Научная школа по ЭММ

Наумов
Илья
Викторович

Введение

Существует три основных класса эконометрических моделей:

1. Модели временных рядов
2. Регрессионные модели с одним уравнением
3. Системы эконометрических уравнений

Модели временных рядов – представляют собой зависимость результативной переменной от переменной времени или переменных, относящихся к другим моментам времени:

- **модель тренда** (зависимость переменной Y от трендовой компоненты);
- **модель сезонности** (зависимость переменной Y от сезонной компоненты);
- **модель тренда и сезонности**.
- модели временных рядов, в которых результативная переменная Y зависит от переменных, датированных другими моментами времени:
  - **модели с распределенным лагом**, объясняющие изменение переменной Y в зависимости от предыдущих значений факторных переменных;
  - **модели авторегрессии**, объясняющие изменение переменной Y в зависимости от предыдущих значений результативных переменных;
  - **модели ожидания**, объясняющие изменение переменной Y в зависимости от будущих значений факторных или результативных переменных.

Регрессионные модели с одним уравнением, в которых зависимая переменная может быть представлена в виде функции факторных (независимых) переменных:

$$y = f(x1, x2, ..., xn, b1, b2, ..., bn),$$

По количеству факторных переменных регрессионные модели делятся на:

- **парные регрессии** (с одной факторной переменной);
- **множественные регрессии** (с двумя и более факторными переменными).

По виду функции f (x1, x2, ..., xn, b1, b2, ..., bn) регрессионные модели делятся на:

- **линейные и**
- **нелинейные регрессионные модели.**

Системы эконометрических уравнений:

- предназначены для исследования тех экономических процессов, которые невозможно описать одним уравнением регрессии.
- в этом случае строятся несколько эконометрических уравнений, которые в результате образуют систему.

Для решения эконометрической задачи необходимо последовательно выполнить несколько **этапов экономико-математического моделирования**

1. **Постановочный этап** - определяются конечные цели и задачи исследования, а также число включенных в модель факторных и результативных экономических переменных.

   Цели эконометрического исследования:
   - анализ изучаемого экономического процесса (явления, объекта);
   - прогноз экономических показателей, характеризующих изучаемый процесс;
   - моделирование поведения процесса при различных значениях факторных переменных;
   - формирование управленческих решений

   Количество переменных, включенных в модель:
   - не должно быть слишком большим
   - должно быть теоретически обоснованным
   - в модели должна отсутствовать функциональная или корреляционная связь между факторами.

2. **Априорный этап** – осуществляется теоретический анализ сущности изучаемого процесса

3. **Этап параметризации** – происходит выбор общего вида модели, а также определяется состав и формы формирующих ее связей.

   Основные задачи данного этапа:

   - выбор наиболее подходящего вида функциональной зависимости результативной переменной от факторных переменных (линейная или нелинейная).
   - спецификация модели:

– выявление связей и соотношений между параметрами модели;

– определение зависимых и независимых переменных;

– выражение исходных предпосылок и ограничений регрессионной модели.

4. **Информационный этап** – собирается требуемая статистическая информация и осуществляется анализ качества собранных данных.
5. **Этап идентификации модели** – проводится статистический анализ модели и происходит оценивание ее параметров.
6. **Этап оценки качества модели** – проверяются достоверность и адекватность модели реальному экономическому процессу.
7. **Этап интерпретации результатов моделирования**.

В рамках регрессионного анализа необходимо решить 4 задачи:

- Определение числовых значений параметров модели;
- Определение статистической достоверности параметров модели;
- Расчет и анализ показателей качества построенной регрессионной модели;
- Определение статистической достоверности построенной модели.

Эконометрика занимается:

- изучением количественных взаимосвязей экономических явлений и процессов,
- имеет дело со случайными событиями, которые характеризуются случайными величинами поскольку большинство взаимосвязей в экономике носит не детерминированный (строго определенный), а стохастический (вероятностный) характер.

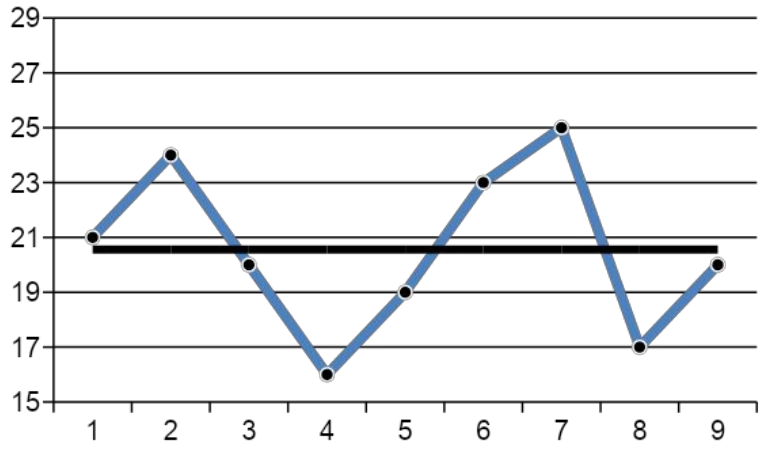
Каждая случайная величина оценивается числовыми характеристиками:

- Математическое ожидание
- Дисперсия
- Стандартное отклонение
- Вариация

Математическое ожидание:

- это среднее ожидаемое значение, принимаемое случайной величиной в больших сериях испытаний.
- оно используется в случаях, когда необходимо сравнить несколько альтернативных стратегий в однотипных ситуациях множество раз (при проведении больших серий испытаний).
- показывает какое значение случайная величина принимает «в среднем» (функция СРЗНАЧ в Excel).

$$\mu = MX = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n, \text{ или } MX = \sum_{t=1}^{n} x_t p_t$$



Дисперсия:

$$DX = \sum_{i=1}^{m} (x_i - MX)^2 \cdot p_i$$

- Используется для оценки разброса значений случайной величины вокруг ее среднего значения (математического ожидания).
- Это показатель степени, или мера отклонения случайной величины от ее математического ожидания, характеризующая вариативность значений случайной величины.
- Это показатель риска выбора случайной величины. Чем больше величина дисперсии случайной величины, тем выше риск в случае выбора именно этой альтернативы.
- если математическое ожидание может быть любым числом, даже отрицательным, то дисперсия всегда неотрицательна.
- Дисперсия не всегда удобна для анализа и оценки риска той или иной альтернативы из-за за высокой размерности (единицы измерения случайной величины в квадрате).
- Рассчитывается с помощью функции ДИСП в Excel.

Стандартное (среднеквадратичное) отклонение:

- Как и дисперсия используется в качестве меры абсолютного разброса случайной величины возле ее математического ожидания.
- используется для приведения размерности числовых характеристик к уровню размерности случайной величины.
- равно квадратному корню из дисперсии: $\sigma = \sqrt{DX}$
- Рассчитывается с помощью функции СТАНДТОТКЛОН в Excel.

Коэффициент вариации:

$$V = \frac{\text{СТАНДОТКЛОН(диапазон)}}{\text{СРЗНАЧ(диапазон)}}$$

- используется не для абсолютного, а относительного разброса.
- показывает какую долю среднего значения случайной величины составляет ее средний разброс.
- если коэффициент вариации менее 33%, то совокупность данных является однородной, если более 33%, то – неоднородной.

Чтобы совокупность случайных величин можно было использовать для регрессионного анализа и строить точные прогнозы, **необходимо, чтобы случайная составляющая была однородной и нормально распределена.**

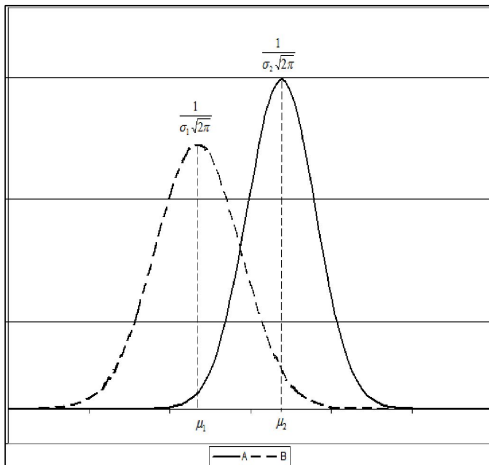
Это позволяет прогнозировать их поведение:

- проверять статистические гипотезы,
- строить интервальные оценки.

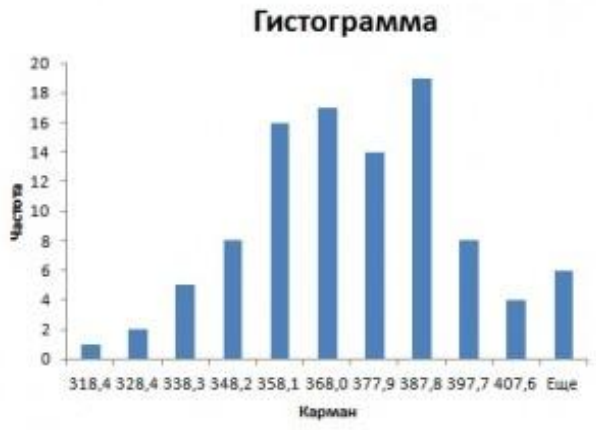
Нормальное распределение (распределение Гаусса) одной случайной величины X характеризуется лишь двумя параметрами:

- средним значением (математическим ожиданием μ)
- стандартным отклонением (σ).

График плотности вероятности нормального распределения имеет вид колокола:



- Максимум этой функции, а также центр симметрии находится в точке x=математическому ожиданию (μ) а «растянутость» вдоль оси X определяется параметром σ (среднеквадратическим отклонением)
- Чем больше значение Математического ожидания, тем правее расположен график при одинаковых значениях σ (μ2 > μ1).
- Чем меньше значение параметра СКО тем более острый и высокий максимум имеет плотность нормального распределения (σ2< σ1).
- Разброс среднего арифметического нормально распределенных случайных величин при неограниченном увеличении их числа стремится к нулю.



Такой график может быть получен только при бесконечно большом количестве измерений (при увеличении количества измерений приближается к графику нормального распределения Гаусса).

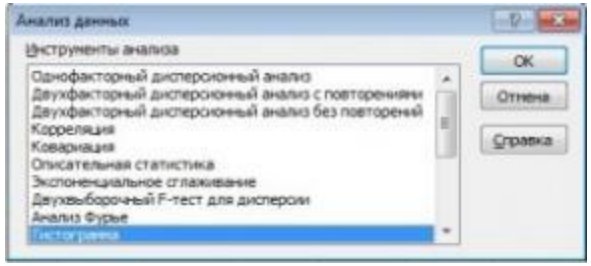
Например:

Построение гистограмм является очень быстрым способом проверки стабильности работы оборудования и добросовестности коллектива:

- если получим «кривую» гистограмму,
- значит, либо прибор не исправен или
- мы данные неверно собрали,
- либо кто-то где-то преднамеренно мухлюет или
- неверно использует оборудование.

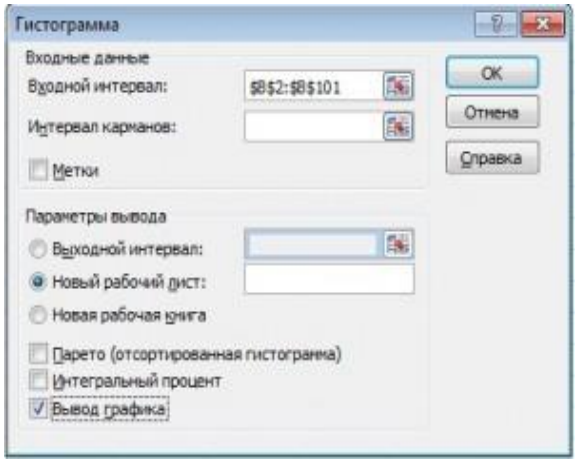
Построение гистограммы с помощью программы Excel.

1. Идем во вкладку «Анализ данных» и выбираем «Гистограмма».



2. Выбираем входной интервал.
   - Необходимо задать интервал карманов (т.е. те диапазоны, в пределах которых будут лежать наши значения).
   - Чем больше значений в интервале тем выше столбик гистограммы.
   - Если мы оставим поле «Интервалы карманов» пустым, то программа вычислит границы интервалов за нас.

3. Вывод графика - ставим соответствующую галочку напротив «Вывод графика».
   - Нажимаем «ОК».
   - Гистограмма готова.



4. Теперь нужно сделать так, чтобы по вертикальной оси отображалась не абсолютная частота, а относительная.
5. Под появившейся таблицей со столбцами «Карман» и «Частота» введем формулу «=СУММ» и сложим все абсолютные частоты.
6. К появившейся таблице со столбцами «Карман» и «Частота» добавим еще один столбец и назовем его «Относительная частота».
7. Во всех ячейках нового столбца введем формулу, которая будет рассчитывать относительную частоту:

100 * абсолютная частота /
/сумму, которую мы вычислили в п. 5.

Корреляция и ковариация

Важнейшая задача эконометрики – исследование существующих связей между социально-экономическими явлениями и процессами.

В процессе статистического исследования зависимостей:

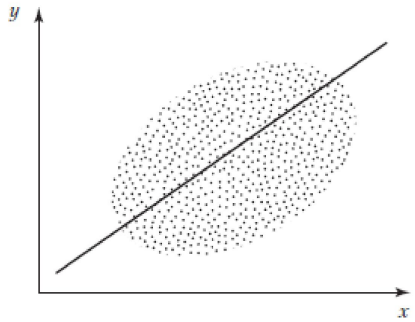
- вскрываются причинно-следственные связи между явлениями,
- это позволяет выявить факторы, оказывающие влияние на вариацию изучаемых явлений и процессов.

Причинно-следственные отношения — это связь явлений и процессов, при которой изменение одного из них (причины) ведет к изменению другого (следствия).

- Социально-экономические явления являются результатом одновременного воздействия большого числа факторов.
- Главной задачей эконометрики является нахождение основных причин и второстепенных.

Виды взаимосвязей между признаками, которые исследуются статистикой:

- **Функциональная** – зависимость, при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака.
- **Стохастическая** – проявляется не в каждом отдельном случае, а в общем, среднем при большом числе наблюдений.
- **Корреляционная** – является частным случаем стохастической связи, при которой изменение среднего значения результативного признака обусловлено изменением факторных признаков.



Корреляция — это статистическая зависимость между случайными величинами, не имеющая строго функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

Принято различать следующие **виды корреляции**:

1. **Парная** — связь между двумя признаками (результативным и факторным, или двумя факторными);
2. **Частная** — зависимость между результативным и одним факторным признаками при фиксированном значении других факторных признаков;
3. **Множественная** — зависимость результативного и двух или более факторных признаков, включенных в исследование.

Корреляционный метод анализа:

- используют для количественного определения тесноты и направления связи между:
  - двумя признаками (при парной связи) и
  - результативным и множеством факторных признаков(при многофакторной связи).
- Теснота связи количественно выражается величиной коэффициентов корреляции.
- Знаки при коэффициентах корреляции характеризуют направление связи между признаками.

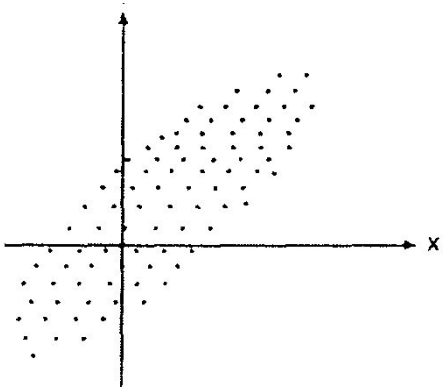
Ковариация выражает степень статистической зависимости между двумя множествами данных, измеряется в тех же единицах что и переменные:

$$Cov(X,Y) = \frac{1}{m}\sum_{i=1}^{m}(X_i - M(X))(Y_i - M(Y))$$

где X, Y - множества значений случайных величин размерности m;
M(X) - математическое ожидание случайной величины X;
M(Y) - математическое ожидание случайной величины Y.

Ковариация:

- характеризует связь двух переменных,
- дает количественную характеристику диаграммы рассеивания:



– По облаку рассеивания можно судить о связи переменных.

– Чем связь больше, тем более вытянуто облако.

