ЭКОНОМЕТРИЧЕСКИЙ АНАЛИЗ РЫНКА СТРОИТЕЛЬСТВА И НЕДВИЖИМОСТИ

Управление недвижимостью

БГТУ, 2012

Рекомендуемая литература

- Доугерти, К. Введение в эконометрику : учеб. для студ. эконом. спец. вузов : пер. с англ. / К. Доугерти. – 3-е изд. – М. : ИНФРА-М, 2009. – 452 с.
- Handbook of Applied Econometrics and Statistical Inference / ed. by A. Ullah, A. T.K. Wan, A. Chaturvedi. – Houndmills: Palgrave Macmillan, 2002. – 718 p.
- Patterson, K. Palgrave Handbook of Econometrics: Vol. 2: Applied Econometrics / K. Patterson, T. C. Mills. – Palgrave Macmillan, 2009. – 1128 p.
- Wang, P. Econometric Analysis of the Real Estate Market and Investment / P.Wang. – Routledge, 2001. – 208 p.
- Wooldridge, J. Introductory Econometrics: A Modern Approach / J. Wooldridge. – Belmont: South-Western, 2009.

План курса

1	Введение в дисциплину «Эконометрический анализ рынка строительства и недвижимости».	2
2	Парный регрессионный анализ	4
3	Свойства коэффициентов регрессии и проверка гипотез	4
4	Множественный регрессионный анализ	4
5	Преобразования и спецификация переменных регрессии	4
6	Модели двоичного выбора	4
7	Моделирование и свойства регрессионных моделей с временными рядами	6
8	Модели с панельными данными	6
	ИТОГО	**34**

1. Вводная лекция

- Место эконометрики в системе наук
- Особенности статистических данных
- Случайные переменные и теория выборок
- Ковариация, дисперсия и корреляция

Введение

- Современное экономическое образование держится на трех китах:
  - макроэкономике,
  - микроэкономике
  - и эконометрике.

- Поскольку модели **неполны**, а используемые данные **несовершенны**, эконометрика посвящена **методам**, которые могут работать с такими моделями и данными.

Случайные факторы в модели

- Факторы, не учтенные явно в модели рассматриваются как **случайные**.
- Например в модели спроса **q=f(p,I)+u**

q- количество блага, p – цена, I – доход, **u – показывает суммарное влияние всех неучтенных (случайных) факторов**.

Определение эконометрики

- Эконометрика как наука расположена между
  - **экономикой**
  - **статистикой**
  - **и математикой**.

- Поскольку **эконометрика тесно связана с экономикой, математикой и статистикой**, то исследователь, использующий эконометрические методы, должен быть :
  - **Экономистом**, чтобы применять экономическую теорию к анализу эмпирических данных;

  - **Математиком**, чтобы формулировать экономическую теорию на математическом языке;
  - **Специалистом в экономической статистике**, чтобы разбираться в процессах сбора и обработки экономических данных;

  - **Специалистом в математической статистике**, чтобы применять для анализа эмпирических данных статистические методы;
  - Уметь работать со статистическими или эконометрическими пакетами, без использования которых сегодня невозможно исследование.

- **Таким образом, для понимания и применения эконометрики нужно быть в достаточной степени образованным по широкому спектру экономико-математических дисциплин.**

Место эконометрики в системе наук

- Выражая экономические законы в форме математических соотношений, **математическая экономика** не измеряет входящие в эти соотношения переменные и не проверяет теорию на практике.
- Это - задача **эконометрики**, для этого приходится приводить уравнения к допускающей эмпирическую проверку форме.

- Особенно много общего у **эконометрики и статистики**.
- Но если экономическая статистика ограничивается обработкой и представлением эмпирических данных в виде таблиц и графиков, то эконометрика использует их как **первичные данные для проверки экономической теории**.

- Для этого применяется **математическая статистика и теория вероятности**.
- Кроме этого, эконометрика разрабатывает специфические методы, учитывающие **природу экономических данных**, которые не являются результатом специально поставленного эксперимента.

- **Основные результаты экономической теории носят не количественный, а качественный характер.**

- Так, из теории следует, что при прочих равных условиях **повышение цены товара ведет к уменьшению спроса на него**.
- Однако, вопрос о величине снижения спроса при увеличении цены конкретного товара в конкретных условиях выходит за рамки теории.
- **Ответ на него пытается дать эконометрика, внося эмпирическое содержание в экономическую теорию**.

- Эконометрика связана с **эмпирическим** выводом экономических законов.
- Т.е. исследователь **использует данные или наблюдения для того, чтобы получить количественные зависимости для экономических соотношений.**

- Данные, как правило, не являются экспериментальными, так как в экономике невозможно проводить **многократные эксперименты**.

Виды данных

- Перекрестные данные (cross-section data)
- Временные ряды (time series)
- Панельные данные (panel data)

- **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 принадлежность к определенному моменту времени несущественна.

