, (1)), отклонения точек наблюдений от нее будут существенными и неслучайными;
- на графике (c) явная взаимосвязь между переменными x и y отсутствует.

И какую бы мы ни выбрали формулу связи, результаты ее параметризации будут здесь не удачными. В частности, прямые линии (1) и (2), проведенные через «центр облака» точек наблюдений и имеющие противоположный наклон, одинаково плохи для того, чтобы делать выводы об ожидаемых значениях переменной y по значениям переменной x.

3.2 Парная линейная регрессия. Метод наименьших квадратов

Начальным пунктом эконометрического анализа зависимостей обычно является оценка линейной зависимости переменных. Если имеется некоторое "облако" точек наблюдений, через него всегда можно попытаться провести такую прямую линию, которая является наилучшей в определенном смысле среди всех прямых линий, то есть "ближайшей" к точкам наблюдений по их совокупности. Для этого мы вначале должны определить понятие близости прямой к некоторому множеству точек на плоскости; меры такой близости могут быть различными. Однако любая разумная мера должна быть, очевидно, связана с расстояниями от точек наблюдений до рассматриваемой прямой линии, задаваемой уравнением $y = a + bx$.

Обычно в качестве критерия близости используется минимум суммы квадратов разностей наблюдений зависимой переменной y_i и теоретических, рассчитанных по уравнению регрессии значений ($a + bx_i$):

$$Q = \sum_i e_i^2 = \sum_i \left(y_i - (a + bx_i)\right)^2 \to \min .$$

Здесь считается, что y_i и x_i - известные данные наблюдений, a и b - неизвестные параметры линии регрессии. Поскольку функция Q непрерывна, выпукла и ограничена снизу нулем, она имеет минимум. Для соответствующих точке этого минимума значений a и b могут быть найдены простые и удобные формулы (они будут при ведены ниже). Метод оценивания параметров линейной регрессии, минимизирующий сумму квадратов отклонений наблюдений зависимой переменной от искомой линейной функции, называется ***методом наименьших квадратов (МНК)*** **или *Least Squares Method (LS).***

«Наилучшая» по МНК прямая линия всегда существует, но даже наилучшая не всегда является достаточно хорошей. Если в действительности зависимость $y = f(x)$ является, например, квадратичной (как на рисунке 3 (b)), то ее не сможет адекватно описать ни какая линейная функция, хотя среди всех таких функций обязательно найдется «наилучшая». Если величины x и y вообще не связаны (рисунок 3 (c)), мы также всегда сможем найти "наилучшую" линейную функцию $y = a + bx$ для данной совокупности наблюдений, но в этом случае конкретные значения a и b определяются только случайными отклонениями переменных и сами будут очень сильно меняться для различных выборок из одной и той же генеральной совокупности. Возможно, на рисунке 3 (c) прямая (1) является наилучшей среди всех прямых линий (в смысле минимального значения функции Q), но любая другая прямая, проходящая через центральную точку "облака" (например, линия (2)), ненамного в этом смысле хуже, чем прямая (1), и может стать наилучшей в результате небольшого изменения выборки.

Рассмотрим теперь задачу оценки коэффициентов парной линейной регрессии более формально. Предположим, что связь между x и y линейна: $y = \alpha + \beta x$. Здесь имеется в виду связь между всеми возможными значениями величин x и y, то есть для генеральной совокупности. Наличие случайных отклонений, вызванных воздействием на переменную y множества других, неучтенных в нашем уравнении факторов и ошибок измерения, приведет к тому, что связь наблюдаемых величин x и y приобретет вид $y_i = \alpha + \beta x_i + \varepsilon_i$. Здесь ε_i - случайные ошибки (отклонения, возмущения). Задача состоит в следующем: по имеющимся данным наблюдений $\{x_i\},\{y_i\}$ оценить значения параметров α и β, обеспечивающие минимум величины Q . Если бы были известны точные значения отклонений ε_i, то можно было бы (в случае правильности предполагаемой линейной формулы) рассчитать значения параметров α и β. Однако значения случайных отклонений в выборке неизвестны, и по наблюдениям x и y можно получить оценки параметров α и β, которые сами являются случайными величинами, поскольку соответствуют случайной выборке. Пусть a - оценка параметра α, b - оценка параметра β. Тогда оцененное уравнение регрессии будет иметь вид: $y_i = a + bx_i + e_i$, где e_i - наблюдаемые значения ошибок ε_i.

Для оценки параметров α и β воспользуемся МНК, который минимизирует сумму квадратов отклонений фактических значений y_i от расчетных. Минимум ищется по переменным a и b.

При использовании МНК к ошибкам ε_i предъявляются следующие требования, называемые условиями Гаусса-Маркова:

1) величина ε_i является случайной переменной;

2) математическое ожидание ε_i равно нулю - $M(\varepsilon_i) = 0$;

3) дисперсия ε_i постоянна - $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ для всех i, j;

4) значения ε_i, независимы между собой; откуда вытекает, в частности, что

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, i \neq j, \\ \sigma^2, i = j; \end{cases}$$

5) величины ε_i статистически независимы со значениями x_i.