Оценка связи по ковариации:

1. **Положительная ковариация** наблюдается когда большим значениям случайной величины X соответствуют большие значения случайной величины Y (между ними существует тесная прямая взаимосвязь).
2. **Отрицательная ковариация** наблюдается когда малым значениям случайной величины X соответствуют большие значений случайной величины Y.
3. **Показатель ковариации близок к нулю** при слабо выраженной зависимости.

Значение ковариации зависит не только от “тесноты” связи случайных величин, но **и от самих значений этих величин** (например, от единиц измерения этих значений).

Для исключения этой зависимости вместо ковариации используется **безразмерный коэффициент корреляции R -** отношение полученной ковариации к максимально возможной:

$$R = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Коэффициент корреляции принимает значения от -1 до +1 :

1. **Если R<0**, то связь между изучаемыми показателями xt и yt является **обратной**, (с увеличением xt значение yt уменьшается, и наоборот);
2. **Если R>0**, то связь между изучаемыми показателями xt и yt является **прямой** (с увеличением xt значение yt увеличивается);
3. **Если R=0**, то линейная связь между изучаемыми показателями xt и yt **отсутствует**;
4. **Если R близок к нулю**, то может присутствовать нелинейная связь переменных, либо зависимость вообще отсутствует.
5. **Если R=1 (-1)**, то линейная связь между изучаемыми показателями xt и yt является **строго функциональной** (изменение факторного признака xt определяет изменение результативного признака yt).

$0 < \lvert R \rvert < 0{,}3$	**Связь слабая**
$0{,}3 \le \lvert R \rvert < 0{,}7$	**Связь средняя**
$0{,}7 \le \lvert R \rvert < 1$	**Связь тесная**

Парные регрессионные модели

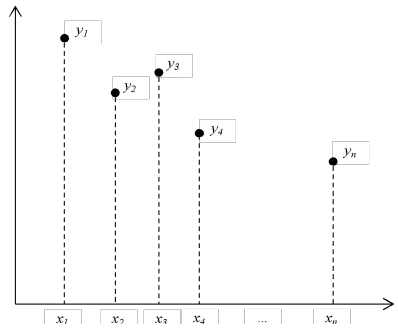
Модели парной регрессии

В регрессионной модели **все переменные делятся на**:

- зависимые, эндогенные (y) и
- независимые, экзогенные переменные-факторы (x).

Регрессионный анализ:

- предназначен для количественного измерения выявленной связи между этими переменными,
- уточнения выводов самого качественного анализа.



Анализ начинается с установления вида зависимости между x и y:

- необходимо найти такой вид уравнения регрессии, который наилучшим образом соответствует характеру изучаемой связи.
- от вида изучаемой связи между переменными зависит тип формируемой модели (линейный или нелинейный).
- самый простой способ определения вида связи между показателями – визуальный – для этого **строится корреляционное поле**.

Если на поле между точками можно провести прямую линию, то для моделирования связи **можно использовать линейную зависимость:**

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Существует несколько причин появления в модели случайной составляющей:

1. **Не включение объясняющих переменных**.
   - Соотношение между y_t и x_t является упрощением.
   - В действительности существуют другие факторы, влияющие на y_t, которые не учтены в модели $y_t = \alpha + \beta x_t + e_t$, их суммарное влияние представлено в уравнении случайной составляющей $\boldsymbol{e_t}$.
   - Часто возникает ситуация, когда имеются переменные, которые мы хотели бы включить в регрессионное уравнение, но не можем этого сделать потому, что не знаем, как их измерить.
   - Возможно, существуют также другие факторы, которые мы можем измерить, но которые оказывают такое слабое влияние, что их не стоит учитывать.
   - Могут существовать факторы, которые являются существенными, но которые мы из-за отсутствия опыта таковыми не считаем.

2. **Агрегирование переменных.**
   - Во многих случаях рассматриваемая зависимость – это попытка объединить вместе некоторое число экономических соотношений.
   - Однако отдельные соотношения имеют различные параметры, в результате, любая попытка определить точное соотношение между зависимой и независимыми переменными является лишь аппроксимацией.
   - Наблюдаемое расхождение при этом приписывается наличию случайной составляющей.

3. **Неправильная функциональная спецификация.**
   - соотношение между yt и xt математически может быть определено неверно.
   - истинная зависимость может являться не линейной, а более сложной.
   - любая самая изощренная формула является лишь приближением, и существующее расхождение также вносит вклад в случайную составляющую.

4. **Ошибки измерения** – если в измерении переменных имеются (статистические) ошибки, то наблюдаемые значения не будут соответствовать точному соотношению, и существующее расхождение будет вносить вклад в случайную составляющую.

Случайная составляющая – это суммарное проявление всех перечисленных причин.

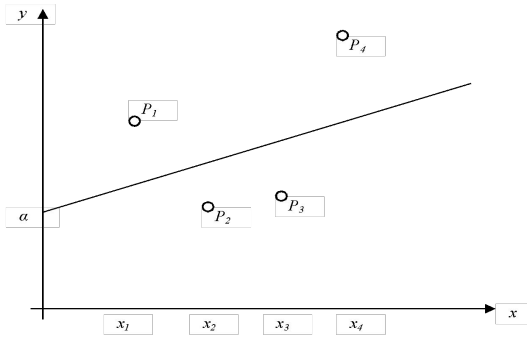
- Чем меньше ее значения, тем точнее оценки коэффициентов α и β.
- Если бы случайных ошибок не было, мы бы смогли точнее измерить влияние xt на yt.
- Однако в действительности каждое изменение yt отчасти вызвано изменением случайной ошибки et, и это значительно усложняет исследования.
- По этой причине et иногда интерпретируется как шум.

Решение первой задачи регрессионного анализа – поиск коэффициентов регрессии

- Предположим, что у нас имеется n наблюдений для xt и yt,
- Имеющиеся переменные имеют линейную динамику,
- Необходимо определить значения α и β в уравнении $y_t = \alpha + \beta x_t + e_t$. поскольку именно эти коэффициенты однозначно и полностью определяют положение прямой на плоскости.
- Для поиска значений a и b, являющихся оценками истинных параметров α и β, используется метод наименьших квадратов.

Особенности применения метода наименьших квадратов.

- Допустим, у нас имеется 4 наблюдения для x и y,
- Они представлены на графике,
- Необходимо определить значения коэффициентов a и b.



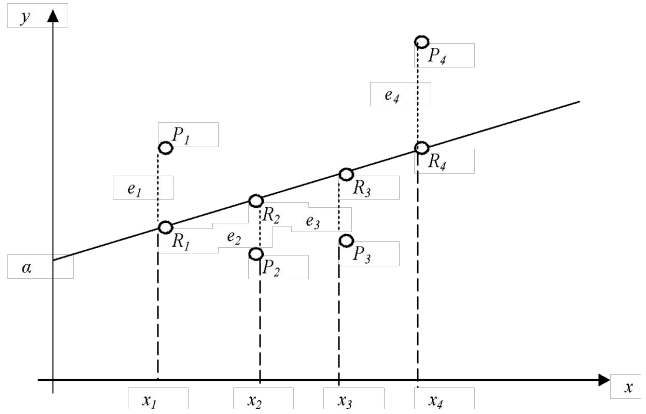
- Это можно сделать очень приблизительно, отложив 4 точки P и построив прямую, соответствующую этим точкам – **линию регрессии:**
  - отрезок прямой на оси y, представляет собой оценку α и обозначен a,
  - угловой коэффициент прямой – оценка β и обозначен b.

Недостатки такого подхода:

- Построение линии регрессии без точных расчетов является достаточно субъективным.
- Более того, если переменная зависит не от одной или двух, а от большего количества независимых переменных, это просто невозможно.
- Через корреляционное поле можно провести бесконечное множество прямых линий.
- Определить какая из них наилучшим образом согласуются с реальными данными сложно

Алгоритм нахождения параметры регрессии:

1. ***Первый этап – определение остатка для каждого наблюдения.***
   - Построенная линия регрессии в нашем случае не совпадает с точками наблюдения.
   - В результате в каждом наблюдении формируются отклонения от прямой (остатки).
   - Для наблюдений остатки обозначены как e1, e2, e3 и e4.



   - В идеальном случае линия регрессии должна быть построена таким образом, чтобы эти остатки были минимальными.
   - Сделать это достаточно сложно, так как линия, строго соответствующая одним наблюдениям, не будет соответствовать другим, и наоборот.

2. ***Второй этап – Необходимо выбрать какой-то критерий подбора, который будет одновременно учитывать величину всех остатков.***

- Один из способов решения поставленной проблемы состоит в минимизации суммы квадратов остатков S: $S = e_1^2 + e_2^2 + e_3^2 + e_4^2$
- Этот метод оценивания параметров называется методом наименьших квадратов (МНК). Его суть заключается в том, чтобы сумма квадратов отклонений фактических значений зависимой переменной от найденных по уравнению регрессии была наименьшей.
- Величина S будет зависеть от выбора a и b, так как они определяют положение линии регрессии:
  - В соответствии с этим критерием, чем меньше S, тем строже соответствие.
  - Если S=0, то получено абсолютно точное соответствие, так как это означает, что все остатки равны нулю.
  - В этом случае линия регрессии будет проходить через все точки, однако, это невозможно из-за наличия случайной составляющей.
- Таким образом, мы стремимся найти такие a и b, чтобы значение S было минимальным.

3. ***Третий этап – Нахождение параметров уравнения регрессии методом наименьших квадратов:***

- Минимизируется сумма квадратов отклонений фактических значений результативного признака от теоретических, полученных по выбранному уравнению регрессии:

$$S = \sum (y_i - y_x)^2 \to \min.$$

- Система нормальных уравнений для нахождения параметров линейной парной регрессии методом наименьших квадратов имеет следующий вид:

$$\begin{cases} na_0 + a_1 \sum x_i = \sum y_i; \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i, \end{cases}$$

где n — объем исследуемой совокупности (число единиц наблюдения).

a0 – показывает усредненное влияние неучтенных факторов на результативный признак.

a1 – показывает, насколько в среднем изменяется значение результативного признака при изменении факторного признака на единицу собственного измерения.

Например:

Субъект	Полная учетная стоимость основных фондов (*yi*), трлн руб.	Инвестиции в основной капитал (*xi*), млрд руб.
Белгородская область	586,0	51,4
Брянская область	357,8	19,9
Владимирская область	369,2	36,6
Воронежская область	738,6	61,3
Ивановская область	299,8	20,4
Калужская область	383,1	49,3
Костромская область	279,9	8,8
Курская область	399,6	30,6
Липецкая область	579,3	62,1
Орловская область	237,8	13,2
Рязанская область	515,7	27,3
Смоленская область	441,0	22,6
Тамбовская область	418,3	24,3
Тверская область	663,8	51,5
Тульская область	492,1	46,1
Ярославская область	770,9	43,0
Итого	7532,9	568,4

Предположим наличие линейной зависимости между рассматриваемыми переменными.

$$\begin{cases} na_0 + a_1 \sum x_i = \sum y_i; \\ a_0 \sum x_i + a_1 \sum x_i^2 = \sum x_i y_i, \end{cases} \qquad \begin{cases} 16a_0 + 568{,}4a_1 = 7532{,}9; \\ 568{,}4a_0 + 24\,499{,}32a_1 = 299\,066{,}75. \end{cases}$$

Отсюда получается:

- a0 = 211,296
- a1 = 7,305
- Y = 211,296 + 7,305 * X

Коэффициент регрессии a1 = 7,305 означает, что при увеличении инвестиций в основной капитал на 1 млрд руб. полная учетная стоимость основных фондов субъектов возрастет в среднем на 7,305 трлн руб.

Другой способ нахождения коэффициентов регрессии:

Таблица 2.
Промежуточные расчеты для получения числовых значений коэффициентов регрессии a и b

x	y	$x \cdot y$	x^2
1	3	3	1
2	5	10	4
3	6	18	9
сумма		31	14
сумма / n		10,34	4,67

$$\bar{x} = (1+2+3)/3 = 2; \qquad \bar{x}^2 = 4; \qquad \bar{y} = (3+5+6)/3 = 4{,}67;$$

$$b = \frac{\overline{yx} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{10{,}34 - 4{,}67 \cdot 2}{4{,}67 - 4} = \frac{1}{0{,}67} = 1{,}5;$$

$$a = \bar{y} - b\bar{x} = 4{,}67 - 1{,}5 \cdot 2 = 4{,}67 - 3 = 1{,}67.$$

Рассчитанные коэффициенты позволяют записать уравнение регрессии:

$$\hat{y} = 1{,}67 + 1{,}5x \quad (7).$$

Экономико-математическая интерпретация построенной регрессионной модели

После записи уравнения регрессии необходимо выполнить экономико-математическую интерпретацию полученной модели: **y = a + b*x .**

- Формально **коэффициент регрессии «a»** дает прогнозируемое значение **«y»** при нулевом значении **«x»**.
- Однако в экономических задачах показатель **«x»** редко принимает нулевое значение и буквальная интерпретация может привести к неверным результатам.
- Поэтому в процессе интерпретации модели основное внимание следует уделять не величине, а **знаку коэффициента «a»**, который здесь определяет относительную скорость изменения показателей, включенных в модель.
  - Если a > 0, то относительное изменение **«x»** происходит быстрее, чем изменение **«y»**.
  - Если a < 0, то относительное изменение **«y»** происходит быстрее, чем изменение **«x»**.

- Если величина показателя **«x»** увеличилась на 1 единицу, тогда уравнение изменяется следующим образом: $y = a + b*(x + 1) = a + b*x + b$.
- То есть, увеличение **«x»** на 1 единицу приводит к изменению зависимой переменной **«y»** на величину **«b»**.
- Важную роль в интерпретации коэффициента **«b»** играет его знак.
  - Если b > 0, с ростом **«x»** растет **«y»**, и **связь между показателями является прямой**.
  - Если b < 0, с ростом **«x»** величина **«y»** падает, и **связь между показателями является обратной.**

Решение второй задачи регрессионного анализа – Проверка статистической достоверности параметров построенной модели

Математически параметры **a** и **b** можно рассчитать для любого массива статистической информации, однако необходимо проверить, можно ли доверять найденным значениям:

- Исследователем выдвигается гипотеза о том, что две сравниваемые совокупности не отличаются (нулевая гипотеза, или нуль-гипотеза).
- При этом предполагается, что различие сравниваемых величин равно нулю, а выявленное по данным выборки отличие от нуля носит случайный характер.
- Нулевая гипотеза отвергается тогда, когда получается результат, который маловероятен.
- Границей маловероятного обычно считают значение 0,05 (5%).