- Например: данные бюджетных обследований населения в определенный момент времени.

- **Временные ряды** – это данные, характеризующие один и тот же объект, но в различные моменты времени.
- Например: данные о динамике уровня инфляции за определенный период.

- Данные **временных рядов** характеризуются зависимостями их последовательных значений, например, могут быть связаны между собой последовательные отклонения от общей тенденции развития, могут быть задержки (временные лаги).

- Поэтому для временных рядов применяются специальные методы обработки и анализа по сравнению с перекрестными данными.

Панельные данные

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	5.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2

Источники статистических данных

Существуют различные методы сбора экономических данных:

- опрос;
- анкетирование и интервьюирование;
- статистическая отчетность и т.д.

Источники:

- Белстат
- Отраслевые министерства (Госкомимущества)
- Кадастровое агентство

- Основные источники статистических данных можно разделить на две группы:
  - Внутренние
  - Внешние

- К внутренним источникам относятся те виды и формы статистического наблюдения, которые организует Белстат:

A. Отчетность предприятий
B. Регистр предприятий
C. Переписи и обследования

- К внешним источникам относят те виды и формы статистического наблюдения, которые организуют другие ведомства:

A. Административные источники
B. Денежная и банковская статистика
C. Платежный баланс
D. Таможенная статистика

Случайные переменные и теория выборок

Дискретная случайная величина

- Случайная переменная - это любая переменная, значение которой не может быть точно предсказано.
- Дискретной называется случайная величина, имеющая определенный набор возможных значений.

- Пример - сумма выпавших очков при бросании двух игральных костей.
- Пример непрерывной случайной величины, - температура в комнате.
- Она может принять любое из непрерывного диапазона значений.

Дискретная случайная величина

- Рассмотрим пример с двумя игральными костями. Предположим, что одна из костей зеленая, а другая - красная.
- Если их бросить, то возможны 36 элементарных исходов эксперимента, поскольку на зеленой кости может выпасть любое число от 1 до 6 и то же самое - на красной.
- Случайная переменная, определенная как их **сумма (x),** может принимать одно из **11 числовых значений - от 2 до 12.**

Красная	Зеленая					
	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

- Поскольку на костях имеется 36 различных комбинаций, каждый исход имеет вероятность 1/36. Лишь одна из возможных комбинаций {зеленая=1, красная=1} дает сумму, равную 2, так что вероятность x=2 равна **1/36**.
- Чтобы получить сумму **x=7**, нам потребуются сочетания {зеленая=1, красная=6}, либо {зеленая=2, красная=5}, либо {зеленая=3, красная=4}, либо {зеленая=4, красная=3}, либо {зеленая=5, красная=2}, либо {зеленая=6, красная=1}.
- 6 возможных исходов, и поэтому вероятность получения 7 равна **6/36**.

- Вероятности приведены в следующей таблице:

x	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Совокупность всех возможных значений случайной переменной описывается **генеральной совокупностью**, из которой извлекаются эти значения. В нашем случае генеральная совокупность - это набор чисел от 2 до 12.

Значения случайной переменной и их вероятности – закон распределения случайной величины

Выборки

- В большинстве случаев рассматривают только часть наблюдений, взятых из генеральной совокупности, которое называют **выборкой.**
- Выборка объема **n** — это результат наблюдений случайной величины в вероятностном эксперименте, который повторяется **n** раз в одних и тех же условиях, а следовательно при неизменном распределении **x**.

- Выборку называют **репрезентативной** (представительной), если она достаточно полно представляет изучаемые признаки и параметры генеральной совокупности.
- Для репрезентативности выборки важно **обеспечить случайность отбора**, с тем, чтобы все объекты генеральной совокупности имели равные вероятности попасть в выборку.

- Для обеспечения репрезентативности выборки применяют следующие способы отбора:
  - **Простой отбор** — последовательно отбирается первый случайно попавший объект
  - **Типический отбор** — объекты отбираются пропорционально представительству различных типов объектов в генеральной совокупности
  - **Случайный отбор** — например, с помощью таблицы случайных чисел и т.д.

- В эконометрике всегда известна **только выборка** из некоторого количества наблюдений случайной величины и по данным выборки можно рассчитать только **выборочные а не теоретические** характеристики случайной величины.

Математическое ожидание случайной величины

- **МО - это взвешенное среднее всех ее возможных значений, причем в качестве весового коэффициента берется вероятность соответствующего исхода**.
- МО можно рассчитать, перемножив все возможные значения случайной величины на их вероятности и просуммировав полученные произведения.
- МО случайной величины обозначают **E(x)** или **M(x)** или μ_x.