Известно, что если условия 1) - 5) выполняются, то оценки, сделанные с помощью МНК, обладают следующими свойствами:

1) оценки являются несмещенными, то есть математическое ожидание оценки каждого параметра равно его истинному значению - $M(a) = \alpha, M(b) = \beta$. Это вытекает из того, что $M(\varepsilon_i) = 0$, и говорит об отсутствии систематической ошибки в определении положения линии регрессии;

2) оценки состоятельны, так как дисперсия оценок параметров при возрастании числа наблюдений стремится к нулю: $\lim\limits_{n\to\infty} D(a) = 0, \lim\limits_{n\to\infty} D(b) = 0$. Иначе говоря, если *n* (количество наблюдений) достаточно велико, то практически наверняка a близко к α, а b близко к β. Надежность оценки при увеличении выборки растет;

3) оценки эффективны, они имеют наименьшую дисперсию по сравнению с любыми другими оценками данного параметра, линейными относительно величин y_i. В англоязычной литературе такие оценки называются BLUE (***Best Linear Unbiased Estimators* - наилучшие линейные несмещенные оценки).**

Перечисленные свойства не зависят от конкретного вида распределения величин ε_i, тем не менее, обычно предполагается, что они распределены нормально $N(0, \sigma^2)$ ([1]). Эта предпосылка необходима для проверки статистической значимости сделанных оценок и определения для них доверительных интервалов. При ее выполнении оценки МНК имеют наименьшую дисперсию не только среди линейных, но среди всех несмещенных оценок.

Если предположения 3) и 4) нарушены, то есть дисперсия возмущений непостоянна и/или значения ε_i связаны друг с другом, то свойства несмещенности и состоятельности сохраняются, но свойство эффективности - нет.

При невыполнении предположения 5) может нарушаться и свойство несмещенности оценок, являющееся наиболее важным в эконометрическом анализе. Значительная часть современной эконометрической теории посвящена анализу выполнения данного свойства (в совокупности с остальными) в различных конкретных ситуациях, а также выяснению и корректировке последствий его не выполнения.

Рассмотрим теперь процедуру оценивания параметров парной линейной регрессии a и b. Для того, чтобы функция $Q = \sum_i e_i^2 = \sum_i (y_i - (a + bx_i))^2$ достигала минимума, необходимо равенство нулю ее частных производных:

$$\begin{cases} Q'_a = -2\sum_i (y_i - a - bx_i) = 0, \\ Q'_b = -2\sum_i (y_i - a - bx_i)x_i = 0, \end{cases} \Rightarrow \begin{cases} \sum_i y_i - na - b\sum_i x_i = 0, & (1) \\ \sum_i y_i x_i - a\sum_i x_i - b\sum_i x_i^2 = 0. & (2) \end{cases}$$

Если уравнение (1) разделить на n, то получим $\bar{y} = a + b\bar{x}$ (здесь $\bar{y} = \frac{\sum_i y_i}{n}; \bar{x} = \frac{\sum_i x_i}{n}$ - средние значения x и y). Таким образом, линия регрессии проходит через точку со средними значениями x и y. Подставив величину a из (1) во (2), получаем:

$$\sum_i y_i x_i = \sum_i x_i(\bar{y} - b\bar{x}) + b\sum_i x_i^2 = n\bar{x}(\bar{y} - b\bar{x}) + b\sum_i x_i^2 .$$

Откуда:

$$b = \frac{\sum_i y_i x_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}; \qquad a = \bar{y} - b\bar{x} .$$

Иначе можно записать, что $b = \frac{\operatorname{cov}(x,y)}{D(x)} = r\frac{\sqrt{D(y)}}{\sqrt{D(x)}}$, где r - коэффициент корреляции x и y. Таким образом, коэффициент регрессии пропорционален показателю ковариации и коэффициенту корреляции x и y, а коэффициенты этой пропорциональности служат для соизмерения перечисленных разноразмерных величин. Оценки a и b, очевидно, являются линейными относительно y_i (если x_i считать коэффициентами) - выше об этом упоминалось.

Итак, если коэффициент r уже рассчитан, то легко рассчитать коэффициент парной регрессии, не решая системы уравнений. Ясно также, что если рассчитаны линейные регрессии $x(y)$ и $y(x)$, то произведение коэффициентов b_x и b_y равно r^2 - $b_x b_y = r\frac{\sqrt{D(y)}}{\sqrt{D(x)}} \cdot r\frac{\sqrt{D(x)}}{\sqrt{D(y)}} = r^2$.

3.3 Анализ статистической значимости коэффициентов линейной регрессии

Величины y_i, соответствующие данным x_i при некоторых теоретических значениях α и β, являются случайными. Следовательно, случайными являются и рассчитанные по ним значения коэффициентов a и b.

Их математические ожидания при выполнении предпосылок об отклонениях e_i равны, соответственно, α и β. При этом оценки тем надежнее, чем меньше их разброс вокруг α и β, то есть дисперсия. По определению дисперсии $D(b) = M(b-\beta)^2$; $D(a) = M(a-\alpha)^2$. Надежность получаемых оценок a и b зависит, очевидно, от дисперсии случайных отклонений ε_i, но поскольку по данным выборки эти отклонения (и, соответственно, их дисперсия) оценены быть не могут, они заменяются при анализе надежности оценок коэффициентов регрессии на отклонения переменной y от оцененной линии регрессии $e_i = y_i - a - bx_i$. Можно доказать (доказательство опускаем), что

$$D(b) = S_b^2 = \frac{S^2}{\sum_i (x_i - \bar{x})^2}; \quad D(a) = S_a^2 = \frac{S^2 \sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}, \text{ где } S^2 = \frac{\sum_i e_i^2}{n-2} \text{ - мера разброса}$$

зависимой переменной вокруг линии регрессии (***необъясненная дисперсия***). S_a и S_b - стандартные отклонения случайных величин a и b. Полученный результат можно интерпретировать следующим образом.