Алгоритм проверки статистической гипотезы о достоверности параметра b:

1. **Выдвигается нулевая гипотеза** H0 (b): b = 0,
   - согласно которой при неограниченном увеличении объема статистической информации коэффициент **b** будет **= 0**,
   - а при анализе имеющегося ограниченного набора статистических данных **получится не равным нулю**;
2. **Необходимо определить, существенно ли найденное значение параметра b отличается от нуля.**
   - В качестве базиса для проверки используются имеющиеся статистические данные.
   - Для этого необходимо ввести такую переменную, по значению которой можно было бы судить о справедливости нулевой гипотезы.
   - Такой переменной является **статистика Стьюдента**, обозначаемая **t**:

$$t = \frac{b}{\sigma_b}.$$

   - Статистика – это случайная переменная, распределение вероятностей которой лежит в основе проверки выполнения различных гипотез,
   - Статистика Стьюдента имеет так называемое t-распределение, которое стремится к нормальному при увеличении объема статистических данных;

3. **По таблице распределения Стьюдента определяется критическое значение t-статистики для оцениваемого коэффициента регрессии.**

Распределение Стьюдента, t-распределение с k степенями свободы:

α	0.200	0.150	0.100	0.050	0.025	0.020	0.010	0.005
t_5	0.92	1.16	1.48	2.02	2.57	2.76	3.36	4.03
t_6	0.91	1.13	1.44	1.94	2.45	2.61	3.14	3.71
t_7	0.90	1.12	1.41	1.89	2.36	2.52	3.00	3.50
t_8	0.89	1.11	1.40	1.86	2.31	2.45	2.90	3.36
t_9	0.88	1.10	1.38	1.83	2.26	2.40	2.82	3.25
t_{10}	0.88	1.09	1.37	1.81	2.23	2.36	2.76	3.17
t_{13}	0.87	1.08	1.35	1.77	2.16	2.28	2.65	3.01
t_{14}	0.87	1.08	1.35	1.76	2.14	2.26	2.62	2.98
t_{15}	0.87	1.07	1.34	1.75	2.13	2.25	2.60	2.95
t_{16}	0.86	1.07	1.34	1.75	2.12	2.24	2.58	2.92
t_{20}	0.86	1.06	1.33	1.72	2.09	2.20	2.53	2.85
t_{21}	0.86	1.06	1.32	1.72	2.08	2.19	2.52	2.83
t_{22}	0.86	1.06	1.32	1.72	2.07	2.18	2.51	2.82
t_{23}	0.86	1.06	1.32	1.71	2.07	2.18	2.50	2.81
t_{24}	0.86	1.06	1.32	1.71	2.06	2.17	2.49	2.80
t_{25}	0.86	1.06	1.32	1.71	2.06	2.17	2.49	2.79
t_{26}	0.86	1.06	1.31	1.71	2.06	2.16	2.48	2.78
t_{27}	0.86	1.06	1.31	1.70	2.05	2.16	2.47	2.77
t_{28}	0.85	1.06	1.31	1.70	2.05	2.15	2.47	2.76
t_{29}	0.85	1.06	1.31	1.70	2.05	2.15	2.46	2.76

t крит = (n-1; α/2)

- Если значение анализируемого коэффициента регрессии по модулю больше значения t-статистики для него, то нулевая гипотеза отвергается.
- В противном случае нулевую гипотезу отвергнуть нельзя.
- Это не означает, что мы ее принимаем, мы только не можем ее отвергнуть и, следовательно, нужны дополнительные исследования

4. **В большинстве случаев определяется не только величина статистики Стьюдента, но и вероятность выполнения нулевой гипотезы.**
   - Вероятность выполнения нулевой гипотезы для соответствующего коэффициента регрессии определяется с помощью **P-Значения**:
     - Нулевая гипотеза отвергается, если вероятность ее выполнения < 5%.
     - Если данная вероятность =>5%, нуль-гипотезу отвергнуть нельзя и, следовательно, между **xt** и **yt** нет линейной связи.
5. **Аналогично проверяется выполнение нулевой гипотезы для параметра a.**
   - Если нулевую гипотезу для параметра a нельзя отвергнуть, то коэффициент a признается не достоверным, а зависимость между **xt** и **yt** превращается в простую пропорциональную зависимость.
5. **Для оценки параметров α и β не всегда достаточно точечного анализа.**
6. **Важно определить, в какой интервал в 95% будут попадать истинные значения параметров α и β при изменении набора данных.**
   - Зная табличное значение статистики Стьюдента (t табл.), можно определить границы искомых интервалов.

$$\text{Для прам етра } \alpha\text{: } [a - t_{табл} \cdot \sigma_a \, , a + t_{табла}\text{: } \sigma_a].$$

$$\text{Для прам етра } \beta\text{: } [b - t_{табл} \cdot \sigma_b \, , b + t_{табл} \cdot \sigma_b].$$

   - Записанные интервалы называются доверительными интервалами с 95%-м уровнем доверия.

Например:

По 25 наблюдениям получено уравнение регрессии:

$$\underset{(se)}{\hat{y}_i} = \underset{(1.4)}{1.2} - \underset{(0.2)}{2.8}x_i + \underset{(3.0)}{4.5}z_i, \quad R^2 = 0.77$$

Необходимо проверить значимость коэффициента при переменной zi на уровне значимости α = 0.05

Решение:

- Для расчета t статистики используем формулу: $t = \frac{b}{\sigma_b}$.
- В результате: t расч = 4,5 / 3 = 1,5
- По таблице распределения Стьюдента
  t крит = (n-1; α/2) = t (25-1; 0,05/2) = t (24; 0,025) = 2,06
- Поскольку t расч < t крит (1,5 < 2,06) то на уровне значимости 5% нулевая гипотеза не отвергается, то есть коэффициент при переменной zi не значим (=0) с надежностью 95%.

Решение третьей задачи регрессионного анализа – расчет и оценка показателей качества построенной регрессионной модели

Мы предположили, что показатели **xt** и **yt** связаны между собой линейной связью, нашли параметры **a** и **b**, оценили их статистическую значимость.

1. Необходимо установить, насколько эта связь является тесной.

В качестве меры степени тесноты линейной связи переменных используется коэффициент корреляции R:

Если на уровне теоретического исследования связь между показателями установлена, но при этом значение коэффициента корреляции $R < 0,7$ то необходимо:

- удалить из анализируемой статистики статистические выбросы
- добавить в регрессионную модель новые наблюдения или факторы, поскольку результирующий показатель **yt** может реально зависеть не только от **xt**, но и от других факторов;
- перейти к нелинейной регрессионной модели, т.к. экономические процессы не могут быть адекватно описаны линейной моделью.

2. Необходимо установить уровень подгонки модели к исходным данным (рассчитать коэффициент детерминации)

- Исходя из этого квадрат полной вариации равен сумме квадратов вариации вследствие регрессии yt на xt (**RSS**) и квадратов остатков (**ESS**):

TSS = RSS + ESS

где, TSS – общая дисперсия регрессионной модели
RSS – дисперсия, объясненная регрессией
ESS – остаточная дисперсия

– Если в модели остатки минимальны (а это основополагающий принцип метода МНК) то связь между показателями считается функциональной:

$$\min \frac{ESS}{TSS} = 0 \Rightarrow \frac{RSS}{TSS} = 1 \Rightarrow RSS = TSS$$

– Если в регрессионной модели остатки максимальны, то размер дисперсии, объясненной регрессией стремится к нулю:

$$\max \frac{ESS}{TSS} = 1 \Rightarrow \frac{RSS}{TSS} = 0 \Rightarrow RSS = 0$$

- Доля дисперсии, объясненная регрессией (RSS) – Коэффициент детерминации показывает:
  - какая доля вариации зависимой переменной может быть объяснена уравнением регрессии.
  - долю разброса данных, объясненного регрессионной моделью;
  - долю наблюдений, попавших под описание регрессионной модели.

$$R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS},$$

$$0 \le R^2 \le 1.$$

  - Если предположить, что вся вариация в **yt** полностью определяется случайными возмущениями и не связана с изменением **xt**, тогда RSS=0,
  - В результате ESS=TSS, то есть R2 = 0.

Коэффициент детерминации показывает качество «подгонки» регрессионной модели к значениям yt, однако полагаться только на этот коэффициент нельзя, поскольку:

- Коэффициент детерминации возрастает при добавлении еще одного фактора;
- Он изменяется даже в результате простейшего преобразования зависимой переменной.
- Если взять число факторов, равное количеству наблюдений, всегда можно добиться, чтобы величина коэффициента детерминации равнялась единице.

- Для устранения эффекта, связанного с ростом коэффициента детерминации при увеличении количества факторов **используется нормированный коэффициент детерминации**.

$$R_{норм} = 1 - \frac{ESS/(n-k)}{TSS/(n-1)} \qquad R_{норм} = 1 - (1 - R^2) \cdot \frac{n-1}{n-k}$$

Основные свойства уточненного коэффициента детерминации Rнорм:

- Rнорм $\leq$ R2;
- Rнорм ≤ 1,
- В некоторых случаях может быть отрицательным.

Уточнённый коэффициент детерминации:

- используется для сравнения регрессий при изменении количества переменных.
- показывает, какая доля общей дисперсии объясняется факторами, включенными в регрессионную модель.

3. Необходимо проанализировать выбросы в модели

Статистический выброс – это аномальное наблюдение, для которого реальное значение результирующего показателя **yt** резко отклоняется от линии регрессии. Наблюдение является статистическим выбросом, его стандартный остаток по абсолютной величине больше или равен 2.

- Выбросы удаляются если коэффициент корреляции меньше 0,7
- количество удаляемых наблюдений не должно превышать 1/8 общего объема данных.
- при регрессионном анализе динамических рядов не следует удалять последнее наблюдение.

- если последующие наблюдения не приближаются к линии регрессии, то можно сделать вывод о том, что изучаемый процесс вследствие каких-либо причин стал развиваться по иному закону
- поэтому, построенную регрессионную модель нельзя использовать для его дальнейшего исследования.

Решение четвертой задачи регрессионного анализа – определение статистической достоверности построенной модели

Величина, с помощью которой проверяется нулевая гипотеза для коэффициента детерминации, называется **статистикой Фишера**. Для ее расчета отношение RSS / TSS преобразуется с учетом соответствующих степеней свободы:

$$F = \frac{RSS/(k-1)}{ESS/(n-k)}.$$

$$Fрасч = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

$$F = \frac{SS^1_{ost} - SS^2_{ost}}{1} : \frac{SS^2_{ost}}{n-m-1}$$

Величина F подчиняется F-распределению Фишера. Зная его можно рассчитанную статистику Фишера сравнить с табличным значением.

- **Если F табличное < F фактическое**, **то нулевая гипотеза для коэффициента детерминации отвергается**, т.е., вариация **yt** обусловлена не только случайными возмущениями, но и вариацией **xt**.
- **Если Fтабличное > Fфактическое**, **то нулевую гипотезу для коэффициента детерминации отвергнуть нельзя**. Это не означает, что **xt** не влияет на **yt**, просто на анализируемых статистических данных это влияние установить не удалось.

Случайное превышение табличного значения маловероятно.

F крит = F(k-1; n-k)

	$n=5$	$n=10$	$n=19$	$n=20$	$n=21$	$n=22$	$n=23$	$n=21$	$n=25$
$k=1$	6.61	4.96	4.38	4.35	4.32	4.30	4.28	4.26	4.24
$k=2$	5.79	4.10	3.52	3.49	3.47	3.44	3.42	3.40	3.39
$k=3$	5.41	3.71	3.13	3.10	3.07	3.05	3.03	3.01	2.99
$k=4$	5.19	3.48	2.90	2.87	2.84	2.82	2.80	2.78	2.76
$k=5$	5.05	3.33	2.74	2.71	2.68	2.66	2.64	2.62	2.60
$k=10$	4.74	2.98	2.38	2.35	2.32	2.30	2.27	2.25	2.24
$k=11$	4.70	2.94	2.34	2.31	2.28	2.26	2.24	2.22	2.20
$k=12$	4.68	2.91	2.31	2.28	2.25	2.23	2.20	2.18	2.16
$k=13$	4.66	2.89	2.28	2.25	2.22	2.20	2.18	2.15	2.14
$k=14$	4.64	2.86	2.26	2.22	2.20	2.17	2.15	2.13	2.11
$k=15$	4.62	2.85	2.23	2.20	2.18	2.15	2.13	2.11	2.09
$k=16$	4.60	2.83	2.21	2.18	2.16	2.13	2.11	2.09	2.07
$k=17$	4.59	2.81	2.20	2.17	2.14	2.11	2.09	2.07	2.05
$k=18$	4.58	2.80	2.18	2.15	2.12	2.10	2.08	2.05	2.04
$k=19$	4.57	2.79	2.17	2.14	2.11	2.08	2.06	2.04	2.02
$k=20$	4.56	2.77	2.16	2.12	2.10	2.07	2.05	2.03	2.01
$k=21$	4.55	2.76	2.14	2.11	2.08	2.06	2.04	2.01	2.00
$k=22$	4.54	2.75	2.13	2.10	2.07	2.05	2.02	2.00	1.98
$k=23$	4.53	2.75	2.12	2.09	2.06	2.04	2.01	1.99	1.97
$k=24$	4.53	2.74	2.11	2.08	2.05	2.03	2.01	1.98	1.96
$k=25$	4.52	2.73	2.11	2.07	2.05	2.02	2.00	1.97	1.96
$k-30$	4.50	2.70	2.07	2.04	2.01	1.98	1.96	1.94	1.92
$k=31$	4.49	2.69	2.07	2.03	2.00	1.98	1.95	1.93	1.91
$k=32$	4.49	2.69	2.06	2.03	2.00	1.97	1.95	1.93	1.91
$k=33$	4.48	2.69	2.06	2.02	1.99	1.97	1.94	1.92	1.90
$k=34$	4.48	2.68	2.05	2.02	1.99	1.96	1.94	1.92	1.90
$k=35$	4.48	2.68	2.05	2.01	1.98	1.96	1.93	1.91	1.89

По распределению Фишера определяют вероятность нулевой гипотезы для коэффициента детерминации:

1. Сначала выдвигается нуль-гипотеза, согласно которой R2=0, а его расчетное значение отлично от нуля из-за ограниченности имеющегося набора статистических данных;
2. Затем определяется статистика Фишера, имеющая F-распределение;
3. По распределению статистики Фишера рассчитывается вероятность выполнения нулевой гипотезы:
   - если вероятность больше или равна 5%, то:
     - нулевую гипотезу отвергнуть нельзя,
     - установленная линейная связь между xt и yt не является статистически достоверной,
     - необходимо увеличить количество наблюдений;
   - если вероятность меньше 5%, то:
     - нулевая гипотеза отвергается на 95%-м уровне значимости,
     - найденному значению коэффициента детерминации можно доверять,
     - размер используемой выборки признается достаточным.

Например:

По 25 наблюдениям получено уравнение: $\underset{(se)}{\hat{y}_i} = \underset{(1.4)}{1.2} - \underset{(0.2)}{2.8}x_i + \underset{(3.0)}{4.5}z_i, \quad R^2 = 0.77$

Необходимо проверить гипотезу о значимости регрессии на уровне значимости α=0.05.

Решение:

Для расчета Fрасч используем формулу $F_{расч} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$

Поскольку k=3 (у нас 3 коэффициента регрессии, то есть 3 степени свободы)

То по таблице распределения Фишера

$$F_{крит} = F(k-1; n-k) = F(2; 22) = 3{,}44$$

Так как Fрасч >Fкрит (36,83 > 3,44), то на уровне значимости 5% нулевая гипотеза отвергается.

Следовательно, с надежностью 95% регрессия значима.

Множественные регрессионные модели

Модели множественной линейной регрессии

- строятся когда величина исследуемого показателя складывается под влиянием не одного, а многих различных факторов,
- каждый из факторов в отдельности может не оказывать решающего воздействия.
- используются для измерения совместного влияния ряда показателей факторов на величину анализируемого показателя.

Основная цель множественной регрессии – построение модели с большим числом факторов. При этом:

- Необходимо определить влияние каждого фактора в отдельности на результирующий показатель, а также в совокупности.
- Выбор факторов производится исходя из экономического анализа и связан с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями.
- Факторы, включаемые в модель должны быть количественно измеримы и не должны коррелировать между собой.
- Для получения надежных оценок в модель не следует включать слишком много факторов (их число не должно превышать 1/3 объема имеющихся данных).

В таких моделях зависимая переменная **y** рассматривается как функция не одной, а нескольких независимых переменных **xt**:

$$y_t = a + b_1 x_{t1} + b_2 x_{t2} + \ldots + b_m x_{tm} + e_t.$$

Множественный регрессионный анализ выполняется аналогично парной линейной регрессии, однако:

- в качестве независимой (экзогенной) переменной выбран не один, а несколько факторов.
- при выделении входного интервала X, помечаются столбцы значений всех независимых переменных вместе с названиями.
- по величине P-значений определяется вероятность отсутствия влияния каждого введенного в модель фактора a, b1, b2, ..., bm на зависимую переменную:
  - Если величина P-Значения для фактора больше или равна 5%, то фактор исключается из модели.
  - Если факторов, имеющих высокое P-значение несколько, то их исключение проводится последовательно.
  - В первую очередь удаляется фактор, имеющий наибольшее P-Значение, после чего процедура регрессионного анализа проводится заново, на оставшихся факторах.