- Если x принимает n конкретных значений (x1, x2,...,xn) и вероятность получения x_i равна p_i тогда

$$E(x) = x_1 p_1 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i$$

- В случае с двумя костями величинами от x_1 до x_n были числа от 2 до 12: x1=2, x2=3,...,x11=12

и p1=1/36, p2=3/36,...,p11=1/36.

МО рассчитывается так:

(2*1/36)+(3*2/36)+(4*3/36)+...+(11*2/36)+(12*1/36)=7

- В случае с одной костью x меняется от 1 до 6 с равной вероятностью 1/6.
- E(x)= 1·1/6+2·1/6+ 3·1/6+4·1/6+ 5·1/6+6·1/6=**3,5**
- МО случайной величины часто называют ее **средним по генеральной совокупности**.

- МО **функций** дискретных случайных переменных вычисляются по формуле:

E{g(x)}=Σg(xi)*pi,

где суммирование производится по всем возможным значениям x и g(x) - некоторая функция от x.

- Существует 3 правила расчета МО и они одинаковы применимы для дискретных и непрерывных случайных переменных.

- **Правило 1**: Математическое ожидание суммы нескольких переменных равно сумме математических ожиданий. Если имеются 3 случайные переменные, то **E(x+y+z)=E(x)+E(y)+E(z).**
- **Правило 2**: Если случайная переменная умножается на константу, то ее математическое ожидание умножается на ту же константу. Если x - случайная переменная и a - константа, то **E(ax)=aE(x).**
- **Правило 3**:Математическое ожидание константы есть она сама. Если a - константа, то **E(a)=a**.

- Две случайные переменные x и y называются независимыми, если **E{f(x)g(y)}=E{f(x)}E{g(y)}** для любых функций f(x) и g(y).
- Из независимости следует как важный частный случай, что

E(xy)= E(x)E(y).

Теоретическая дисперсия

- Теоретическая дисперсия является **мерой разброса** для вероятностного распределения, обозначается σ_x^2 или **pop.var(x)**
- Она определяется как МО квадрата разности между величиной x и ее средним, т.е. величины $(x-\mu)^2$, где μ- математическое ожидание x, :

$$(\sigma_x)^2 = pop.var(x) = E[(x-\mu)^2] = \sum_{i=1}^{n}(x_i-\mu)^2 p_i = (x_1-\mu)^2 p_1 + \ldots + (x_n-\mu)^2 p_n$$

- Из σ_x^2 можно получить σ_x - теоретическое стандартное отклонение случайной величины— **квадратный корень из ее дисперсии**.

- Расчет дисперсии на примере с одной игральной костью. Поскольку μ=E(x)=3,5, то $(x-\mu)^2$ в этом случае равно $(x-3,5)^2$.

X_i	p_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$(xi-\mu)^2*pi$
1	1/6	-2,5	6,25	1,042
2	1/6	-1,5	2,25	0,375
3	1/6	-0,5	0,25	0,042
4	1/6	0,5	0,25	0,042
5	1/6	1,5	2,25	0,375
6	1/6	2,5	6,25	1,042
всего				2,92

- Формула расчета теоретической дисперсии случайной переменной, может быть записана $\sigma^2=E(x^2)-\mu^2$.
- Это выражение иногда более удобно, чем первоначальное определение.

Постоянная и случайная составляющие случайной переменной

- Случайную величину можно разбить на постоянную (МО) и чисто случайную составляющие.
- Если **x** - случайная переменная и **μ** - ее МО, то декомпозиция случайной величины записывается следующим образом:

x=μ+u, где u - чисто случайная составляющая,

либо **u=x-μ.**

- Из определения следует, что МО величины **u** равно нулю.
- Из уравнения u=x-μ имеем:

E(u)=E(x-μ)=E(x)-E(μ)=E(x)-μ=μ-μ=0.

- Поскольку весь разброс значений x обусловлен u, то теоретическая дисперсия x равна теоретической дисперсии u.
- По определению,

$\sigma_x^2=E\{(x-\mu)^2\}=E\{u\}^2$

$\sigma_u^2=E\{(U-MO(u))^2\}=E\{(u-0)^2\}=E\{u\}^2$.

- Таким образом, **σ^2** может быть эквивалентно определена как дисперсия x или u.
- Итак, если x - случайная переменная, определенная по формуле $x=\mu+u$, где μ - заданное число и u - случайный член с $E(u)=0$ и дисперсией σ^2, то МО величины x равно μ, а дисперсия - σ^2.

Способы оценивания и оценки

- На практике, за исключением искусственно простых случайных величин, мы не знаем точного вероятностного распределения или плотности распределения вероятностей.
- Это означает, что неизвестны также и теоретическое МО, и дисперсия.
- Тем не менее, необходимы **оценки** этих или других теоретических характеристик генеральной совокупности.