Коэффициент b есть мера наклона линии регрессии. Очевидно, чем больше разброс значений y вокруг линии регрессии, тем больше (в среднем) ошибка в определении наклона линии регрессии. Если такого разброса нет совсем ($\varepsilon_i = 0$ и, следовательно, $\sigma^2 = 0$), то прямая определяется однозначно и ошибки в расчете коэффициентов a и b отсутствуют (а отсюда и значение S^2, «замещающее» σ^2, равное нулю).

На рисунке 4 (*a*) отклонения в значениях переменной y от линии регрессии отсутствуют, и через три точки проводится та же прямая, что и через любые две из них. На рисунке 4 (*b*) через три точки проводится такая же линия регрессии, но колебания значений переменной y вокруг этой линии значительны. Поэтому через пары точек (1; 2) и (1; 3) проходят совершенно разные прямые, отличные от общей прямой. Следовательно, стандартные ошибки коэффициентов регрессии в этом случае будут значительными.

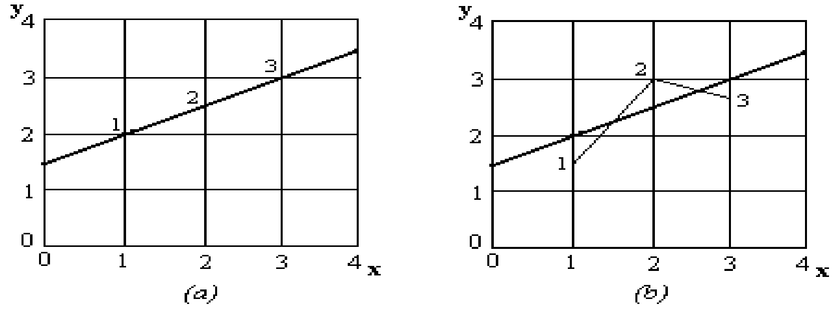

Рисунок 4

В знаменателе величины $D(b)$ стоит сумма квадратов отклонений x от среднего значения $\bar{x}$. Эта сумма велика в том случае, если регрессия оценена на достаточно широком диапазоне значений переменной *x,* и в этом случае, при данном уровне разброса y, очевидно, ошибка в оценке величины наклона прямой будет меньше, чем при малом диапазоне изменения переменной x. Попробуйте провести прямую по двум точкам: если x_i и x_j лежат рядом, то даже небольшое изменение одного из y_i существенно меняет наклон прямой (если x_i и x_j далеки друг от друга - ситуация обратная).

Формально значимость оцененного коэффициента регрессии b может быть проверена с помощью анализа его отношения к своему стандартному отклонению $S_b = \sqrt{D(b)}$. Эта величина в случае выполнения исходных предпосылок модели имеет t-распределение Стьюдента с $(n-2)$ степенями свободы (n - число наблюдений). Она (величина) называется ***t* -статистикой**:

$$t = \frac{b}{\sqrt{D(b)}} = \frac{b}{S_b}.$$

При оценке значимости коэффициента линейной регрессии можно использовать следующее грубое правило. Если стандартная ошибка коэффициента больше его модуля $(t < 1)$, то он не может быть признан хорошим (значимым). Если стандартная ошибка меньше модуля коэффициента, но больше его половины ($1 < t < 2$), то сделанная оценка может рассматриваться как более или менее значимая. Значение t от 2 до 3 свидетельствует о весьма значимой связи, и $t > 3$ есть практически стопроцентное свидетельство ее наличия. В каждом случае играет роль число наблюдений: чем их больше, тем надежнее при прочих равных выводы о наличии связи.

3.4 Множественная линейная регрессия

Значения экономических переменных определяются обычно влиянием не одного, а нескольких объясняющих факторов. В таком случае зависимость $y = f(x)$ означает, что x - вектор, содержащий m компонентов: $x = (x_1, x_2, ..., x_m)$. Задача оценки статистической взаимосвязи переменных y и $x = (x_1, x_2, ..., x_m)$ формулируется аналогично случаю парной регрессии. Записывается функция $y = f(x, \alpha) + \varepsilon$, где α - вектор параметров, ε - случайная ошибка. Предполагается, что эта функция связывает переменную y с вектором независимых переменных x для данных генеральной совокупности. Как и в случае парной регрессии, предполагается, что ошибки ε_i являются случайными величинами с нулевым математическим ожиданием и постоянной дисперсией; ε_i и ε_j статистически независимы при различных i, j. Кроме того, для проверки статистической значимости оценок α обычно предполагается, что ошибки ε_i нормально распределены. По данным наблюдений выборки размерности n

требуется оценить значения параметров α, то есть провести параметризацию выбранной формулы (спецификации) зависимости.

Мы будем говорить о линейной зависимости y от x, то есть о множественной линейной регрессии. Теоретическое уравнение регрессии имеет вид: $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_m x_m + \varepsilon$.