Оценка качества модели:

1. **Связь между изучаемыми факторами и зависимой переменной должна быть тесной**:
   - коэффициент корреляции (Множественный R) должен быть ≥ 0,7;
   - если он меньше 0,7 значит необходимо удалить выбросы
   - если удаление выбросов не помогает улучшить тесноту связи, значит необходимо добавить новые наблюдения.
2. **Регрессионная модель в целом должна быть достоверна:**
   - количество наблюдений должно быть достаточным, т.е. величина Значимость F должна быть < 5%;
   - Отсюда делаем вывод о том, что наблюдений достаточно или нет для построения регрессионной модели
3. **Коэффициенты модели, определяющие меру влияния факторов на результат, должны быть достоверными**:
   - все P-значения должны быть < 5%;
   - Отсюда делаем вывод: ВЛИЯЮТ либо НЕ ВЛИЯЮТ факторы на зависимую переменную
4. **Результаты регрессионного анализа не должны содержать статистических выбросов, которые могут быть удалены.**

Нелинейные регрессионные модели

Модели нелинейной регрессии

- Соотношения, существующие между социально-экономическими показателями и процессами не всегда описываются линейными функциями,
- Зачастую для моделирования используют нелинейную (по независимой переменной) регрессию.
- В случае неправильного выбора типа регрессионной модели могут возникать большие ошибки.

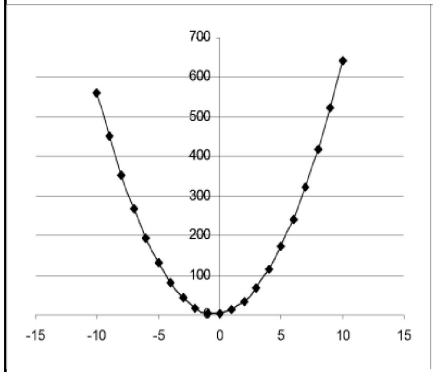
Основные этапы нелинейного моделирования:

1. **Этап спецификации модели – определяют вид уравнения регрессии:**
   - для этого используется опыт предыдущих исследований,
   - визуальное наблюдение расположения точек корреляционного поля.
   - строится графики динамики всех показателей, используемых в моделировании, для того чтобы определить какие переменные необходимо преобразовывать.
   - Среди множества моделей нелинейной регрессии можно выделить два вида:
     - модели, нелинейные относительно независимых переменных, но линейные относительно параметров регрессии, и
     - модели, нелинейные как относительно переменных, так и относительно параметров.

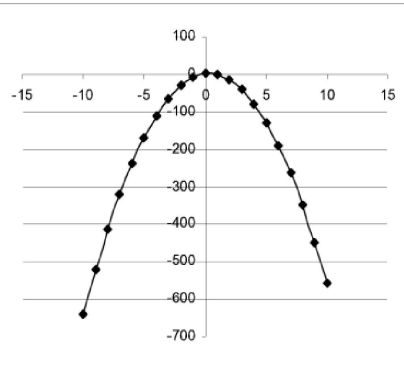
2. Этап линеаризации – преобразования переменных к линейному виду.
3. Этап линеаризации – это переход от нелинейной связи (гиперболической, показательной, степенной, логарифмической и т.п.) к линейной.
4. Этап регрессионного анализа
5. Этап оценки качества модели
6. Этап обратного преобразования переменных модели к нелинейному виду.

Основные виды преобразования нелинейных моделей в линейные

Связь квадратичная: $y = a + b_1x + b_2x^2$



$b_2 \geq 0$



$b_2 \leq 0$

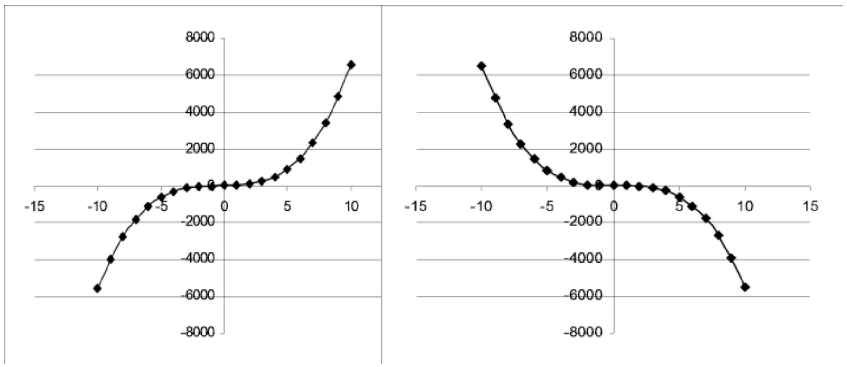
- Модель не линейна относительно независимой переменной **x**
- В этом случае линеаризация выполняется с помощью замены переменной:

$$x1 = x2.$$

- В результате исходное уравнение принимает вид:

$$y = a + b1x + b2x1$$

Связь кубическая: $y = a + b_1x + b_2x^2 + b_3x^3$

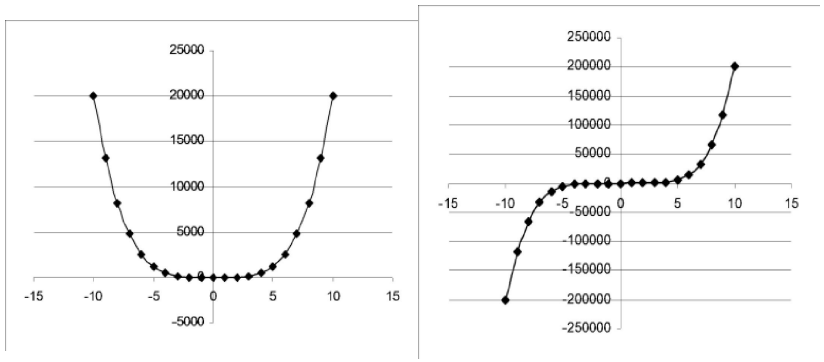


b3 ≥ 0 b3 ≤ 0

- Модель не линейна относительно независимых переменных x.
- В этом случае линеаризация выполняется с помощью замен переменных: $x_1=x^2$ и $x_2=x^3$.
- В результате исходное уравнение принимает вид:

$$y = a + b_1x + b_2x_1 + b_3x_2$$

Связь степенная: $y = a * x^b$ (b≥2 и целое)



b – четкое b – нечеткое

- Модель не линейна относительно параметра – коэффициента b. Логарифмируя, получаем:

$$\ln y = \ln a + b \ln x .$$

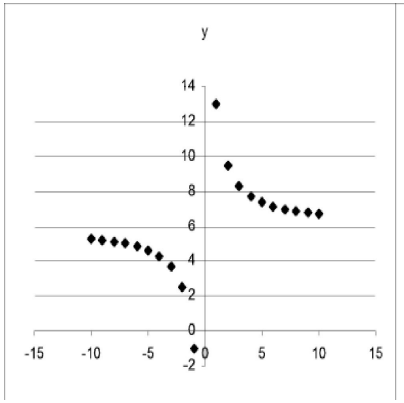
- Линеаризуется последнее выражение заменой:

$y1 = \ln y$, $a1 = \ln a$, $x1 = \ln x$.

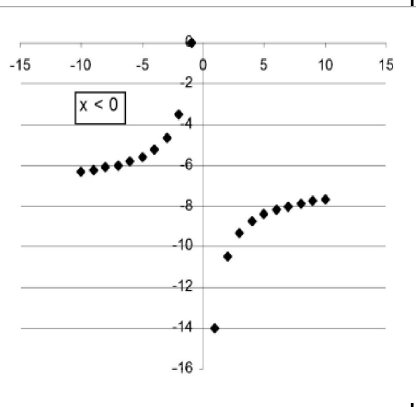
- В результате получаем:

$$y_1 = a_1 + b\,x_1$$

Связь гиперболическая: $y = a + b / x$ $(x \neq 0,\ b \neq 0)$



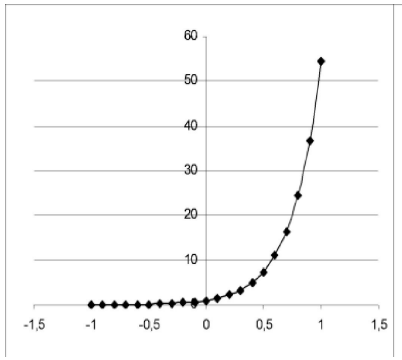
$b \geq 0$



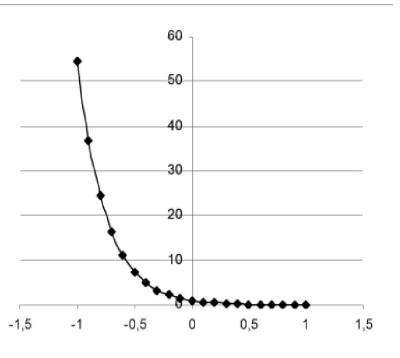
$b \leq 0$

- Модель не линейна относительно независимой переменной x.
- Линеаризация выполняется с помощью замены переменной: **x1 = 1/x**
- Исходное уравнение принимает вид: $y = a + b * x_1$
- Частный вид функции – обратно-пропорциональная: $y = b / x$ $(x, b \neq 0)$

Связь экспоненциальная: $y = e^{bx}$ $(b \neq 0)$



$b \geq 0$

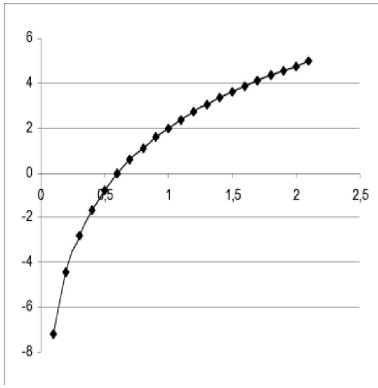


$b \leq 0$

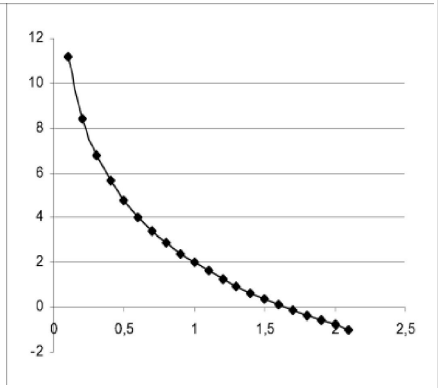
- Модель не линейна как относительно независимой переменной x, так и относительно коэффициента b.
- Линеаризация выполняется логарифмированием выражения:

  ln y = b * x.
- Замена переменных: $y1 = \ln y$,
- В результате исходное уравнение принимает вид: **y1 = b * x**

Связь логарифмическая (обратная экспоненциальной): $y = a + b \cdot \ln x$



$b \geq 0$

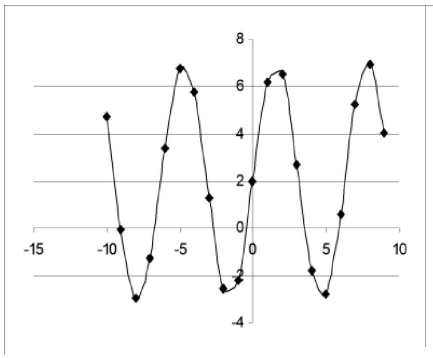


$b \leq 0$

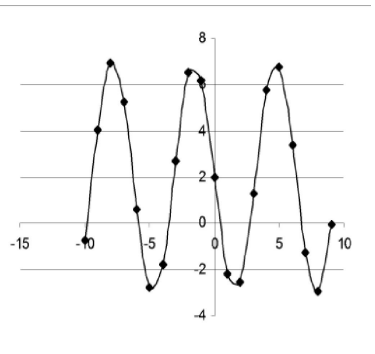
- Модель не линейна относительно независимой переменной x.
- Линеаризация выполняется с помощью замены: **x1 = ln x.**
- В результате получаем:

$$y = a + b\,x_1$$

Связь тригонометрическая с функцией синуса:



$b \geq 0$



$b \leq 0$

- Модель не линейна относительно независимой переменной x.
- Линеаризация выполняется с помощью замены:

x1 = sin x.

- В результате получаем:

$$y = a + b\,x_1$$

Функция Кобба-Дугласа

- характеризует связь между совокупным выпуском (доходом) и объемами используемых ресурсов.
- применяются для описания технологических процессов, в целом производственной деятельности предприятий, отрасли или экономики страны в целом.
- отражает устойчивую количественную связь между затратами и выпуском продукции.

Основные переменные модели:

- Капитал К (фактически использованный объем капитала),
- Труд L (численность занятых или отработанное время).
- Национальный доход (выпуск) – зависима переменная Y.

Производственные функции обладают следующими свойствами:

- Выпуск растет при росте затрат каждого фактора, т.е., первая производная от выпуска по каждому из факторов строго положительна: $Y1K > 0$, $Y1L > 0$
- Предельная производительность каждого фактора убывает, т.е., вторая производная от выпуска по каждому из факторов строго отрицательна: $Y11K < 0$, $Y11L < 0$
- Предельная производительность каждого фактора возрастает при росте затрат другого фактора, т.е., производная второго порядка по обоим факторам строго положительна: $Y11KL > 0$, $Y11LK > 0$
- Если один из факторов отсутствует, то выпуск равен нулю.

- По результатам модели:
  - увеличение затрат труда на 1% повлечет за собой рост национального дохода на **b** %,
  - а увеличение затрат капитала на 1% увеличит национальный доход на **a** %.
  - Таким образом, a и b являются эластичностями национального дохода по факторам производства.
- В случае, когда **a + b = 1** говорят **о постоянной отдаче от масштабов производства** – во сколько раз увеличиваются затраты ресурсов, во столько же раз увеличивается выпуск.
- При **a + b <1** имеет место **убывающая отдача от масштабов производства** – увеличение объема выпуска меньше увеличения затрат ресурсов (экономия на масштабах производства).
- При **a + b > 1** – **возрастающая отдача от масштабов производства** – увеличение объема выпуска больше увеличения затрат ресурсов (рост удельных издержек).

Алгоритм построения нелинейной модели

1. **Перевод модели Кобба-Дугласа в линейную выполняется с использованием процедуры логарифмирования:**
   - Для этого берутся логарифмы от всех значений указанных переменных.
   - Прологарифмированные значения будут играть роль переменных для построения регрессионной модели.