- Процедура оценивания всегда одинакова.
- Берется выборка из **n** наблюдений, и с помощью подходящей формулы рассчитывается оценка нужной характеристики.
- **Способ оценивания - это общее правило, или формула, а значение оценки - это конкретное число, которое меняется от выборки к выборке**.

- Выборочное среднее $\bar{x}$ обычно дает оценку для МО, а формула **s^2** - оценку дисперсии генеральной совокупности.

Характеристики генеральной совокупности	Формулы оценивания
Математическое ожидание μ	$\bar{x}=\frac{1}{n}\sum x_i$
Дисперсия, σ^2	$s^2=\frac{1}{n-1}\sum (x_i-\bar{x})^2$

- Это обычные формулы оценки МО и дисперсии генеральной совокупности, однако не единственные.
- Причина, по которой они используется, в том, что эта оценка в наилучшей степени соответствует двум очень важным критериям - **несмещенность и эффективности**.

Оценки как случайные величины

- Получаемая оценка представляет частный случай случайной переменной.
- Причина здесь в том, что сочетание значений **x** в выборке случайно, поскольку **x** - случайная переменная и, следовательно, случайной величиной является и функция набора ее значений.

- Возьмем, например, $\bar{x}$ - оценку МО:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

Величина x в i-ом наблюдении может быть разложена на две составляющие: постоянную часть μ и чисто случайную составляющую u_i $x_i = \mu + u_i$. Следовательно

$$\bar{x} = \mu + \bar{u}$$

$\bar{u}$ — выборочное среднее величин u_i.

- Можно увидеть, что $\bar{x}$ подобно x, имеет как фиксированную, так и чисто случайную составляющие.
- Ее фиксированная составляющая - μ, то есть МО x, а ее случайная составляющая - $\bar{u}$ то есть среднее значение чисто случайной составляющей в выборке.

- Величина $\mathbf{s^2}$ - оценка теоретической дисперсии **x** - также является случайной переменной.

Вычитая из $x_i = \mu + u_i$ $\bar{x} = \mu + \bar{u}$ имеем

$$x_i - \bar{x} = u_i - \bar{u}$$

следовательно $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{n-1}\sum \{(u_i - \bar{u})^2\}$

Таким образом, s^2 зависит только от чисто случайной составляющей наблюдений x в выборке. Поскольку эти составляющие меняются от выборки к выборке, **также от выборки к выборке меняется и величина оценки s^2**.

Несмещенность, эффективность, состоятельность

- Поскольку оценки являются случайными переменными, их значения лишь по случайному совпадению могут в точности равняться характеристикам генеральной совокупности.
- Обычно будет присутствовать определенная ошибка, которая может быть большой или малой, положительной или отрицательной, в зависимости от чисто случайных составляющих величин x в выборке.

- Желательно, чтобы оценка в среднем за достаточно длительный период была **аккуратной**.
- Выражаясь формально, мы хотим, чтобы **МО оценки равнялось бы соответствующей характеристике генеральной совокупности**.
- Если это так, то оценка называется **несмещенной**.
- Если это не так, то оценка называется **смещенной** , и разница между ее МО и соответствующей теоретической характеристикой генеральной совокупности называется **смещением**.

- Рассмотрим выборочное среднее. Является ли оно несмещенной оценкой теоретического среднего?
- Да, это так, что вытекает из

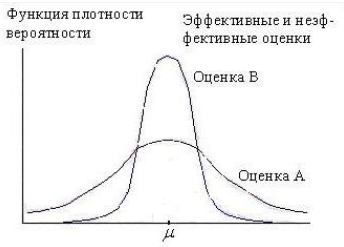
$$\bar{x} = \mu + \bar{u}$$

$$E(\bar{x}) = E(\mu + \bar{u}) = E(\mu) + E(\bar{u}) = \mu + 0 = \mu$$

- Величина s^2 является оценкой теоретической дисперсии σ^2.
- Можно показать, что МО s^2 равно σ^2, и эта величина является **несмещенной** оценкой теоретической дисперсии, если наблюдения в выборке независимы друг от друга.

- Еще одна важная сторона оценок - это **надежность**.
- Необходимо, чтобы оценка с максимально возможной вероятностью давала бы близкое значение к теоретической характеристике, что означает желание **получить функцию плотности вероятности, как можно более "сжатую" вокруг истинного значения**.
- Один из способов выразить это требование - сказать, что мы хотим получить сколь возможно **малую дисперсию**.

- Предположим, что мы имеем две оценки теоретического среднего, рассчитанные на основе одной и той же информации, что обе они являются несмещенными.



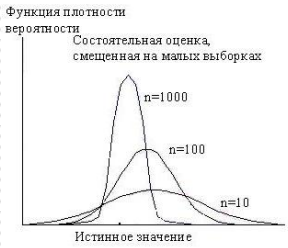
Функция плотности вероятности для оценки B более "сжата", чем для оценки A, с ее помощью получим более точное значение. Эта оценка более **эффективна.**

- Хотя оценка B более эффективна, это не означает, что она всегда дает более точное значение.
- При определенном стечении обстоятельств значение оценки A может быть ближе к истине.
- **Эффективная** оценка - это та, у которой **дисперсия минимальна**.