Здесь α - вектор неизвестных параметров размерности $(m+1)$. Пусть имеется n наблюдений вектора x и зависимой переменной y. Для того, чтобы формально можно было решить задачу, то есть найти некоторый наилучший вектор параметров, должно быть $n \geq (m+1)$. Если это условие не выполняется, то можно найти бесконечно много разных векторов коэффициентов, при которых линейная формула связывает между собой x и y для имеющихся наблюдений абсолютно точно. Если в частном случае $n = m+1$ (например, при двух объясняющих переменных в уравнении $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ и трех наблюдениях), то оценки коэффициентов α рассчитываются единственным образом – путем решения системы линейных уравнений $\{y_j = \alpha_0 + \alpha_1 x_{j1} + \alpha_2 x_{j2} + ... + \alpha_m x_{jm}\}, j = 1,2,...,n;$ n - индекс наблюдения. Так, через три точки наблюдения в трехмерном пространстве можно провести единственную плоскость, определяемую параметрами $\alpha_0, \alpha_1, \alpha_2$. Если число наблюдений больше минимально необходимого, то есть $n > m+1$, то уже нельзя подобрать линейную формулу, в точности удовлетворяющую всем наблюдениям, и возникает необходимость оптимизации, то есть выбора наилучшей формулы приближения для имеющихся наблюдений. Положительная разность $(n-m-1)$ в этом случае называется числом степеней свободы. Если число степеней свободы мало, то статистическая надежность оцениваемой формулы невысока. Так, если проведена плоскость «в точности» через имеющиеся три точки наблюдений, любая четвертая точка-наблюдение из той же генеральной совокупности будет практически наверняка лежать вне этой плоскости, возможно, достаточно далеко от нее. Обычно при оценке множественной регрессии для обеспечения статистической надежности требуется, чтобы число наблюдений, по крайней мере в 3 раза, превосходило число оцениваемых параметров.

Задача построения множественной линейной регрессии состоит в нахождении $(m+1)$-мерного вектора a, элементы которого есть оценки соответствующих элементов вектора α. Критерии оценивания, как и в случае парной регрессии, могут быть различными; мы будем вновь использовать метод наименьших квадратов (МНК).

Уравнение регрессии с оцененными параметрами имеет вид:

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + e.$$

Критерием для нахождения вектора a является - $\min\sum_i e_i^2$. Оцененное уравнение должно описать как общий тренд (тенденцию) изменения зависимой переменной y, так и отклонения от этого тренда. Проблема здесь состоит не только в том, чтобы объяснить возможно большую долю колебаний переменной y, но и отделить влияние каждого из факторов, рассматриваемых как объясняющие переменные.

При выполнении предпосылок 1) - 4) (смотри предыдущий пункт) относительно ошибок e_i оценки параметров множественной линейной регрессии являются несмещенными, состоятельными и эффективными. Отклонение зависимой переменной y в j-м наблюдении от линии регрессии e_j записывается следующим образом: $e_j = y_j - a_0 - a_1x_{j1} - a_2x_{j2} - ... - a_mx_{jm}$. Обозначим сумму квадратов этих величин, которую нужно минимизировать в соответствии с методом наименьших квадратов, через Q:

$$Q = \sum_j e_j^2 = \sum_j \left(y_j - \left(a_0 + \sum_i a_i x_{ji} \right) \right)^2 \to \min .$$

Минимизируемая функция Q является квадратичной относительно неизвестных величин a_i. Необходимым условием ее минимума является равенство нулю всех ее частных производных по a_i. Частные производные квадратичной функции являются линейными функциями и, приравнивая их все к нулю, мы получим систему из $(m+1)$ линейных уравнений с $(m+1)$ неизвестными. Такая система имеет обычно единственное решение. За исключением особого случая, когда столбцы ее линейно зависимы и решения нет или их бесконечно много. Однако данные реальных статистических наблюдений к такому особому случаю никогда не приводят. Данная система называется системой нормальных уравнений. Ее решение в явном виде удобнее всего выписать в векторно-матричной форме, иначе оно становится слишком громоздким. Векторно-матричная запись и вывод решения системы нормальных уравнений приведены в Приложении. При начальном ознакомлении с проблемой оно может быть опущено.

Для анализа статистической значимости полученных коэффициентов множественной линейной регрессии необходимо, как и в случае парной регрессии, оценить дисперсию и стандартные отклонения коэффициентов a_i. В случае парной регрессии - $D(b) = S_b^2 = \dfrac{\sum_i e_i^2}{(n-2)\sum_i (x_i - \bar{x})^2}$,

в общем случае - $D(a_j) = \dfrac{\sum_i e_i^2}{(n-m-1)} \cdot Z_{jj}$.

Где Z_{jj} - диагональный элемент матрицы $\left(X^T X\right)^{-1}$. Соответственно, стандартное отклонение $S_{aj} = \sqrt{D\left(a_j\right)}$, и для проверки нулевой гипотезы (проверки на равенство нулю) для каждого из коэффициентов a_j рассчитываются, как и в случае парной регрессии, t-статистики: $t = \dfrac{a_j}{S_{aj}}$, имеющие распределение Стьюдента с $(n - m - 1)$ степенями свободы.

3.5 Проверка общего качества уравнения регрессии

Для анализа общего качества оцененной регрессии используют обычно коэффициент детерминации R^2 , называемый также, квадратом коэффициента множественной корреляции. Для случая парной регрессии это квадрат коэффициента корреляции переменных x и y. Коэффициент детерминации рассчитывается по формуле $R^2 = 1 - \dfrac{\sum_i e_i^2}{\sum_i \left(y_i - \bar{y}\right)^2}$. Он характеризует долю вариации (разброса) зависимой переменной, объясненной с помощью данного уравнения. Иногда при расчете коэффициента детерминации для получения несмещенных оценок дисперсии в числителе и знаменателе вычитаемой из единицы дроби делается поправка на число степеней свободы. Тогда:

$$R^2 = 1 - \left(\frac{\sum_i e_i^2}{n - m - 1}\right) : \left(\frac{\sum_i \left(y_i - \bar{y}\right)^2}{n - 1}\right)$$