2. **При построении модели:**
   - в качестве Входного интервала **Y** выбираются значения из столбца **ln Y**, а
   - в качестве Входного интервала **X** – значения из столбцов **ln L** и **ln K**.
   - после процесса линеаризации проводится регрессионный анализ.
3. **Интерпретация уравнения линеаризованной модели Кобба-Дугласа:**

$$\ln Y = 2{,}529 + 0{,}616 \ln L + 0{,}370 \ln K.$$

   - с увеличением трудозатрат на 1% возрастает выпуск на 0,616%,
   - при увеличении капиталовложений на 1% следует ожидать роста выпуска на 0,37%.
   - Поскольку сумма коэффициентов перед факторными переменными не превышает 1 (0,616+0,370=0,986), можно говорить об убывающей отдаче от масштабов производства.
4. **Конвертация функции Кобба-Дугласа в исходный, не линеаризованном вид,**
   - необходимо пропотенцировать константу линеаризованного уравнения, поскольку константа совпадает с величиной **lnA**:
   - переменная $A = e^{2{,}529} = 12{,}546$
   - Эластичности выпуска по факторам производства α и β выводятся в линеаризованной модели в явном виде, т.е. α = 0,616, β = 0,370. В итоге функция Кобба-Дугласа для рассматриваемой выборки принимает вид:

$$Y = 12{,}546 * L\ 0{,}616 * K\ 0{,}37$$

Пример нелинейного моделирования

п/п	Выпуск	Капитало вложения	Трудозатраты
1	36209	24191	35963
2	38058	28505	37365
3	32511	17720	34360
4	39445	29892	37951
5	40832	37750	39522
6	34515	20647	35100
7	37442	26194	36672
8	38983	28659	37550
9	39291	28968	37689
10	37750	27581	37026
11	35285	22804	35454
12	33282	19106	34715
13	36980	25578	36471
14	38367	28505	37442
15	39907	30817	38058

Используем функцию LN (каждой ячейки)

п/п	Выпуск	Капитало вложения	Трудозатраты
1	10,50	10,09	10,49
2	10,55	10,26	10,53
3	10,39	9,78	10,44
4	10,58	10,31	10,54
5	10,62	10,54	10,58
6	10,45	9,94	10,47
7	10,53	10,17	10,51
8	10,57	10,26	10,53
9	10,58	10,27	10,54
10	10,54	10,22	10,52
11	10,47	10,03	10,48
12	10,41	9,86	10,45
13	10,52	10,15	10,50
14	10,55	10,26	10,53
15	10,59	10,34	10,55

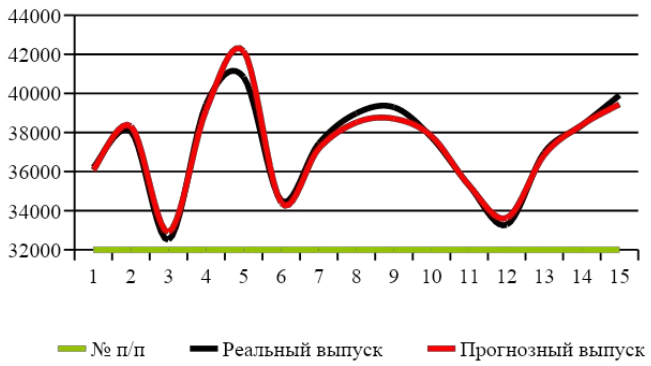
Проводим регрессионный анализ и получаем прогнозные значения для текущих наблюдений

Коэффициенты регрессии	a	b1	b2
Значение		0,173	0,833
Уравнение регрессии	N=C^0,173*L^0,833		

п/п	Выпуск	Прогнозный выпуск
1	10,50	10,49
2	10,55	10,55
3	10,39	10,40
4	10,58	10,57
5	10,62	10,65
6	10,45	10,45
7	10,53	10,52
8	10,57	10,56
9	10,58	10,56
10	10,54	10,54
11	10,47	10,47
12	10,41	10,42
13	10,52	10,51
14	10,55	10,56
15	10,59	10,58

Используем функцию EXP (каждой ячейки)

п/п	Прогнозный выпуск	Выпуск
1	36079	36209
2	38322	38058
3	32909	32511
4	39142	39445
5	42160	40832
6	34398	34515
7	37179	37442
8	38516	38983
9	38706	39291
10	37815	37750
11	35290	35285
12	33628	33282
13	36858	36980
14	38388	38367
15	39443	39907



Увеличение затрат на производство приводит к постоянному возрастанию отдачи от масштаба (0,173+0,833 = 1,006)

Регрессионные модели с фиктивными переменными

Использование фиктивных переменных в регрессионном анализе

- До сих пор в качестве факторов мы рассматривали экономические переменные, принимающие количественные значения.
- Однако результирующий признак может зависеть и от неколичественных (качественных) факторных признаков.
- Переменные, входящие в состав регрессионной модели, могут принимать как конечное, так и бесконечное множество значений.
- Для включения неколичественной переменной в модель необходимо перевести ее качественные значения в числовые величины.
- Это можно сделать с помощью фиктивных переменных.

Фиктивные переменные – это переменные бинарного типа, при котором переменная может принимать всего два значения: 1 или 0.

Фиктивная переменная **d** – такая же «равноправная» переменная, как и любая другая экзогенная переменная (**x**).

Ее «фиктивность» состоит только в том, что она количественным образом описывает качественный признак.

Например:
Имеется бинарная модель:

$$\text{Пробег} = 41{,}98 - 1{,}5 * \text{Возраст} + 1{,}11 * \text{Пол}$$

Фиктивная переменная «Пол»:

- принимает значение 1 – если водитель – женщина,
- принимает значение 0 – если водитель – мужчина.

Согласно построенной модели:

- увеличение срока эксплуатации автомобиля на 1 год приводит к снижению пробега на 1,5 км.
- переменная Пол принимает значение 1, если водитель – женщина,
- если водителем автомобиля является женщина, то пробег увеличивается на 1,11 км.
- если водителем автомобиля является мужчина, то пробег снижается на 1,11 км.

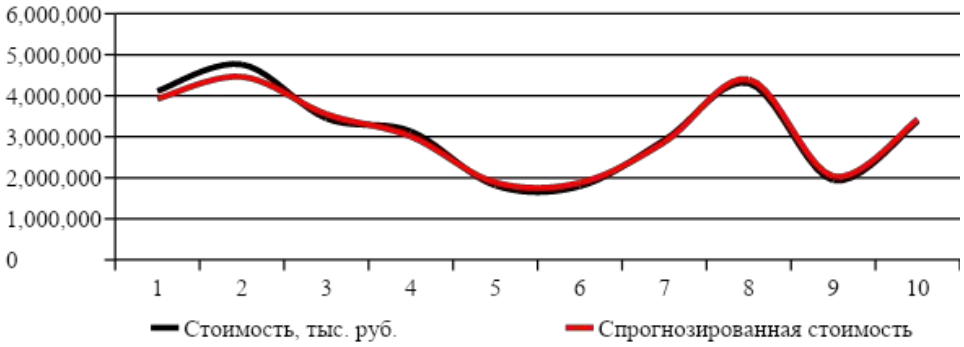
Например:

№Н	Площадь	Кухня	Этаж	Телефон	Стоимость
	ОП	К	Э	Т	С
1	92,4	16,4	крайний	есть	4 112 856
2	99,1	16,4	средний		4 762 126
3	82,8	11,3	крайний		3 447 405
4	61,2	9,4	средний		3 133 113
5	32,1	3,9	средний	есть	1 812 795
6	39,2	6,3	крайний	есть	1 793 342
7	57,5	6,5	средний	есть	2 917 295
8	97,5	10,8	средний	есть	4 303 481
9	35,9	6,6	средний		1 938 923
10	79,5	15,0	крайний		3 380 757

Используем формулу:
ЕСЛИ (ячейка = «средний»; 1; 0)
ЕСЛИ (ячейка = «есть»; 1; 0)

	Э	Т
1	0	1
2	1	0
3	0	0
4	1	0
5	1	1
6	0	1
7	1	1
8	1	1
9	1	0
10	0	0

Стоимость = 372939 + 38404 * Общая площадь + 282936 * Этаж



Использование фиктивных переменных в анализе сезонных колебаний

- Иногда заметное влияние на регрессионную зависимость оказывает сезонный характер изменения зависимой переменной.
- Если его воздействия не учитывать, то он вносит свой вклад в величину ошибки ε,
- Это приводит к снижению качественных характеристик регрессионной модели.

Основные этапы построения модели:

1. **Предполагаем наличие некоторого результативного признака yt в сезон t, изменение которого зависит от времени года.**
2. Для выявления влияния сезонности **вводим фиктивные переменные** d1, d2, d3.
3. **Полагаем, что**
   - d1 = 1, если сезон является зимним и d1 = 0 в остальных случаях;
   - d2 = 1, если сезон является весенним и d2 = 0 в остальных случаях;
   - d3 = 1, если сезон является летним и d3 = 0 в остальных случаях.
   - Четвертая фиктивная переменная осеннего сезона не вводится, поскольку:
     - ее добавление приведет к тому, что для любого сезона будет выполняться

$$d1 + d2 + d3 + d4 = 1,$$

     - что означает линейную зависимость регрессоров и
     - в результате делает невозможным получение оценок по МНК.

4. **Переходим к оценке уравнения**

$$y = a + b1\ d1 + b2\ d2 + b3\ d3 + \varepsilon.$$

- В нашем случае в качестве эталонной категории выбран осенний сезон.
- Выбор эталонной категории не оказывает воздействия на сущность уравнения регрессии
- Но от этого выбора зависит, какие тесты необходимо провести.
- В нашем случае фиктивные переменные будут использоваться для оценки различия в величине результативного показателя между осенним периодом и другими сезонами.

5. **С использованием МНК находятся числовые оценки параметров a, b1, b2, b3.**
   - Величины b1, b2, b3 (коэффициенты при фиктивных переменных) дают численную величину эффекта изменения объема потребления, вызываемого сменой сезона
   - Коэффициент b1 показывает изменение результативного показателя y в зимний период относительно осеннего,
   - Коэффициенты b2, b3 показывают изменение результативного показателя y в весеннем и летнем периодах относительно осеннего.

- Таким образом, среднее значение результативного показателя в каждый из сезонов достигает значения:

Для осеннего периода = a
Для зимнего периода = a + b1
Для весеннего периода = a + b2
Для летнего периода = a + b3

6. Тестируя нуль-гипотезу b1 = 0, проверяется предположение о несущественном различии в величине изменения результирующего показателя y между зимним и осенним сезонами.

7. Тестируя нуль-гипотезы для параметров b2 и b3, мы проверяем предположение о несущественном различии в величине изменения результирующего показателя y между весенним и осенним, а также летним и осенним сезонами.

Например:

- Предполагается проведение исследований сезонных колебаний цены на акции компании «Лукойл».
- Выделяются четыре сезона: зима, весна, лето, осень.
- В качестве **эталонного сезона** можно выбрать произвольный сезон.
  - Пусть это будет осень.
  - Эталонный сезон не включается в данные для построения регрессионной модели.
- Таким образом, модель будет включать:
  - в качестве результативного показателя цену закрытия,
  - в качестве факторных переменных – показатели сезонов зима, весна и лето.
- При выполнении регрессионного анализа:
  - в качестве Входного интервала Y в данном случае выделяются все значения цены закрытия (Last price),
  - в качестве Входного интервала X – все значения переменных зима, весна и лето.

	A	B	C	D	E	F	G	H
1	Date	Open price, USD	Last price, USD	Min price, USD	Max price, USD	зима	весна	лето
2	01 Aug	22,7	22,93	22,07	23,15	0	0	1
3	04 Aug	22,75	23,25	22,07	23,5	0	0	1
4	05 Aug	23,09	23,9	23,9	24,2	0	0	1
5	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6	01 Sep	22,125	22	21,5	22,35	0	0	0
7	02 Sep	22,05	23,65	22	23,75	0	0	0
8	03 Sep	24,5	24,5	21,95	25,5	0	0	0
9	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	01 Dec	19	18,45	18,4	19,33	1	0	0
11	02 Dec	18,4	19	18,4	19,22	1	0	0
12	03 Dec	19,7	20,8	19,3	20,55	1	0	0
13	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14	02 Mar	17,8	18,19	17,49	18,2	0	1	0
15	03 Mar	17,8	17,8	17,55	18,29	0	1	0
16	04 Mar	17,8	18,47	17,7	18,55	0	1	0
17	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

	A	B	C	D	E	F	G
1							
2	ВЫВОД ИТОГОВ						
3							
4	*Регрессионная статистика*						
5	Множественный R	0,771					
6	R-квадрат	0,594					
7	Нормированный R-квадрат	0,587					
8	Стандартная ошибка	2,332					
9	Наблюдения	183					
10							
11	Дисперсионный анализ						
12		*df*	*SS*	*MS*	*F*	*Значимость F*	
13	Регрессия	3	1422	474	87	0,000	
14	Остаток	179	973	5			
15	Итого	182	2396				
16							
17		*Коэффициенты*	*Стандартная*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
18	Y-пересечение	23,51	0,29	80,67	0,00	22,94	24,09
19	зима	-3,93	0,42	-9,42	0,00	-4,75	-3,11
20	весна	-5,92	0,48	-12,30	0,00	-6,88	-4,97
21	лето	2,26	0,59	3,85	0,00	1,10	3,41

Last price = 23,51 – 3,93 * зима – 5,92 * весна + 2,26 * лето.

- Константа регрессионной модели:
  - определяет величину результирующего показателя в эталонном сезоне.
  - Таким образом, среднее значение цены закрытия осенью составляет 23,51
- Остальные коэффициенты модели показывают величину отклонения средней цены закрытия в другие периоды от цены закрытия в эталонном периоде.
- Тогда чтобы рассчитать среднее значение цены закрытия в зимний период следует:
  - в регрессионную модель вместо показателя зима подставить 1,
  - а вместо всех остальных показателей подставить 0:

Last price (зима) = 23,51 – 3,93*1 – 5,92*0 + 2,26*0 = 19,58

- Аналогично можно получить средние значения цены закрытия в другие сезоны:
  Last price (весна) = 23,51 – 3,93*0 – 5,92*1 + 2,26*0 = 17,59.
  Last price (лето) = 23,51 – 3,93*0 – 5,92*0 + 2,26*1 = 25,77.
- Ориентируясь на средние значения цены закрытия в разные сезоны, можно сделать вывод:
  - что при долгосрочном инвестировании в ценные бумаги «Лукойла» целесообразно покупать акции весной,
  - а продавать выгодно летом.
- В этом случае появляется возможность заработать на разности цен покупки и продажи с каждой акции денежную сумму в размере 25,77 – 17,59 = 8,18 долларов.
- Этот заработок обусловлен правильным выбором времени покупки и продажи акций благодаря использованию построенной регрессионной модели.

Устранение трендовых компонент с помощью регрессионных моделей

Пример освобождения динамических рядов от сезонных колебаний

В задаче необходимо:

- Исследовать зависимость производства товаров двух заводов.
- Проверить динамические ряды на наличие тренда.
- Освободить показатели от тренда.
- Сравнить полученные результаты. Сделать выводы.

1. **Этап №1 – Проведение регрессионного анализа с целью выявления зависимости между переменными**

Временной период t	Производство чулков Yt	Производство женской обуви Xt
1	265,19	304,08
2	266,39	302,94
3	267,48	301,26
4	267,74	298,43
5	269,08	298,48
6	269,41	299,39
7	270,72	298,16
8	269,14	297,24
9	270,69	296,27
10	271,33	294,23

Y = 488,245 – 0,734*x

- R-квадрат = 0,97
- Коэффициент корреляции = 0,98
- P-значения переменных < 5%
- Значимость F = 0
- Полученные математические результаты противоречат экономическому смыслу.
- Скорее всего, зависимости между переменными нет,
- Наблюдается зависимость данных временных рядов от периодов времени.

2. **Этап №2 - Исключение трендовой составляющей и нахождение реальной регрессионной зависимости временных рядов.**

Удаление трендовой составляющей осуществляться двумя методами:

- методом аналитического выравнивания временных рядов;
- методом последовательных разностей.