- Замечания:
- Эффективность оценок можно сравнивать лишь тогда, когда они используют одну и ту же информацию, например один и тот же набор наблюдений нескольких случайных переменных.
- Если одна из оценок использует в 10 раз больше информации, чем другая, то она вполне может иметь меньшую дисперсию, но было бы неправильно считать ее более эффективной.

- Если **предел оценки** по вероятности равен истинному значению характеристики генеральной совокупности, то эта оценка называется **состоятельной**.
- Иначе говоря, **состоятельной** называется такая оценка, которая дает точное значение для **большой выборки** независимо от входящих в нее конкретных наблюдений.

- На рисунке показано как при различных размерах выборки может выглядеть распределение вероятностей.



- При увеличении размера выборки распределение становится симметричным вокруг истинного значения.
- Это **состоятельная оценка.**

- Иногда невозможно найти оценку, несмещенную на малых выборках.
- Если при этом вы можете найти хотя бы состоятельную оценку, это может быть лучше, чем не иметь никакой оценки.

МО и дисперсия непрерывной случайной величины

- Определение МО непрерывной случайной переменной :

$E(x)=\int x*f(x)dx,$

где интегрирование производится на всем интервале, где определена функция f(x).

Возможные значения x взвешиваются по соответствующим им вероятностям.

- σ^2 - теоретическая дисперсия x:

$\sigma^2=E\{(x-\mu)^2\}=\int(x-\mu)^2f(x)dx.$

- Теоретическое отклонение (σ) получаем извлечением квадратного корня из дисперсии.

Ковариация, дисперсия и корреляция

Выборочная и теоретическая ковариации

- Ковариация является мерой взаимосвязи между двумя переменными
- Если x и y - случайные величины, то **теоретическая ковариация** определяется как математическое ожидание произведения отклонений этих величин от их средних значений:

$$\mathbf{cov}(x,y) = \sigma_{xy} = E[(x-\mu_x)(y-\mu_y)]$$

•где μ_x и μ_y - теоретические средние значения x и y соответственно.

- При наличии n наблюдений двух переменных (x и y) **выборочная ковариация** между x и y задается формулой:

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

- Если теоретическая ковариация неизвестна, то для ее оценки может быть использована **выборочная ковариация**, вычисленная по ряду наблюдений.

- Эта оценка будет иметь отрицательное смещение.
- Причина заключается в том, что выборочные отклонения измеряются по отношению к выборочным средним значениям величин x и y и **имеют тенденцию к занижению отклонений** от истинных средних значений.

- Можно рассчитать несмещенную оценку путем умножения выборочной оценки на **n / (n - 1)** .
- **Если x и y независимы, то их теоретическая ковариация равна нулю**.

Пример расчета ковариации

- Со времен нефтяного кризиса 1973 г. реальная цена на бензин, т.е. цена бензина, отнесенная к уровню общей инфляции, значительно возросла, и это оказало заметное воздействие на потребительский спрос.
- В период между 1963 и 1972 гг. потребительский спрос на бензин устойчиво повышался.
- Эта тенденция прекратилась в 1973 г., а затем последовали нерегулярные колебания спроса с незначительным его падением в целом.

Потребительские расходы на бензин и его реальная цена в США		
Год	Расходы (млрд. долл., цены 1972 г.)	Индекс реальных цен (1972 = 100)
1973	26,2	103,5
1974	24,8	127,0
1975	25,6	126,0
1976	26,8	124,8
1977	27,7	124,7
1978	28,3	121,6
1979	27,4	149,7
1980	25,1	188,8
1981	25,2	193,6
1982	25,6	173,9

- В таблице приведены данные о потребительском спросе и реальных ценах после нефтяного кризиса

- Реальная цена вычислялась путем деления индекса номинальной цены на бензин, на общий индекс потребительских цен и умножения результата на 100.
- Индексы основаны на данных 1972 г.; индекс реальной цены показывает повышение цены бензина относительно общей инфляции начиная с 1972г.

Эти данные показаны в виде диаграммы рассеяния.

Можно видеть отрицательную связь между потребительским спросом на бензин и его реальной ценой.

- Показатель выборочной ковариации позволяет выразить данную связь единым числом.
- Для его вычисления мы сначала находим средние значения цены и спроса на бензин.