или для парной регрессии, где число независимых переменных m равно 1:

$$R^2 = 1 - \left(\frac{\sum_i e_i^2}{n - 2}\right) : \left(\frac{\sum_i \left(y_i - \bar{y}\right)^2}{n - 1}\right).$$

Если существует статистически значимая линейная связь величин x и y, то коэффициент R^2 близок к единице. Однако он может быть близким к единице просто в силу того, что эти обе величины имеют выраженный временной тренд, не связанный с их причинно-следственной взаимозависимостью. В экономике обычно объемные показатели (доход, потребление, инвестиции) имеют такой тренд, а темповые и относительные (производительность, темпы роста, доли, отношения) – не всегда. Поэтому при оценивании линейных регрессий по временным рядам объемных показателей величина R^2 обычно очень близка к единице. Это говорит о том, что зависимую переменную нельзя описать просто как равную своему среднему значению.

Если имеются не временные ряды, а перекрестная выборка, то есть данные об однотипных объектах в один и тот же момент времени, то для оцененного по ним уравнения линейной регрессии величина R^2 не превышает обычно уровня 0,6–0,7. То же самое имеет место и для регрессии по временным рядам, если они не имеют выраженного тренда. В макроэкономике примерами таких зависимостей являются связи относительных, удельных, темповых показателей: зависимость темпа инфляции от уровня безработицы, нормы накопления от процентной ставки, темпа прироста выпуска от темпов прироста затрат ресурсов.

3.6 Пример решения задачи с применением уравнения регрессии

Имеются данные по годовой производительности труда (в расчете на 1 рабочего) и по энерговооруженности на 14 предприятиях одной отрасли.

С помощью уравнения регрессии спрогнозировать производительность труда на 15-м предприятии при его энерговооруженности 4,75 кВт на 1 рабочего. Данные приведены в таблице 9.

Таблица 9

№ предприятия	Производительность труда, тысяч рублей на 1 рабочего	Энерговооруженность, кВт на 1 рабочего
1	6,7	2,8
2	6,9	2,8
3	3,0	0,5
4	7,3	2,9
5	10,9	3,4
6	8,8	0,1
7	9,1	4,0
8	13,8	4,8
9	10,6	4,9
10	10,7	5,2
11	11,1	7,9
12	11,8	5,5
13	12,1	6,2
14	12,4	7,0

Решение Предположим, что между величинами x (энерговооруженность) и y (производительность труда) существует линейная зависимость вида $y = a + bx$. Рассчитаем необходимые для уравнения регрессии величины. Количество наблюдений (n) равно 14.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{135,2}{14} = 9,6571; \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{58}{14} = 4,1429;$$

$$b = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{63,2057}{65,8143} = 0,9604\,; \quad a = \overline{y} - b\overline{x} = 9,6571 - 0,9604 \cdot 4,1429 = 5,6783\,.$$

Уравнение регрессии с оцененными параметрами будет иметь вид:

$$y = 5,6783 - 0,9604x\,.$$

Подставляя значение энерговооруженности 15-го предприятия в найденное уравнение регрессии, можем спрогнозировать производительность труда на этом предприятии:

$$\hat{y} = 5,6783 - 0,9604 \cdot 4,75 = 1,12 \text{ (тысяч рублей на 1 рабочего).}$$

4 Этапы построения многофакторной корреляционно-регрессионной модели

Разработка модели и исследование экономических процессов должны выполняться по следующим этапам:

- априорное исследование экономической проблемы;
- формирование перечня факторов и их логический анализ;
- сбор исходных данных и их первичная обработка;
- спецификация функции регрессии;
- оценка функции регрессии;
- отбор главных факторов;
- проверка адекватности модели;
- экономическая интерпретация;
- прогнозирование неизвестных значений зависимой переменной.

Рассмотрим подробнее содержание этапов.

1 Априорное исследование экономической проблемы

В соответствии с целью работы на основе знаний макро-и микроэкономики конкретизируются явления, процессы, зависимость между которыми подлежит оценке. При этом подразумевается прежде всего четкое определение экономических явлений, установление объектов и периода исследования.

На этом этапе исследования должны быть сформулированы экономически осмысленные и приемлемые гипотезы о зависимости экономических явлений.

2 Формирование перечня факторов и их логический анализ

Для определения наиболее разумного числа переменных в регрессионной модели прежде всего ориентируются на соображения профессиональнотеоретического характера. Исходя из физического смысла явления, производят классификацию переменных на зависимую и объясняющую.

3 Сбор исходных данных и их первичная обработка

При построении модели исходная информация может быть собрана в трех видах:

- динамические (временные) ряды;
- пространственная информация – информация о работе нескольких объектов в одном разрезе времени;
- сменная — табличная форма. Информация о работе нескольких объектов за разные периоды.

4 Спецификация функции регрессии

На данном этапе исследования дается конкретная формулировка гипотезы о форме связи (линейная или нелинейная, простая или множественная и так далее). Для этого используются различные критерии для проверки состоятельности гипотетического вида зависимости. На этом этапе проверяются предпосылки корреляционно-регрессионного анализа.

5 Оценка функции регрессии

Определяются числовые значения параметров регрессии и вычисление ряда показателей, характеризующих точность регрессионнго анализа.

6 Отбор главных факторов Выбор факторов — основа для построения многофакторной корреляционно-регрессионной модели.