Метод аналитического выравнивания временных рядов

Временной период (t)	Производство женской обуви (Xt)
1	304,08
2	302,94
3	301,26
4	298,43
5	298,48
6	299,39
7	298,16
8	297,24
9	296,27
10	294,23

Связь между переменными тесная:

X = 303,85 - 0,897 * t

- R-квадрат = 0,99
- Коэффициент корреляции = 0,99
- P-значения переменных < 5%
- Значимость F = 0
- Анализ показал, что Производство женской обуви тесно коррелирует с временными периодами.
- **Поправочный коэффициент для удаления трендовой зависимости 0,897**

Временной период (t)	Производство чулков (Yt)
1	265,19
2	266,39
3	267,48
4	267,74
5	269,08
6	269,41
7	270,72
8	269,14
9	270,69
10	271,33

Связь между переменными тесная:

$$Y = 265{,}009 + \mathbf{0{,}667}*t$$

- R-квадрат = 0,99
- Коэффициент корреляции = 0,99
- P-значения переменных < 5%
- Значимость F = 0
- Анализ показал, что Производство чулков тесно коррелирует с временными периодами.
- **Поправочный коэффициент для удаления трендовой зависимости 0,667**

3. **Этап №3 – Освобождение исходных данных от трендовой компоненты.**
   Освобождение от трендовой компоненты необходимо осуществлять по **формуле**:

- Ячейка Xt – Временной период * 0,897 (для производителей итальянской обуви)
- Ячейка Yt – Временной период * 0,667 (для производителей чулков)

Исходные Данные		Освобожденные от трендовой компоненты	
Производство чулков	Производство обуви	Yt	Xt
265,19	304,08	**264,52** = 265,19 –1*0,667	**304,98** = 304,08 – 1 * (-0,897)
266,39	302,94	**265,06** = 266, 39 – 2*0,667	**304,73** = 302,94 – 2 * (-0,897)
267,48	301,26	**265,48** = 267,48 – 3*0,667	**303,95** = 301,26 – 3 * (-0,897)
267,74	298,43	**265,07**	**302,02**
269,08	298,48	**265,74**	**302,96**
269,41	299,39	**265,41**	**304,77**
270,72	298,16	**266,05**	**304,44**
269,14	297,24	**263,80**	**304,41**
270,69	296,27	**264,68**	**304,34**
271,33	294,23	**264,66**	**303,20**

4. **Этап №4 – Нахождение регрессионной зависимости по данным, освобожденным от влияния трендовой компоненты.**

Временной период	Производство чулков Yt	Производство обуви Xt
1	264,52	304,98
2	265,06	304,73
3	265,48	303,95
4	265,07	302,02
5	265,74	302,96
6	265,41	304,77
7	266,05	304,44
8	263,80	304,41
9	264,68	304,34
10	264,66	303,20

Связь между переменными отсутствует:

Y = 221,38 + 0,144 * x

- R-квадрат = 0,025
- Коэффициент корреляции = 0,158
- P-значение переменной Xt > 5%
- Значимость F = 0,404 выборка нерепрезентативна
- Анализ показал, что после удаления зависимости от сезонных трендов **связь между переменными исчезла.**
- Можно сделать вывод, что в реальности **связи между ними не было.**

Метод последовательных разностей

Производство		Первые разности	
чулков	обуви	Δt Y	Δt X
265,19	304,08		
266,39	302,94	**1,20** = 266,39 – 265,19	**-1,14** = 302,94 – 304,08
267,48	301,26	**1,09** = 267,48 – 266,39	**-1,68** = 301,26 – 302,94
267,74	298,43	**0,26**	**-2,83**
269,08	298,48	**1,34**	**0,05**
269,41	299,39	**0,33**	**0,91**
270,72	298,16	**1,31**	**-1,23**
269,14	297,24	**-1,58**	**-0,92**
270,69	296,27	**1,55**	**-0,97**
271,33	294,23	**0,64**	**-2,04**

Связь между переменными отсутствует:

- Δy = 0,88 + 0,227 * Δx
- R-квадрат = 0,04
- Коэффициент корреляции = 0,21
- P-значение переменной Xt > 5%
- Значимость F = 0,274 выборка нерепрезентативна
- Анализ показал, что после удаления зависимости от сезонных трендов связь между переменными исчезла.
- Можно сделать вывод, что в реальности **связи между ними не было**.

Предпосылки МНК

Предпосылки метода наименьших квадратов

Для того чтобы регрессионный анализ давал наилучшие результаты должны выполняться **условия Гаусса-Маркова**, являющиеся предпосылками МНК.

Полученные в результате регрессионного анализа коэффициенты должны быть:

1. **Несмещенными - математическое ожидание остатков должно быть равно нулю.**
   - В результате при большом числе наблюдений остатки не будут накапливаться
   - Если оценки обладают свойством несмещенности, то их можно сравнивать по разным исследованиям.
2. **Эффективными - должны обладать наименьшей дисперсией.**
   - это означает возможность перехода от точечного оценивания к интервальному.
3. **Состоятельными – их точность должна увеличиваться при увеличении объема выборки.**
4. Значения случайной составляющей должны быть **независимы и случайно распределены**.

Предпосылки МНК

1. **Математическое ожидание случайной составляющей (остатков) в любом наблюдении должно быть равно нулю.**

$$M(e_t) = 0\,,\ \forall t.$$

- Иногда случайная составляющая будет положительной, иногда – отрицательной,
- но она не должна иметь систематического смещения ни в одну сторону.
- Если уравнение регрессии включает в себя константу, то это условие выполняется автоматически,
- роль константы состоит в определении систематической тенденции в **y**, которую не учитывают объясняющие переменные **x**, включенные в уравнение регрессии.

2. **Гомоскедастичность (постоянство дисперсии отклонений). Дисперсия случайной составляющей должна быть постоянна для всех наблюдений:**

$$\sigma^2(e_t) = \sigma^2(e_k) = \sigma^2(e)\,,\ \forall t \text{ и } k.$$

- Иногда случайная составляющая будет больше, иногда – меньше,
- однако не должно ситуации когда она бы порождала большую ошибку в одних наблюдениях, чем в других.
- Если рассматриваемое условие не выполняется, то коэффициенты регрессии, найденные по МНК, будут неэффективны.

3. **Отсутствие автокорреляции остатков. Любые случайные отклонения ut и uk должны быть независимыми друг от друга.** $\sigma(e_t e_k) = cov(e_t e_k) = 0\,,\ \forall t \neq k.$
   - Здесь **cov (et ek)** – это ковариация, т.е. среднее отклонение многомерной случайной величины от ее среднего значения.
   - Если случайная составляющая велика и положительна в одном наблюдении, это не должно вести к тому, что она будет большой и положительной в следующем наблюдении, и наоборот.
   - Случайные составляющие должны быть абсолютно независимы друг от друга.
4. **Значение любой независимой переменной в каждом наблюдении должно считаться экзогенным (полностью определяться внешними причинами, не учитываемыми в уравнении регрессии).**
   - Если это условие выполнено, то теоретическая ковариация между независимой переменной и случайной составляющей равна нулю. $\sigma(x_t, e_t) = 0\,,\ \forall t.$
5. **Линейность модели относительно параметров.**
6. **Отсутствие мультиколлинеарности.**
   - Между переменными должна отсутствовать сильная линейная зависимость.
7. **Нормальное распределение случайной составляющей.**
   - Если случайная составляющая нормально распределена, то так же будут распределены и коэффициенты регрессии.
   - Это позволяет прогнозировать их поведение (проверять статистические гипотезы и строить интервальные оценки).

Мультиколлинеарность

Мультиколлинеарность

- это сильная коррелированность двух или нескольких объясняющих переменных.
- в этом случае переменные меняются синхронно
- оказывается сложным, а иногда и невозможным, разделить их влияние на зависимую переменную.
- при наличии мультиколлинеарности оценки по МНК обладают неудовлетворительными свойствами.

Очень часто приходится сталкиваться с **несовершенной мультиколлинеарностью:**

- это стохастическая (вероятностная, случайная) связь между переменными.
- чем ближе по модулю коэффициент парной корреляции к 1, тем ближе мультиколлинеарность к совершенной и тем труднее разделить влияние каждой из объясняющих переменных на результирующий показатель.

Основная причина мультиколлинеарности – несколько независимых переменных могут иметь общий временной тренд, относительно которого они совершают малые колебания.

Признаки мультиколлинеарности:

- незначительное изменение исходных данных приводит к существенному изменению коэффициентов регрессионной модели.
- коэффициенты имеют большие стандартные ошибки и малую статистическую значимость (Р-значения больше 5%), в то время, как регрессионная модель в целом является значимой:
  - коэффициент детерминации стремится к единице
  - является статистически достоверным (значимость F меньше 5%);
- коэффициенты регрессии имеют нелогичные, с точки зрения теории, знаки
- коэффициенты регрессии имеют неоправданно большие значения (в этом случае незначительное изменение значений независимых переменных, входящих в модель, приводит к значительному изменению величины зависимой переменной).

Отрицательные последствия мультиколлинеарности:

- усложняется процедура отбора факторов, оказывающих влияние на результирующий показатель;
- искажается смысл коэффициента множественной корреляции, при расчете которого предполагается независимость регрессоров;
- искажается экономический смысл коэффициентов регрессии: в случае мультиколлинеарности значения коэффициентов ненадежны, и их нельзя использовать для интерпретации меры воздействия фактора на зависимую переменную;
- снижается точность оценки параметров регрессионной зависимости;
- критерии статистической значимости становятся ненадежными.

Для измерения мультиколлинеарности можно использовать коэффициент множественной детерминации:

- При отсутствии мультиколлинеарности факторов коэффициент множественной детерминации рассчитывается по формуле:

$$R^2 = \sum_{i=1}^{k} r_{yi}^2$$

где r^2_{yi} – коэффициент детерминации между **i-м** фактором и зависимой переменной **y**.

- При наличии мультиколлинеарности данное равенство не выполняется.
- Поэтому в качестве меры мультиколлинеарности можно использовать следующую разность:

$$M = R^2 - \sum_{i=1}^{k} r_{yi}^2 .$$

- Чем меньше величина M, тем меньше величина мультиколлинеарности.

Для устранения мультиколлинеарности используется метод исключения переменных:

- высоко коррелированные объясняющие переменные поэтапно удаляются из регрессионной модели, и она заново оценивается.
- Отбор переменных, подлежащих исключению, производится с помощью коэффициентов **парной корреляции** (это коэффициенты корреляции между парами объясняющих переменных).
- Если коэффициент парной корреляции $\geq$ **0,7** то одну из переменных можно исключить.
- Выбор исключаемой переменной проводят, исходя из управляемости факторов.
  - Обычно в модели оставляют тот фактор, для которого можно разработать мероприятия, обеспечивающие улучшение значения этого фактора в плановом периоде.
  - В ходе логического анализа на основе экономических знаний исследователь должен сделать вывод: можно ли разработать организационно-технические мероприятия, направленные на улучшение выбранных факторов?
  - Если это возможно, то факторы управляемы. Неуправляемые факторы могут быть исключены из модели.

Процедура отбора удаляемых факторов включает следующие этапы:

1. Проводится анализ рассчитанных значений коэффициентов парной корреляции между объясняющими факторами.
2. Проводится анализ тесноты взаимосвязи каждого объясняющего фактора с зависимой переменной:
   - факторы, для которых коэффициент парной корреляции с зависимой переменной **y**, равен нулю, **подлежат исключению в первую очередь**;
   - факторы, имеющие невысокое значение коэффициента парной корреляции с зависимой переменной, могут быть исключены из модели, но для них дополнительно рассчитывается **коэффициент β:**
     - он учитывает влияние анализируемых факторов на зависимую переменную с учетом различий в уровне их колеблемости.
     - показывает, на какую величину среднеквадратического отклонения – СКО (σ) изменяется зависимая переменная с изменением соответствующего фактора при фиксированном значении остальных факторов:

$$\beta = b_k * \sigma_{xk} / \sigma_y$$

где b_k – коэффициент регрессии при k-м факторе,
σ_{xk} – СКО (дисперсия) для k-фактора
σ_y – СКО (дисперсия) для результирующего показателя y

   - Из двух объясняющих факторов исключается тот, который имеет меньшее значение коэффициента β.

3. Прежде, чем вынести решение об исключении факторов, проводят дополнительное исследование с помощью **статистики Фишера F**.

- Рассчитывают значения F-статистики для переменных,
- также по таблице распределений Фишера находят **критическое значение F**.

$$F = \frac{(R_m^2 - R_{m_1}^2) \cdot (n - m - 1)}{(m - m_1) \cdot (1 - R_m^2)}$$

$R_{m_1}^2$ - коэффициент детерминации в модели с m1 факторами (m2 факторов удалено)

R_m^2 - коэффициент детерминации в модели без удаленных факторов

m1 – количество оставшихся после удаления факторов (m1= m – m2)

m2 – количество удаляемых факторов

n – количество наблюдений

- **Если рассчитанное F ≤ критического значения F,** то включение в регрессионную модель факторов не оказывает значимого влияния на зависимую переменную y, **и их можно удалить**.
- **Если рассчитанное F ≥ критического значения F**, то факторы совместно оказывают существенное влияние на зависимую переменную **y**, и, следовательно, **оба фактора исключать из регрессионной модели нельзя**.

Например:

	Безработные	Численность населения	Число занятых	Кол-во предприятий	Средний налог
	Безр	Нас	Труд	ЧП	Налог
1	46 474	226 765	108 121	3 849	52,25%
2	44 782	194 548	79 767	2 221	72,74%
3	27 651	141 104	89 930	4 168	48,51%
4	22 304	120 697	67 668	4 595	43,10%
5	20 061	98 629	41 369	3 815	52,66%
6	24 387	120 204	69 892	3 738	53,55%
7	41 249	209 438	76 398	4 371	45,74%
8	26 570	130 708	83 417	3 913	51,21%
9	30 348	168 817	102 269	5 236	35,06%
10	43 847	191 256	80 092	2 387	70,27%

Этап 1 – Проведение регрессионного анализа и исследование его результатов.

Коэффициенты регрессии	a	b1	b2	b3	b4
Значение	-2041,074	0,209	-0,004	-1,355	12981,596
Уравнение регрессии	Безр=-2041,074+0,209*Нас-0,004*Труд-1,355*ЧП+12981,596*Налог				
P-значение	93,227%	0,000%	21,082%	65,127%	58,916%
Вывод	незначим	значим	незначим	незначим	незначим
Надо исключить	**A**				

Регрессионная статистика	
Множественный R	0,999778527
R-квадрат	0,999557102
Нормированный R-квадрат	0,999486239
Стандартная ошибка	224,0654786
Наблюдения	30

Этап 2 - Построение матрицы попарных корреляций и ее анализ

	Коэф. Корреляции (R)				
	Безр	**Нас**	**Труд**	**ЧП**	**Налог**
Безр	1				
Нас	**0,946463658**	1			
Труд	0,673775441	**0,772484819**	1		
ЧП	-0,502683859	-0,19753633	0,011790401	1	
Налог	0,50337302	0,198313646	-0,011341572	**-0,999916077**	1

Из построенной матрицы мы видим:

- наиболее тесная корреляционная связь (R>0,7) наблюдается между факторами «Население» и «Труд», а также между «ЧП» и «Налог».
- сильную связь между зависимой эндогенной переменной «Безр» и «Нас».

Этап 3 – Расчет меры мультиколлинеарности (М)

$$M = R^2 - \sum_{j=1}^{m} r_{yj}^2$$

где R2 – коэф. детерминации регрессионного уравнения, полученного на Первом этапе,
r2 – коэффициенты детерминации в парных регрессиях **y** на **x** (коэффициенты корреляции, полученные на втором этапе, в квадрате).

- Коэффициенты детерминации в парных регрессиях y на x:

Регрессия Безр на	**r** ^{2}
Нас	**r** ^{2} = 0,946 ^ 2 = 0,896
Труд	**r** ^{2} = 0,673 ^ 2 = 0,454
ЧП	**r** ^{2} = - 0,503 ^ 2 = 0,253
Налог	**r** ^{2} = 0,503 ^ 2 = 0,253
Сумма	1,856

- Мера мультиколлинеарности = 0,999557102 – 1,856 = **- 0,856**

Этап 4 – Результаты регрессионного и корреляционного анализа:

- из модели необходимо исключить случайную переменную «А».
- однако после корреляционного анализа выяснилось, что в модели присутствует мультиколлинеарность.
- Наиболее тесная линейная взаимосвязь наблюдается между переменными «ЧП» и «Налог».
- Для того, чтобы определиться с тем, какую переменную необходимо исключить из модели, необходимо провести расчет и анализ коэффициента β.

Этап 5 – Расчет коэффициента Бета (β).

$$\beta = b_k \cdot \frac{D(x_k)}{D(y)}$$

где bk – коэффициент регрессии при k-м факторе,
Dxk – дисперсия для k-фактора
Dy – дисперсия для результирующего показателя y

		Дисперсия	Коэф. регрессии	Коэф. Бета
Y	**Безр**	**97721124,326**		
X_k	Нас	1771891706,593	0,209	**3,7835**
X_k	Труд	420133295,385	-0,004	**-0,0183**
X_k	ЧП	1181062,815	-1,355	**-0,0164**
X_k	Налог	0,018	12981,596	**0,0000**

- По результатам расчетов необходимо из нескольких коррелированных между собой факторов исключить тот, для которого коэффициент **β – наименьший**.
- В результате, из модели необходимо исключить переменную «Налог».

Этап 6 – Расчет статистики Фишера и выбор удаляемого фактора

$$F = \frac{(R_m^2 - R_{m_1}^2) \cdot (n - m - 1)}{(m - m_1) \cdot (1 - R_m^2)}$$

F расч = $\frac{(0{,}998 - 0{,}997) * (30 - 4 - 1)}{(4 - 2) * (1 - 0{,}998)}$ 29

F крит = F (m-1; n-m) F крит = F (3; 27) = 2,99 **F расч > F крит**

- **Поэтому, удалять переменные нельзя**.
- **НО!!!** Поскольку мультиколлинеарность все равно высокая, то мы удаляем фактор по коэф. β

Этап 7 – Удаление из модели переменной и повтор всех проведенных процедур для поиска качественной модели.