Наблюдение	p	y
1973	103,5	26,2
1974	127,0	24,8
1975	126,0	25,6
1976	124,8	26,8
1977	124,7	27,7
1978	121,6	28,3
1979	149,7	27,4
1980	188,8	25,1
1981	193,6	25,2
1982	173,9	25,6
Сумма	**1433,6**	**262,7**
Среднее	**143,36**	**26,27**

- Обозначив цену через **p** и спрос через **y**, определяем средние значения, которые оказываются равными соответственно **143,36 и 26,27.**

- Затем для каждого года вычисляем отклонение величин **p** и **y** от средних и перемножаем их.

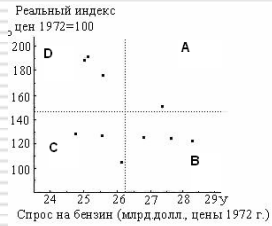
Наблюдение	p	y	$(p-\bar{p})$	$(y-\bar{y})$	$(p-\bar{p})(y-\bar{y})$
1973	103,5	26,2	-39,86	-0,07	2,79
1974	127,0	24,8	-16,36	-1,47	24,05
1975	126,0	25,6	-17,36	-0,67	11,63
1976	124,8	26,8	-18,56	0,53	-9,84
1977	124,7	27,7	-18,66	1,43	-26,68
1978	121,6	28,3	-21,76	2,03	-44,17
1979	149,7	27,4	6,34	1,13	7,16
1980	188,8	25,1	45,44	-1,17	-53,16
1981	193,6	25,2	50,24	-1,07	-53,76
1982	173,9	25,6	30,54	-0,67	-20,46
Сумма	**1433,6**	**262,7**			**-162,44**
Среднее	**143,36**	**26,27**			**-16,24**

В нижней клетке последнего столбца определяется средняя величина (-16,24), она является значением выборочной ковариации

- Ковариация в данном случае отрицательна.
- Так это и должно быть.
- Отрицательная связь, как это имеет место в данном примере, выражается отрицательной ковариацией, а положительная связь - положительной ковариацией.

- На рисунке диаграмма рассеяния наблюдений делится на четыре части вертикальной и горизонтальной линиями, проведенными через средние значения p и v соответственно.

- Пересечение этих линий образует точку , которая показывает **среднюю цену и средний спрос** за период, соответствующий выборке.



Для любого наблюдения, лежащего в квадранте **A**, значения реальной цены и спроса **выше** соответствующих средних значений.

$(p-\bar{p})$ $(y-\bar{y})$

Здесь , и являются положительными, а поэтому должно $(p-\bar{p})(y-\bar{y})$ жительным и

Наблюдения дают положительный вклад в ковариацию.

В квадранте B наблюдения имеют реальную цену **ниже** средней и спрос **выше** среднего. Наблюдения дают отрицательный вклад в ковариацию.

В квадранте C как реальная цена, так и спрос **ниже** своих средних значений. Наблюдения дают положительный вклад в ковариацию.

В квадранте D реальная цена **выше** средней, а спрос **ниже** среднего. Наблюдения дают отрицательный вклад в ковариацию

- Поскольку выборочная ковариация является средней величиной произведения для 10 наблюдений, она будет **положительной, если положительные вклады будут доминировать над отрицательными, и отрицательной, если будут доминировать отрицательные вклады**.
- Положительные вклады исходят из квадрантов A и C, и ковариация будет, скорее всего, положительной, если основной разброс пойдет по наклонной вверх.

- Точно так же отрицательные вклады исходят из квадрантов B и D.
- Поэтому если основное рассеяние идет по наклонной вниз, как в данном примере, то ковариация будет, скорее всего, отрицательной.

Правила расчета ковариации

- Существует несколько правил, которые вытекают непосредственно из определения ковариации.
- Правило 1:

Если y = v + w, то

Cov(x, y) = Cov(x, v) + Cov(x, w).

- Допустим, имеются данные по 6 семьям: общий годовой доход (x); расходы на питание и одежду (y), расходы на питание (v), расходы на одежду (w). Естественно, y = v + w

Семья	Доход семьи (x)	Расходы на питание и одежду (y)	расходы на питание (v)	Расходы на одежду (w)
1	3000	1100	850	250
2	2500	850	700	150
3	4000	1200	950	250
4	6000	1600	1150	450
5	3300	1000	800	200
6	4500	1300	950	350

Семья	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(v-\bar{v})$	$(x-\bar{x})(v-\bar{v})$	$(w-\bar{w})$	$(x-\bar{x})(w-\bar{w})$
1	-883	-75	66250	-50	44167	-25	22083
2	-1383	-325	449583	-200	276667	-125	172917
3	117	25	2917	50	5833	-25	-2917
4	2117	425	899586	250	529167	175	370416
5	-583	-175	102083	-100	58333	-75	43750
6	617	125	77083	50	30833	75	46250
Сумма			**1597500**		**945000**		**652500**
Среднее			**266250**		**157500**		**108750**

Cov(x, v) равна 157500 и Cov(x, w) = 108750.