На этапе «Формирование перечня факторов и их логический анализ» собираются все возможные факторы, обычно более 20 - 30 факторов. Но это неудобно для анализа и модель, включающая 20 - 30 факторов, будет неустойчива. Неустойчивость модели находит выражение в том, что в ней изменение некоторых факторов ведет к увеличению y вместо снижения y.

Мало факторов – тоже плохо. Это может привести к ошибкам при принятии решений в ходе анализа модели. Поэтому необходимо выбирать более рациональный перечень факторов. При этом проводят анализ факторов на ***мультиколлинеарность***.

7 Анализ и способы снижения влияния мультиколлинеарности на значимость модели

Мультиколлинеарность — попарная корреляционная зависимость между факторами. Мультиколлинеарная зависимость присутствует, если коэффициент парнои корреляции $r(x_i;x_j) \in (0{,}7;0{,}8)$.

Отрицательное воздействие мультиколлинеарности состоит в следующем:

a) усложняется процедура выбора главных факторов;

b) искажается смысл коэффициента множественной корреляции (он предполагает независимость факторов);

c) усложняются вычисления при построении самой модели;

d) снижается точность оценки параметров регрессии, искажается оценка дисперсии.

Следствием снижения точности является ненадежность коэффициентов регрессии и, отчасти, неприемлемость их использования для интерпретации как меры воздействия соответствующей объясняющей переменной на зависимую переменную.

Оценки коэффициента становятся очень чувствительными к выборочным наблюдениям. Небольшое увеличение объема выборки может привести к очень сильным сдвигам в значениях оценок. Кроме того, стандартные ошибки оценок входят в формулы критерия значимости, поэтому применение самих критериев становится также ненадежным.

В качестве меры мультиколлинеарности используется следующая разность: $M = R^2 - \sum_{j=1}^{m} r^2(x_j; y)$,

где:

j - число факторов модели;

$r^2(x_j; y)$ - коэффициент парной корреляции между j-м фактором и зависимой переменной y.

Чем меньше эта разность, тем меньше мультиколлинеарность. Для устранения мультиколлинеарности используется ***метод исключения переменных.*** Этот метод заключается в том, что высоко коррелированные объясняющие переменные (факторы) устраняются из регрессии и она заново оценивается. Отбор переменных, подлежащих исключению, производится с помощью коэффициентов парной корреляции. Опыт показывает, что если $\left| r(x_i; x_j) \right| \geq 0{,}7$, то одну из переменных можно исключить, но какую переменную исключить из анализа решают, исходя из управляемости факторов на уровне предприятия.

Обычно в модели оставляют тот фактор, на который можно разработать мероприятие, обеспечивающее улучшение значения этого фактора в планируемом году. Возможна ситуация, когда оба мультиколлинеарных фактора управляемы на уровне предприятия. Решить вопрос об исключении того или иного фактора можно только в соответствии с процедурой отбора главных факторов. Отбор факторов не самостоятельный процесс, он сопровождается построением модели. Принятие решения об исключении факторов производится на основе анализа значений специальных статистических характеристик и с учетом управляемости факторов на уровне предприятия.

Процедура отбора главных факторов обязательно включает следующие этапы.

8 Анализ факторов на мультиколлинеарность и ее исключение

Здесь производится анализ значений коэффициентов парной корреляции между факторами x_i и x_j.

9 Анализ тесноты взаимосвязи факторов x с зависимой переменной y

Для анализа тесноты взаимосвязи x и y используются значения коэффициента парной корреляции между фактором и функцией $r(x_i; y)$. Величина $r(x_i; y)$ определяется на ЭВМ и представлена в ***корреляционной матрице*** вида (таблица 10):

Таблица 10

№ переменной	x_1	x_2	x_3	$\dots$	x_m	y
x_1	1	$r(x_1; x_2)$	$r(x_1; x_3)$	$\dots$	$r(x_1; x_m)$	$r(x_1; y)$
x_2	$r(x_2; x_1)$	1	$r(x_2; x_3)$	$\dots$	$r(x_2; x_m)$	$r(x_2; y)$
x_3	$r(x_3; x_1)$	$r(x_3; x_2)$	1	$\dots$	$r(x_3; x_m)$	$r(x_3; y)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_m	$r(x_m; x_1)$	$r(x_m; x_2)$	$r(x_m; x_3)$	$\dots$	1	$r(x_m; y)$
y	$r(y; x_1)$	$r(y; x_2)$	$r(y; x_3)$	$\dots$	$r(y; x_m)$	1

Факторы, для которых $r(x_i; y) = 0$, то есть не связанные с y, подлежат исключению в первую очередь. Факторы, имеющие наименьшее значение $r(x_i; y)$, могут быть потенциально исключены из модели. Вопрос об их окончательном исключении решается в ходе анализа других статистических характеристик.

10 Анализ коэффициентов β факторов, которые потенциально могут быть исключены

Коэффициент β учитывает влияние анализируемых факторов на y с учетом различий в уровне их колеблемости. Коэффициент β показывает, насколько сигм (средних квадратических отклонений) изменяется функция с изменением соответствующего аргумента на одну сигму при фиксированном значении остальных аргументов:

$$\beta_k = a_k \frac{\sigma_k}{\sigma_y},$$

где:

β_k - коэффициент β k - го фактора;

σ_k - среднее квадратическое отклонение k - го фактора;

σ_y - среднее квадратическое отклонение функции;

a_k - коэффициент регрессии при k - м факторе.