Результаты последующих этапов анализа:

Коэффициенты регрессии	a	b1	b2
Значение	10945,645	0,207	-2,988
Уравнение регрессии	Безр=10945,645+0,207Нас-2,988ЧП		
P-значение	0,000%	0,000%	0,000%
Вывод	значим	значим	значим

Регрессионная статистика

Множественный R	0,999760949
R-квадрат	0,999521954
Нормированный R-квадрат	0,999486543
Стандартная ошибка	223,9990034
Наблюдения	30

Анализ матрицы попарных корреляций:

	Безр	**Нас**	**ЧП**
Безр	1		
Нас	0,946463658	1	
ЧП	-0,502683859	-0,19753633	1

- Сильная корреляция между экзогенными факторами отсутствует.
- Мера мультиколлинеарности = -0,149 (очень низкая)
- Найденная регрессионная модель является качественной и пригодна для построения точных прогнозов

Автокорреляция остатков

Автокорреляция остатков

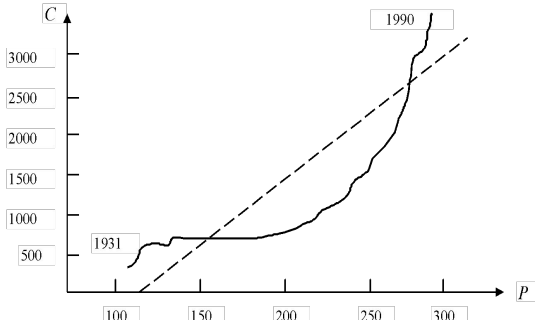
Статистическая значимость коэффициентов регрессии и близкое к 1 значение коэффициента детерминации R2 не всегда гарантируют высокое качество уравнения регрессии.

При анализе динамических рядов следует принимать во внимание:

- что наблюдения в различные моменты времени в определенной мере статистически зависимы (например ежедневный обменный курс доллара по отношению к рублю).
- ошибки, относящиеся к различным наблюдениям (различным моментам времени), могут быть зависимы между собой, т.е., коррелированы.
- в этом случае не выполняется одна из предпосылок метода наименьших квадратов.
- фактор **«е»** в этом случае представляет собой сумму влияния всех переменных, от которых в действительности зависит переменная **у**, но которые не были включены в модель.
- в некоторых случаях эти неучтенные факторы оказывают регулярное воздействие на величину ошибки **«е»**.
- в такой ситуации ошибки уравнения регрессии нельзя считать независимыми.

Например.

Исследуется зависимость объема потребления **C** от численности населения **P** в США в 1931-1990 гг. Корреляционное поле статистических данных выглядит следующим образом:



- Линейное уравнение регрессии имеет вид: C = -1817,3 + 16,7 * P
- Стандартные ошибки коэффициентов регрессии a = 84,7 b = 0,46.
- Их t-статистики и P-значения свидетельствуют о статистической значимости коэффициентов регрессии.
- Коэффициент детерминации R2 = 0,96
- Значимость F меньше 5%.
- Однако по расположению точек на корреляционном поле видно, что зависимость между **P** и **C** является экспоненциальной (не линейной):
  - в рассматриваемый период население США росло почти линейно,
  - а объем потребления – экспоненциально (с почти постоянными темпами прироста),

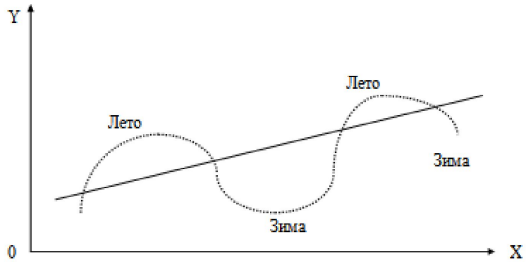
- Если использовать линейную регрессию для прогнозирования дальнейшей динамики потребления, то результат будет неудовлетворительным.
- В нашем примере **распределение отклонений от линии регрессии не случайно**, а обладает определенной закономерностью – знаки двух соседних отклонений одинаковы.
- Такая ситуация может быть следствием:
  - нелинейного характера связи переменных или
  - воздействия какого-либо фактора, не включенного в уравнение регрессии.
- **В данном случае не выполняются условия Гаусса-Маркова** о независимости отклонений реальных статистических данных от линии регрессии.
- **Наблюдается автокорреляция остатков:**
- В рассматриваемом примере отклонения не обладают постоянной дисперсией и не являются взаимно независимыми.
- В результате нарушение предпосылок МНК **делает полученные оценки коэффициентов регрессии неточными** и свидетельствует о неверной спецификации самого уравнения.

Автокорреляция – статистическая зависимость между ошибками различных наблюдений изучаемых показателей, упорядоченных во времени или в пространстве.

Упорядоченность наблюдений оказывается существенной если:

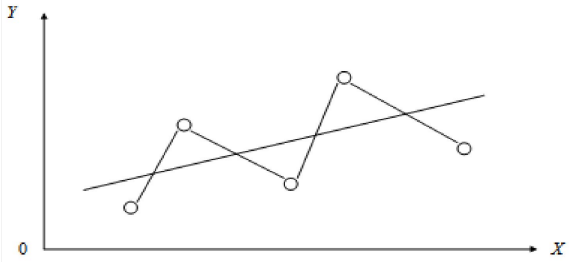
- прослеживается механизм влияния результатов предыдущих наблюдений на результаты последующих.
- случайные величины ошибок в регрессионной модели не оказываются независимыми.
- Часто автокорреляция встречается в регрессионном анализе временных рядов.

Например:



Исследуется спрос Y на напитки в зависимости от дохода X по ежемесячным данным.

- Трендовая зависимость может быть представлена функцией **Y = a + bX**.
- Однако фактические точки наблюдений будут превышать трендовую линию в летом и будут ниже зимой.



На рисунке явно видно:

- каждое следующее наблюдение не является независимым от предыдущего,
- отклонение (остаток) в каждом наблюдении также зависит от предыдущего
- это и есть автокорреляция: зависимость остатков.

Основные причины появления автокорреляции:

1. **Ошибки спецификации:**
   - не учет в модели какой-нибудь важной объясняющей переменной
   - неправильный выбор формы зависимости (например, линейной вместо нелинейной).
2. **Инерция в изменении экономических показателей:**
   - Многие экономические показатели (инфляция, безработица, ВНП и т.п.) обладают определенной цикличностью.
   - Циклическое развитие данных показателей происходит не мгновенно, а обладает определенной инерционностью.
3. **Эффект паутины:**
   - Наблюдается в производственной и других сферах.
   - Многие экономические показатели реагируют на изменение экономических условий с запаздыванием (временным лагом).
4. **Сглаживание данных.**
   - Зачастую данные по продолжительному временному периоду получают путем усреднения данных по интервалам меньшей длительности.
   - Это приводит к сглаживанию колебаний, которые имелись внутри основного периода,
   - В свою очередь может послужить причиной автокорреляции остатков.

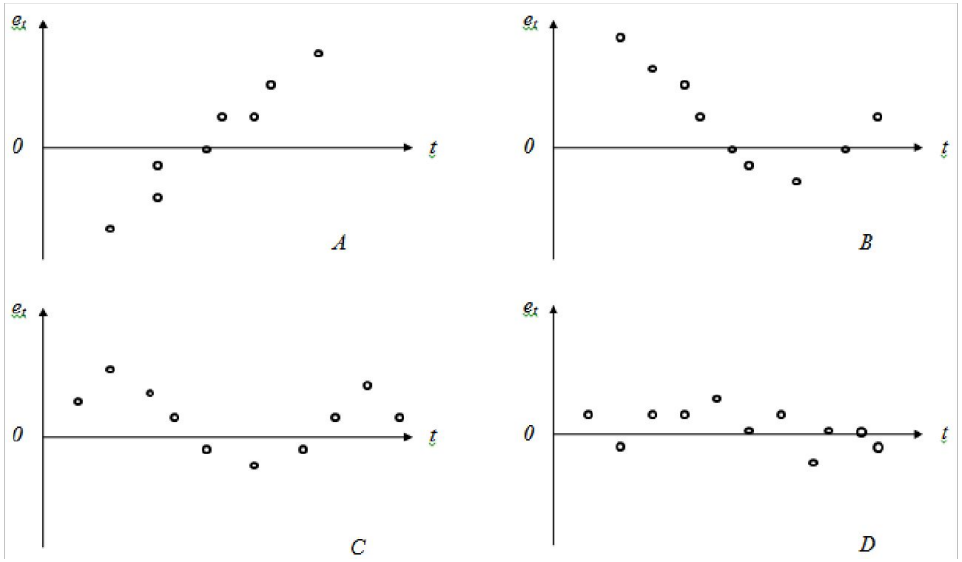
Последствия автокорреляции:

- Оценки коэффициентов регрессии, оставаясь линейными и несмещенными, перестают быть эффективными,
- Дисперсии оценок являются смещенными,
- Во многих случаях занижается оценка дисперсии регрессии.
- Вследствие этого ухудшаются прогнозные качества построенной регрессионной модели.
- Поэтому перед практическим использованием результатов проведенного регрессионного анализа следует выполнить проверку на наличие автокорреляции остатков.

Способы выявления наличия автокорреляции.

Графический метод.

1. **Строится последовательно-временной график.**
   - По оси абсцисс откладывается время получения статистических данных либо порядковый номер наблюдения,
   - а по оси ординат – отклонения **et** .
2. **Анализируется наличие связи между остатками наблюдений:**
   - На фрагментах **A, B, C** рис. видны определенные связи между отклонениями, т.е. имеет место автокорреляция.
   - На фрагменте **D** ее, по всей видимости, нет.



Метод рядов:

1. **Последовательно определяются знаки отклонений et , t = 1,2,...,T.**
2. **Ряд определяется как непрерывная последовательность одинаковых знаков.**
   - Количество знаков в ряду называется длиной ряда.
   - Визуальное распределение знаков свидетельствует о неслучайном характере связей между отклонениями.
   - Если рядов слишком мало или слишком много по сравнению с количеством наблюдений n, то вполне вероятно наличие автокорреляции остатков.
3. **Для более детального анализа предлагается следующая процедура.**
   - При достаточно большом количестве наблюдений (n1>10, n2>10) и отсутствии автокорреляции случайная величина k имеет нормальное распределение с математическим ожиданием и дисперсией, заданными следующими формулами:

$$M(k) = \frac{2n_1n_2}{n_1+n_2} + 1 \qquad \text{и} \qquad D(k) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

n – объем выборки;
n_1 – общее количество знаков «+» при n наблюдениях (положительные отклонения);
n_2 – общее количество знаков «–» при n наблюдениях (отрицательные отклонения);
k – количество рядов.

   - Тогда, если выполняется условие: $M(k) - u_{\alpha/2}D(k) < k < M(k) + u_{\alpha/2}D(k)$, то гипотеза об отсутствии автокорреляции не отклоняется (наблюдается автокорреляция).
4. **Для небольшого числа наблюдений (n1 < 20, n2 < 20)** для определения наличия автокорреляции **можно пользоваться таблицами критических значений Сведа и Эйзенхарта.**

Например:

- Имеется временная последовательность
- (– – – – –) (+ + + + + + + +) (– – –) (+ + + +) (–),
- т.е., 5 «–», 7 «+», 3 «–», 4 «+», 1 «–» при 20 наблюдениях.

Решение:

- n = 20,
- n1 (+) = 11,
- n2 (–) = 9,
- k = 5
- M(k) = (2·11·9) / (11+9) + 1 = 198 / 20 + 1 = 10,9
- D(k) = 35244 / 7600 = 4,63
- 10,9 – 4,63 < k < 10,9 + 4,63
- 6,27 < 5 < 15,53 – ложь
- Соответственно, гипотеза об отсутствии автокорреляции принимается.

Метод Дарбина-Уотсона.

- Это наиболее известный критерий обнаружения автокорреляции первого порядка является
- важнейшая характеристика качества регрессионной модели.
- Статистика DW рассчитывается по формуле:

$$DW = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2},$$

где e_t – остаток в момент времени t,
e_{t-1} – остаток в предыдущий момент времени t-1.

Для определения границ изменения величины DW

- выполним преобразования – в числителе раскроем квадрат разности двух соседних остатков: $\sum (e_t - e_{t-1})^2 = \sum (e_t^2 - 2e_t e_{t-1} + e_{t-1}^2) = \sum e_t^2 - 2\sum e_t e_{t-1} + \sum e_{t-1}^2 \approx$

$$\approx 2\sum e_t^2 - 2\sum e_t e_{t-1}.$$

- Тогда получаем:

$$DW \approx \frac{2(\sum e_t^2 - \sum e_t e_{t-1})}{\sum e_t^2} = 2(1 - r(e_t, e_{t-1})).$$

где r (et, et-1) – коэффициент корреляции соседних отклонений объясняющей переменной x:

$$r(e_t, e_{t-1}) = \frac{\sum e_t e_{t-1}}{\sqrt{\sum e_t^2 \sum e_{t-1}^2}}.$$

- Этот коэффициент также называется коэффициентом автокорреляции первого порядка.

- Расчет интервала коэффициента DW:
  - Если et = et-1, то r (et, et-1) = 1, тогда DW=0.
  - Если et = –et-1, то r (et,et-1) = –1, тогда DW=4.
  - во всех других случаях коэффициент DW будет лежать в интервале $0 < DW < 4$.
- Необходимым условием независимости случайных отклонений является близость значения статистики Дарбина-Уотсона к 2. **Если DW = 2**, то:
  - мы считаем отклонения от регрессии случайными.
  - Это означает, что построенная линейная регрессия отражает реальную зависимость.
  - Скорее всего, не осталось неучтенных существенных факторов, влияющих на зависимую переменную
  - другая нелинейная формула не превосходит по статистическим характеристикам предложенную линейную.
- Зачастую для коэффициентов регрессионной модели указываются два числа:
  - d1 – нижняя граница
  - d2 – верхняя граница.

- Какие значения DW можно считать статистически близкими к 2?
  - Для ответа на этот вопрос разработаны специальные таблицы критических точек Дарбина-Уотсона, позволяющие при данном числе наблюдений **n**, количестве объясняющих переменных **m** и заданном уровне точности определять **границы приемлемости (критические точки) наблюдаемой статистики DW**.
  - **Если DW < d1**, то это свидетельствует о положительной автокорреляции остатков.
  - **Если DW > 4–d1**, то это свидетельствует об отрицательной автокорреляции остатков.
  - **При d2 < DW < 4–d2**, гипотеза об отсутствии автокорреляции остатков принимается.
  - **При d1<DW<d2 или 4–d2 <DW< 4–d1** гипотеза об отсутствии автокорреляции остатков не может быть ни принята, ни отклонена.

- Не обращаясь к таблице критических точек Дарбина-Уотсона, можно пользоваться «грубым» правилом и считать, что **автокорреляция отсутствует, если** $1,5 < DW < 2,5$.

  – В случае, **когда** $DW < 1,5$ имеет место **положительная автокорреляция остатков.** (Положительная автокорреляция – когда за положительным отклонением следует положительное, а за отрицательным – отрицательное, иногда меняя знаки).

  – В случае, **когда** $DW > 2,5$ **имеет место отрицательная автокорреляция** остатков. (Отрицательная автокорреляция – когда за положительным отклонением следует отрицательное, и наоборот).

При использовании критерия Дарбина-Уотсона **необходимо учитывать следующие ограничения:**

- Критерий DW применяется лишь для тех моделей, которые содержат свободный член.
- Предполагается, что случайные отклонения **et** определяются по рекуррентной схеме, называемой авторегрессионной схемой первого порядка AR (1):

$$e_t = \rho \cdot e_t + v_t.$$

где vt – последовательность независимых, нормально распределенных случайных величин с нулевым математическим ожиданием и постоянной дисперсией,

ρ – коэффициент авторегрессии (его значение по модулю меньше 1).