Мы проверили, что Cov(x, y) = Cov(x, v)+ Cov(x, w).

- Именно так и должно быть. Рассмотрим i - ю семью
- Поскольку
- $y_i = v_i + w_i$ и

$$\bar{y} = \bar{v} + \bar{w}$$

$$(x_i - \bar{x})(y_i - \bar{y}) = (x_i - \bar{x})(v_i + w_i - \bar{v} - \bar{w}) = (x_i - \bar{x})(v_i - \bar{v}) + (x_i - \bar{x})(w_i - \bar{w})$$

Таким образом, вклад семьи i в Cov(x, y) является суммой ее вкладов в Cov(x, v) и Cov(x, w).

Тоже самое справедливо для всех семей и, соответственно, для ковариации в целом.

- Правило 2:
- Если y = a z, где a - константа, то Cov(x, y) = a Cov(x, z).

Семья	Доход семьи (x)	Расходы на питание и одежду (y)	Вторая выборка: расходы семьи на питание и одежду (z)
1	3000	1100	2200
2	2500	850	1700
3	4000	1200	2400
4	6000	1600	3200
5	3300	1000	2000
6	4500	1300	2600

- Последняя колонка (z) дает расходы на питание и одежду для второго множества из 6 семей.
- Каждое наблюдение z=2y.
- Предполагается, что значения величины x для второго набора семей являются такими же, как и ранее.

Семья	$(x-\bar{x})$	$(z-\bar{z})$	$(x-\bar{x})(z-\bar{z})$
1	-883	-150	132500
2	-1383	-650	899167
3	117	50	5833
4	2117	850	1700167
5	-583	-350	204167
6	617	250	154167
Сумма			**3195000**
Среднее			**532500**

Из таблицы можно видеть, что **Cov(x, z)** равна **532500**, что равно **2Cov(x, y)**

Таким образом мы проверили, что **Cov(x, 2y) = 2Cov(x, v).**

$$Cov(x,y) = 2Cov(x,y) = \frac{1}{n}\sum(x_i - \bar{x})(z_i - \bar{z}) = \frac{1}{n}\sum(x_i - \bar{x})(qy_i - \bar{y}) = \frac{a}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = aCov(x,y)$$

- **Правило 3**:
- Если y = a, где a - константа,

то **Cov(x, y) = 0**.

Допустим, что каждая семья в выборке имеет по два взрослых человека, и предположим, что по недоразумению вы решили вычислить ковариацию между общим доходом (x) и числом взрослых в семье (a).

Естественно, что $a_1=a_2...=a_6$ =2= среднему значению.

Поэтому Cov(x, a)=0.

Семья	x	a	$(x - \bar{x})$	$(a - \bar{a})$	$(x - \bar{x})(a - \bar{a})$
1	3000	2	-883	0	0
2	2500	2	-1383	0	0
3	4000	2	117	0	0
4	6000	2	2117	0	0
5	3300	2	-583	0	0
6	4500	2	617	0	0
Сумма	**23300**	**12**			**0**
Среднее	**3883**	**2**			**0**

Выборочная дисперсия, правила расчета дисперсии

- Для выборки из n наблюдений x_1, ...,x_n выборочная дисперсия определяется как среднеквадратичное отклонение в выборке:

$$Var(x) = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Ранее была определена исправленная", или несмещенная, выборочная дисперсия :

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

- Заметим, что дисперсия переменной x может рассматриваться как ковариация между двумя величинами x:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = Cov(x,x)$$

Кроме того можно получить другую формулу:

$$Var(x) = [\frac{1}{n}\sum_{i=1}^{n} (x_i)^2] - \bar{x}^2$$

- Существует несколько правил для расчета дисперсии, которые являются аналогами правил для ковариации.
- **Правило 1**: Если y = v + w,

то Var(y) = Var(v) + Var(w) + 2Cov(v, w).

- Доказательство :

Если y = v + w, то

Var(y) = Cov(y, y) = Cov(y, [v + w]) =

= Cov([v + w], v) + Cov([v + w], w), по правилу ковариации 1,

= Cov(v, v) + Cov(w, v) + Cov(v, w) + Cov(w, w), по правилу ковариации 1,

= Var(v) + Var(w) + 2Cov(v, w).

- **Правило 2**: Если y = a z, где a - константа,

то $Var(y) = a^2 Var(z)$.

- Доказательство:

Дважды используя правило ковариации 2, получим:

Var(y) = Cov(y, y) = Cov(y, az) = a Cov(y, z)=

= a Cov(az, z) = a^2 Cov(z, z) = a^2Var(z).