Из двух факторов x_i, и x_j может быть исключен тот фактор, который имеет меньшее значение β. Допустим, исключению подлежит один из мультиколлинеарных факторов x_i или x_j. Оба фактора управляемы на уровне предприятия, коэффициенты регрессии a_i и a_j статистически значимы. Фактор x_i - более тесно связан с y, то есть $r(x_i; y) \geq r(x_j; y)$, но при этом $\beta_i < \beta_j$. В этом случае обычно исключению подлежит фактор x_j.

11 Проверка коэффициентов регрессии на статистическую значимость

Проверка может быть произведена двумя способами.

Проверка статистической значимости a_k по ***критерию Стьюдента*** проводится по следующей формуле:

$$t_k = \frac{a_k}{S_{a_k}},$$

где:

a_k - коэффициент регрессии при *k-м* факторе;

S_{a_k} - стандартное отклонение оценки параметра a_k.

Число степеней свободы статистики t_k равно $f = n - m - 1$, где m — количество факторов, включенных в модель. Значение t_k, вычисляемое по формуле, сравнивают с критическим значением $t_{f,\alpha}$, найденным по таблицам для распределения Стьюдента ([1]) при заданном уровне значимости α и числе степеней свободы f (двухсторонняя критическая область).

Если $t_k \geq t_{f,\alpha}$, то a_k существенно больше 0, а фактор x_k оказывает существенное влияние на y. При этом фактор x_k оставляем в модели. Если $t_k < t_{f,\alpha}$, то фактор исключаем из модели.

Проверка статистической значимости a_k по ***критерию Фишера:***

$$F_k = \left(\frac{a_k}{S_{a_k}}\right)^2 = t^2,$$

где:

t^2 - многомерный аналог критерия Стьюдента.

Число степеней свободы статистики F_k следующее: $f_1 = 1, f_2 = n - m - 1$. Значение F_k, вычисляемое по формуле, сравнивается с критическим значением $F_{f_1, f_2, \alpha}$, найденным по таблицам для распределения Фишера ([1]), при заданных уровнях значимости α и числе степеней свободы f_1, f_2.

Если $F_k \geq F_{f_1,f_2,\alpha}$, то a_k ¯ существенно больше 0, а фактор x_k оказывает существенное влияние на y. При этом фактор x_k оставляем в модели. Если $F_k < F_{f_1,f_2,\alpha}$, то фактор исключаем из модели.

12 Анализ факторов на управляемость

В ходе логического анализа на основе экономических знаний исследователь должен сделать вывод: можно ли разработать организационно-технические мероприятия, направленные на улучшение (изменение) выбранных факторов на уровне предприятия. Если возможно, то данные факторы управляемы. Неуправляемые факторы на уровне предприятия могут быть исключены из модели. Например, из двух факторов x_1 — средняя техническая скорость автомобилей и x_2 — время погрузки-разгрузки на одну ездку при равенстве или близких по значению таких характеристик, как $r(x_1, y), r(x_2, y), \beta_1, \beta_2$, исключению подлежит x_1. На уровне АТП практически невозможно повлиять на значение технической скорости, которая зависит в основном от климатических условий и величины транспортного потока.

13 Проверка адекватности модели

Данный этап анализа включает ***оценку значимости коэффициента детерминации***. Данная оценка необходима для решения вопроса: оказывают ли выбранные факторы влияние на зависимую переменную? Оценку значимости R^2 следует проводить, так как может сложиться такая ситуация, когда величина коэффициента детерминации будет целиком обусловлена случайными колебаниями в выборке, на основании которой он вычислен. Это объясняется тем, что величина R^2 существенно зависит от объема выборки.

Для оценки значимости коэффициента множественной детерминации используется следующая статистика:

$$F = \frac{R^2(n-m-1)}{m(1-R^2)},$$

которая имеет F-распределение (распределение Фишера) с $f_1 = m$ и $f_2 = n-m-1$ степенями свободы. Здесь n - объем выборки, а m - количество учитываемых объясняющих переменных (факторов).

Значение статистики F, вычисленное по эмпирическим данным, сравнивается с табличным значением $F_{f_1,f_2,\alpha}$. Критическое значение определяется по таблицам для распределения Фишера ([1]), при заданных уровнях значимости α и числе степеней свободы f_1, f_2. Если $F_k \geq F_{f_1,f_2,\alpha}$, то вычисленный коэффициент детерминации значимо отличается от 0 и, следовательно, включенные в регрессию переменные достаточно объясняют зависимую переменную, что позволяет говорить о значимости самой регрессии (модели).

14 Проверка качества подбора теоретического уравнения

Она проводится с использованием средней ошибки аппроксимации. Средняя ошибка аппроксимации регрессии определяется по формуле:

$$E = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - y_{iT}}{y_{iT}}\right)\cdot 100\%,$$

где:

y_{iT} - значение, рассчитанное по уравнению регрессии.

Вычисление специальных показателей, которые применяются для характеристики воздействия отдельных факторов на результирующий показатель. Это ***коэффициент эластичности***, который показывает, на сколько процентов в среднем изменяется функция с изменением аргумента на 1% при фиксированных значениях других аргументов:

$$Э_k = a_k\,\frac{\bar{x}_k}{\bar{y}}.$$

Доля влияния каждого фактора x_k в отдельности на вариацию y: $g_j = \beta_j^2$, где β_j - коэффициент учета фактора x_k.

Показатель g_j является мерой вариации результативного признака за счет изолированного влияния фактора x_j. Следует отметить, что система факторов, входящая в модель регрессии, это не простая их сумма, так как система предполагает внутренние связи, взаимодействие составляющих ее элементов. Действие системы не равно арифметической сумме воздействий составляющих ее элементов. Поэтому необходимо определить показатель системного эффекта факторов η_s:

$$\eta_s = R^2 - \sum_{j=1}^{m}\beta_j^2.$$

На основе анализа специальных показателей и значений парной корреляции x с y делают вывод, какие из главных факторов оказывают наибольшее влияние на y. После этого переходят к разработке организационно-технических мероприятий, направленных на улучшение значений этих факторов, с целью повышения (снижения) результативного показателя y.

15 Экономическая интерпретация

Результаты регрессионного анализа сравниваются с гипотезами, сформулированными на первом этапе исследования, и оценивается их правдоподобие с экономической точки зрения.

16 Прогнозирование неизвестных значений зависимой переменной

Полученное уравнение регрессии находит практическое применение в прогностическом анализе. Прогноз получают путем подстановки в регрессию с численно оцененными параметрами значений факторов. Следует подчеркнуть, что прогнозирование результатов по регрессии лучше поддается содержательной интерпретации, чем простая экстраполяция тенденций, так как полнее учитывается природа исследуемого явления.

Приложение Расчет коэффициентов множественной линейной регрессии

Пусть $e_i = y_i - a_0 - a_1 x_{i1} - a_2 x_{i2} - ... - a_m x_{im}$, где i - индекс наблюдения. Сумма квадратов отклонений e_i как произведение вектора–строки $\{e_i\} = e^T$ на вектор–столбец $\{e_i\} = e$ (e^T - вектор–столбец, транспонированный в строку). Вектор–столбец e может быть записан как $e = y - Xa$, где y - вектор-столбец наблюдений зависимой переменной, X - матрица $n(m+1)$, в которой каждая из n строк представляет наблюдение вектора значений независимых переменных x_i:

$$x = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1m} \\ 1 & x_{21} & x_{22} & ... & x_{2m} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nm} \end{pmatrix};$$ a - вектор–столбец $(a_0, a_1, ..., a_m)^T$.

Отсюда:

$$Q = e^T e = (y - Xa)^T (y - Xa) = y^T y - a^T X^T y - y^T Xa + a^T X^T Xa = y^T y - 2a^T X^T y + a^T X^T Xa.$$

Теперь нужно записать необходимые условия экстремума выражения Q. Оно состоит в равенстве нулю всех частных производных $\frac{\partial Q}{\partial a_i}$. Вектор $\frac{\partial Q}{\partial a}$ можно записать компактно как $\frac{\partial Q}{\partial a} = -2X^T y + 2(X^T X)a$. Это можно показать следующим образом. Пусть $(X^T X) = X'$ - матрица размерности $(m+1)(m+1)$.

$$Q_1 = a^T X' a = \left((a_0, a_1, ..., a_m) \cdot \begin{pmatrix} x'_{11} & x'_{12} & ... & x'_{1,m+1} \\ x'_{21} & x'_{22} & ... & x'_{2,m+1} \\ ... & ... & ... & ... \\ x'_{m+1,1} & x'_{m+1,2} & ... & x'_{m+1,m+1} \end{pmatrix} \right) \cdot \begin{pmatrix} a_0 \\ a_1 \\ ... \\ a_m \end{pmatrix} = \left(\sum_j a_j x'_{j1} + \sum_j a_j x'_{j2} + ... + \right.$$ -

$$\left. + \sum_j a_j x'_{j,m+1} \right) \cdot (a) = \sum_i a_i \sum_j a_j x'_{ji} = \sum_{ij} a_i a_j x'_{ij}.$$

Отсюда видно, что $\frac{\partial Q_1}{\partial a_i} = 2\sum_i a_j x'_{ij}$, то есть $\frac{\partial Q_1}{\partial a_i} = 2(X^T X)a$. Если обозначить вектор $X^T y = y'$, то $a^T y' = \sum_j a_j y'_j$ и $\frac{\partial (a^T y')}{\partial a} = y' = X^T y$.

Поскольку $Q = y^T y - 2a^T y' + Q_1$ и $\frac{\partial y^T y}{\partial a} = 0$, $\frac{\partial Q}{\partial a} = 0 \Rightarrow -2X^T y + 2(X^T X) = 0$

и $X^T y = X^T X a$, откуда $\underline{a = \left(X^T X\right)^{-1} X^T y}$. Таким образом мы получили формулу расчета вектора коэффициентов регрессии в векторно–матричной форме.

Список литературы

1. Бережная Е.В., Бережной В.И. Математические методы моделирования экономических систем. –М.: ФиС, 2001.
2. Замков О.О., Толстопятенко А.В., Черемных Ю.Н. Математические методы в экономике. -М.: Дис, 2001.
3. Ивашев-Мусатов О.С. Теория вероятностей и математическая статистика. – М.: Наука, 1979.
4. Красс М.С., Чупрынов Б.П. Основы математики и ее приложения в экономическом образовании. – М.: Дело, 2001.

Содержание

Аликас Вадим Эдуардович
Чеботина Марина Сергеевна

Эконометрические методы исследования моделей

Методические указания
по дисциплине «Эконометрика»
для студентов очной формы обучения специальностей
«Финансы и кредит» - 080105(060400),
«Бухучет, анализ и аудит» - 080109(060500)

Редактор

Подписано к печати	Формат 60x84 1/16	Бумага тип. № 1
Печать трафаретная	Усл. печ. л. 2,25	Уч. - .изд. л. 2,25
Заказ	Тираж 150	Цена свободная

Редакционно-издательский центр КГУ
640669, г. Курган, ул. Гоголя, 25
Курганский государственный университет