- **Правило 3**: Если y = a, где a - константа, то Var(y) = 0.
- По правилу ковариации 3 имеем:

Var(y) = Cov(a, a) = 0

- Действительно, если **y** - постоянная, то ее среднее значение является той же самой постоянной и равняется нулю для всех наблюдений.
- Следовательно, **Var(y)=0.**

- **Правило 4**: Если y = v + a, где a - константа, то **Var(y) = Var(v).**
- Доказательство:
- Если y = v + a, где a - константа, то по правилу ковариации 1, используя затем правила 1 и 3 для дисперсии и правило 3 для ковариации, получаем:

Var(y) = Var(v + a) = Var(v) + Var(a) + 2Cov(v, a) = Var(v).

Коэффициент корреляции

- Более точной мерой зависимости между величинами является **коэффициент корреляции.**
- Подобно дисперсии и ковариации, коэффициент корреляции имеет две формы - теоретическую и выборочную.
- Теоретический коэффициент корреляции ρ для переменных x и y определяется следующим образом:

$$\rho_{x,y} = \frac{pop.\mathrm{cov}(x,y)}{\sqrt{pop.\mathrm{var}(x)\,pop.\mathrm{var}(y)}} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x^2\ \sigma_y^2}}$$

- Если x и y независимы, то $\rho_{x,y}$ **=0**, так как равна нулю теоретическая ковариация.
- Если между переменными существует положительная зависимость, то теоретический **коэффициент корреляции будет положительным**.
- Если существует строгая положительная зависимость, то он примет максимальное значение, равное **1**.
- Аналогичным образом при отрицательной зависимости теоретический коэффициент **корреляции будет отрицательным с минимальным значением -1**.

- Выборочный коэффициент корреляции **r** для переменных x и y определяется путем замены теоретических дисперсий и ковариации в формуле теоретического коэффициента корреляции на их несмещенные оценки

$$r_{x,y} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

- Выборочный коэффициент корреляции имеет максимальное значение, равное 1, которое получается при строгой линейной положительной зависимости между выборочными значениями x и y, и минимальное значение -1, когда существует линейная отрицательная зависимость.
- **Величина r=0 показывает, что зависимость между наблюдениями x и y в выборке отсутствует, но это не говорит о том, что ρ=0, и наоборот.**

- Рассмотрим пример расчета корреляции.
- Уже вычислена **Cov(p, y)= -16,24**, поэтому необходимы вычислить только Var(p) и Var(y).

Наблюдение	p	y	$(p-\bar{p})$	$(y-\bar{y})$	$(p-\bar{p})^2$	$(y-\bar{y})^2$
1	103,5	26,2	-39,86	-0,07	1588,82	0,01
2	127,0	24,8	-16,36	-1,47	267,65	2,16
3	126,0	25,6	-17,36	-0,67	301,37	0,45
4	124,8	26,8	-18,56	0,53	344,47	0,28
5	124,7	27,7	-18,66	1,43	348,20	2,05
6	121,6	28,3	-21,76	2,03	473,50	4,12
7	149,7	27,4	6,34	1,13	40,20	1,28
8	188,8	25,1	45,44	-1,17	2064,79	1,37
9	193,6	25,2	50,24	-1,07	2524,06	1,15
10	173,9	25,6	30,54	-0,67	932,69	0,45
Сумма	**1433,6**	**262,7**			**8885,75**	**13,30**
Среднее	**143,36**	**26,27**			**888,58**	**1,33**

В последних двух колонках таблицы можно найти, что Var(p) составляет 888,58 и Var(y) равна 1,33.

$$r = \frac{-16,24}{\sqrt{888,58 \times 1,33}} = \frac{-16,24}{34,38} = -0,47$$

- Из примера видим, что коэффициент корреляции незначительно отличается от нуля.
- Одна из причин в получении такого результата заключается в очень небольшом размере выборки.

- Еще одна причина - не **учтено влияние увеличения дохода на потребительский спрос в целом и на спрос на бензин в частности.**
- **Положительный эффект увеличения дохода в основном компенсировал отрицательный эффект роста цен, и, таким образом, спрос на бензин оставался стабильным**.

- Чтобы выделить эти два фактора используют коэффициент частной корреляции:

$$r_{xyz} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{[1 - r_{xz}^2][1 - r_{yz}^2]}}$$

где $\mathbf{r_{xy.z}}$ - коэффициент частной корреляции между **x** и **y** в случае постоянства воздействия величины **z**, а $\mathbf{r_{xy}}$, $\mathbf{r_{xz}}$ **и** $\mathbf{r_{yz}}$ - обычные коэффициенты корреляции между x и y, x и z, y и z соответственно.

- В примере со спросом на бензин можно вычислить корреляцию между ценой и располагаемым личным доходом и между спросом и доходом.
- Результаты по данной выборке составят соответственно 0,84 и 0,02.
- Подставим результаты в уравнение частной корреляции.

$$r = \frac{-0,47 - 0,84 \cdot 0,02}{\sqrt{(1-0,84^2)\cdot(1-0,02^2)}} = -0,91$$

Результат получился лучше