- Статистические данные должны иметь одинаковую периодичность (не должно быть пропусков в наблюдениях).
- Критерий Дарбина-Уотсона **не применим для регрессионных моделей, содержащих** в составе объясняющих переменных **зависимую переменную с временным лагом в один период**, т.е., для так называемых авторегрессионных моделей.

Методы устранения автокорреляции.

- Так как автокорреляция чаще всего вызывается неправильной спецификацией модели, то **можно скорректировать саму модель**.
- **Возможно, автокорреляция вызвана отсутствием в модели некоторой важной объясняющей переменной**. Тогда следует определить данный фактор и учесть его в уравнении регрессии.
- Если все процедуры изменения спецификации модели не позволяют избавиться от автокорреляции, то **можно предположить, что она обусловлена какими-то внутренними свойствами ряда** {et}.
- В этом случае **можно воспользоваться авторегрессионным преобразованием**.
- Наиболее целесообразным и простым преобразованием в линейной модели является **авторегрессионная схема первого порядка AR(1)**

Рассмотрим ее.

Рассмотрим модель парной линейной регрессии: $y = \alpha + \beta x + u.$

- Тогда наблюдениям t и (t-1) соответствуют формулы:

$$y_t = a + bx_t + e_t \qquad y_{t-1} = a + bx_{t-1} + e_{t-1}$$

- Пусть случайные отклонения подвержены воздействию авторегрессии первого порядка:

$$e_t = \rho \cdot e_t + v_t$$

где vt (t = 2, 3, ... T) - случайные отклонения, удовлетворяющие всем предпосылкам МНК.

- Предположим, что тем или иным способом удалось определить величину параметра **ρ**
- Вычтем из левого уравнения правое, полученное на первом этапе, умноженное на **ρ**:

$$y_t - \rho \cdot y_{t-1} = a \cdot (1 - \rho) + b \cdot (x_t - \rho \cdot x_{t-1}) + (e_t - \rho \cdot e_{t-1}).$$

- Обозначив $y_t^* = y_t - \rho \cdot y_{t-1}, \quad x_t^* = x_t - \rho \cdot x_{t-1}, \quad a^* = a \cdot (1 - \rho),$ получим линейное уравнение следующего вида:

$$y_t^* = a^* + b\, x_t^* + v_t$$

- Так как коэффициент **ρ** известен, то y_t^*, x_t^*, v_t вычисляются просто.
- Так как случайные отклонения **vt** удовлетворяют предпосылкам МНК, то оценки **a*** и **b** будут обладать свойствами наилучших линейных несмещенных оценок.
- Ситуации, когда параметр **ρ** известен, встречаются крайне редко. Поэтому возникает необходимость найти его величину.

Два подхода к оценке параметра ρ:

1. **На основе статистики Дарбина-Уотсона.**

- Статистика Дарбина-Уотсона тесно связана с коэффициентом корреляции между соседними отклонениями через соотношение:

$$DW \approx 2(1 - r(e_t, e_{t-1})).$$

- Тогда в качестве оценки коэффициента **ρ** может быть взят коэффициент $r = r(e_t, e_{t-1})$, тогда:

$$r \approx 1 - \frac{DW}{2}$$

- Этот метод оценивания эффективен при наличии большого количества наблюдений.
- В этом случае оценка **r** параметра **e** будет достаточно точной.

2. Метод Хилдрета-Лу.

- Рассмотрим зависимость показателя **y** от значений **k** регрессоров:

$$y_t = a + b_1 x_{t,1} + b_2 x_{t,2} + \ldots + b_k x_{t,k} + \varepsilon_t$$

- В данном случае для оценивания системы применяется обобщенный метод наименьших квадратов.
- Запишем систему для момента времени **t-1**:

$$y_{t-1} = a + b_1 x_{t-1,1} + b_2 x_{t-1,2} + \ldots + b_k x_{t-1,k} + \varepsilon_{t-1}$$

- Умножим в уравнении обе части равенства на параметр **ρ** и вычтем почленно из предыдущего равенства. В результате получим:

$$y_t - \rho \cdot y_{t-1} = a - a + b_1 (x_{t,1} - \rho \cdot x_{t-1,1}) + b_2 (x_{t,2} - \rho \cdot x_{t-1,2}) +$$
$$\ldots + b_k (x_{t,k} - \rho \cdot x_{t-1,k}) + (\varepsilon_t - \rho \cdot \varepsilon_{t-1}) \qquad (20).$$

- Согласно приведенным выше утверждениям, $(\varepsilon_t - \rho \cdot \varepsilon_{t-1}) = v_t$
- Поскольку **vt**, по предположению, нормально распределенная случайная величина с нулевым средним и постоянной дисперсией, то к полученному уравнению можно применить обычный метод наименьших квадратов (МНК).

Суть процедуры Хилдрета-Лу достаточно проста:

- Из интервала от –1 до +1 возможного изменения коэффициента **ρ** берутся последовательно некоторые значения.
- Для каждого из них проводится оценивание преобразованной системы.
- Определяется то значение параметра **ρ**, **для которого сумма квадратов отклонений (остаточная дисперсия) в минимальна**.
- При необходимости в некоторой окрестности найденного значения строится более мелкая сетка и процесс повторяется.
- Итерации заканчиваются, когда будет достигнута желаемая точность.

Например:

	A	B	C	D	E	F	G
38							
39	*Регрессионная статистика*						
40	Множественный R	0,689					
41	R-квадрат	0,475					
42	Нормированный R-квадрат	0,456					
43	Стандартная ошибка	2,987					
44	Наблюдения	30					
45							
46	Дисперсионный анализ						
47		*df*	*SS*	*MS*	*F*	*Значимость F*	
48	Регрессия	1	226	226	25	0,000	
49	Остаток	28	250	9			
50	Итого	29	476				
51							
52		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
53	Y-пересечение	34,128	2,906	11,743	0,000	28,175	40,081
54	x	-0,180	0,036	-5,034	0,000	-0,254	-0,107

1. **Этап 1 – По выведенным значениям остатков находится статистика Дарбина-Уотсона и тестируется гипотеза о наличии автокорреляции остатков**
   - для первого остатка: $(e_t - e_{t-1})^2 = (3,3484 - 2,4724)^2 = 0,7673$
     $e_t^2 = (3,3484)^2 = 11,2116$

58	ВЫВОД ОСТАТКА		
59			
60	*Наблюдение*	*Предсказанное y*	*Остатки*
61	1	20,4795	-3,3484
62	2	16,1435	-2,4724
63	3	17,4631	-0,2711
64	4	18,2697	-2,4830
65	5	16,6279	-4,7695
66	6	19,0036	-4,1634
67	7	23,2818	-5,5440
68	8	21,8762	-3,9052
69	9	24,5648	0,2509
70	10	24,3711	5,9377
71	11	17,4550	3,0708
72	12	20,4475	-0,1691
73	13	17,1303	2,3452
74	14	17,6008	-1,6808
75	15	17,5141	-0,0406
76	16	23,3806	-1,2093
77	17	17,9797	2,4989
78	18	20,9736	0,5588
79	19	21,9137	-1,5297
80	20	23,5552	3,7875
81	21	17,2778	0,2624
82	22	21,9047	3,1858
83	23	16,1415	2,9774
84	24	16,6308	1,2407
85	25	20,3748	-3,1694
86	26	16,7614	1,5531
87	27	22,2115	1,2920
88	28	23,4223	-1,3690
89	29	21,3002	3,4306
90	30	16,7098	3,7329

Остатки	$(e_t - e_{t-1})^2$	e_t^2
-3,3484		11,2116
-2,4724	0,7673	6,1128
-0,2711	4,8459	0,0735
-2,4830	4,8929	6,1654
-4,7695	5,2279	22,7480
-4,1634	0,3674	17,3337
-5,5440	1,9061	30,7357
-3,9052	2,6857	15,2503
0,2509	17,2732	0,0630
5,9377	32,3398	35,2568
3,0708	8,2193	9,4300
-0,1691	10,4969	0,0286
2,3452	6,3215	5,4999
-1,6808	16,2082	2,8249
-0,0406	2,6901	0,0016
-1,2093	1,3658	1,4623
2,4989	13,7502	6,2443
0,5588	3,7639	0,3122
-1,5297	4,3616	2,3399
3,7875	28,2719	14,3450
0,2624	12,4260	0,0689
3,1858	8,5460	10,1492
2,9774	0,0434	8,8649
1,2407	3,0162	1,5393
-3,1694	19,4488	10,0451
1,5531	22,3024	2,4122
1,2920	0,0682	1,6693
-1,3690	7,0811	1,8742
3,4306	23,0360	11,7687
3,7329	0,0914	13,9342
Сумма	**261,815**	**238,554**

$$DW = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$$ DW = 261,815 / 238,554 = 1,099

- Поскольку статистика Дарбина-Уотсона DW меньше 1,5 значит наблюдается положительная автокорреляция остатков.

2. **Этап 2 – Преобразование исходных данных с учетом коэффициента авторегрессии ρ**
   - В качестве **ρ** можно взять произвольное число из интервала от –1 до +1.
   - Примем значение **ρ = – 0,9**
   - проведем авторегрессионное преобразование исходных данных с выбранным значением параметра **ρ**

Месяцы	Объем продаж	Цена	Промежуточные итоги	
	V	C	$V_t - pV_{t-1}$	$C_t - pC_{t-1}$
1	17,1	75,6	29,089	167,702
2	13,7	99,6	29,496	182,012
3	17,2	92,3	31,259	170,963
4	15,8	87,9	26,066	176,037
5	11,9	97,0	25,513	171,061
6	14,8	83,8	31,094	135,511
7	17,7	60,1	33,935	121,966
8	18,0	67,9	40,990	114,078
9	24,8	53,0	52,643	101,744
10	30,3	54,1	47,804	141,029
11	20,5	92,4	38,752	158,937
12	20,3	75,8	37,726	162,393
13	19,5	94,2	33,448	176,328
14	15,9	91,6	31,802	174,462
15	17,5	92,0	37,898	142,391
16	22,2	59,5	40,433	143,061
17	20,5	89,5	39,963	153,405
18	21,5	72,9	39,763	133,267
19	20,4	67,7	45,688	119,485
Месяцы	Объем продаж	Цена	Промежуточные итоги	
20	27,3	58,6	42,149	146,080
21	17,5	93,4	40,877	151,746
22	25,1	67,7	41,700	160,606
23	19,1	99,7	35,078	186,633
24	17,9	96,9	33,290	163,449
25	17,2	76,2	33,799	164,800
26	18,3	96,2	39,987	152,622
27	23,5	66,0	43,206	118,736
28	22,1	59,3	44,579	124,456
29	24,7	71,1	42,700	160,472
30	20,4	96,5		

1й месяц:

$V_t - pV_{t-1}$ 13,7–(- 0,9*17,1) = 29,09

$C_t - pC_{t-1}$ 99,6–(- 0,9*75,6) = 167,7

3. Этап 3 – Строится линейная регрессионная зависимость

$$(V_t - \rho \cdot V_{t-1}) \text{ на } (C_t - \rho \cdot C_{t-1}),$$

	A	B	C	D	E	F	G
36							
37	ВЫВОД ИТОГОВ						
38							
39	*Регрессионная статистика*						
40	Множественный R	0,696					
41	R-квадрат	0,484					
42	Нормированный R-квадрат	0,465					
43	Стандартная ошибка	4,769					
44	Наблюдения	29					
45							
46	Дисперсионный анализ						
47		*df*	*SS*	*MS*	*F*	*Значимость F*	
48	Регрессия	1	576,543	577	25	0,000	
49	Остаток	27	614,019	23			
50	Итого	28	1190,562				
51							
52		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
53	Y-пересечение	68,180	6,135	11,113	0,000	55,591	80,769
54	$C_t - \rho \cdot C_{t-1}$	-0,203	0,040	-5,035	0,000	-0,285	-0,120

- По выведенным результатам определяется величина остаточной дисперсии ESS.
- Если коэффициент авторегрессии **ρ = –0,9**, то величина остаточной дисперсии ESS составляет 614,019.

4. Этап 4 – Изменяем величину ρ

- Берем **ρ = –0,8**
- преобразуем исходные данные с учетом нового значения **ρ**.
- Далее строится линейная регрессионная зависимость (Vt – ρ*Vt-1) на (Ct – ρ*Ct-1) и определяется остаточная дисперсия.
- Задается новое значение коэффициента авторегрессии **ρ** и т.д.
- **Составляется таблица значений коэффициента авторегрессии ρ и соответствующих ему значений остаточных дисперсий**.

ρ	*ESS*
-0,9	614,019
-0,8	553,703
-0,7	498,080
-0,6	447,135
-0,5	400,847
-0,4	359,186
-0,3	322,119
-0,2	289,609
-0,1	261,623
0,1	219,154
0,2	204,685
0,3	194,772
0,4	189,468
0,5	**188,830**
0,6	192,907
0,7	201,737
0,8	215,343
0,9	233,735

5. **Этап 5 – Определение оптимального коэффициента ρ**

- Наименьшее значение остаточной дисперсии соответствует значению **ρ = 0,5**
- При этом остаточная дисперсия уменьшалась при изменении **ρ** от 0,4 до 0,5 и увеличивалась при изменении **ρ** от 0,5 до 0,6.
- Если значение **ρ** необходимо найти с точностью до сотых, следует провести исследование величины **ρ** на интервале от 0,4 до 0,6.
- **Примем значение ρ = 0,51**. Тогда остаточная дисперсия ESS = 189,0
- т.е. увеличение оптимального **ρ** на 0,1 привело к увеличению остаточной дисперсии,
- значит, **следует последовательно уменьшать параметр ρ** и последовательно фиксировать значения остаточных дисперсий.
- В результате получаем следующую таблицу.

ρ	*ESS*
0,51	189,024
0,49	188,682
0,48	188,581
0,47	188,528
0,46	**188,522**
0,45	188,562

- **оптимальное значение коэффициента авторегрессии ρ=0,46**

6. **Этап 6 – Необходимо записать уравнение регрессии, скорректировав его параметры с учетом найденного значения коэффициента авторегрессии ρ:**

$$(V_t - \rho V_{t-1}) = (a - \rho a) + b(C_t - \rho C_{t-1}) + (e_t - \rho e_{t-1}).$$

Для этого необходимо провести регрессионный анализ с учетом найденного оптимального значения **ρ**

	A	B	C	D	E	F	G
36							
37	ВЫВОД ИТОГОВ						
38							
39	*Регрессионная статистика*						
40	Множественный R	0,725					
41	R-квадрат	0,525					
42	Нормированный R-квадрат	0,508					
43	Стандартная ошибка	2,642					
44	Наблюдения	29					
45							
46	Дисперсионный анализ						
47		*df*	*SS*	*MS*	*F*	*Значимость F*	
48	Регрессия	1	208,490	208	30	0,000	
49	Остаток	27	188,522	7			
50	Итого	28	397,011				
51							
52		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
53	Y-пересечение	18,037	1,417	12,726	0,000	15,129	20,945
54	C_t - ρ·C_{t-1}	-0,167	0,031	-5,464	0,000	-0,230	-0,105

7. Этап 7 – Определение величины статистики Дарбина-Уотсона

Остатки	$(e_t - e_{t-1})^2$	e_t^2
-1,3870		1,9238
0,6511	4,1539	0,4239
-2,5590	10,3047	6,5485
-3,9736	2,0011	15,7895
-2,0893	3,5506	4,3652
-3,5183	2,0420	12,3784

Остатки	$(e_t - e_{t-1})^2$	e_t^2
-1,4884	4,1205	2,2153
2,1551	13,2751	4,6445
5,8266	13,4799	33,9493
-0,1500	35,7197	0,0225
-1,6246	2,1744	2,6393
2,0406	13,4337	4,1640
-2,9975	25,3825	8,9850
0,4726	12,0416	0,2234
-1,0234	2,2380	1,0473
2,6366	13,3956	6,9517
-0,6130	10,5599	0,3758
-1,8406	1,5070	3,3878
4,5246	40,5158	20,4720
-1,9549	41,9839	3,8216
3,1334	25,8908	9,8182
1,0096	4,5105	1,0193
-0,4042	1,9988	0,1634
-3,7609	11,2674	14,1444
2,6044	40,5170	6,7829
0,6853	3,6829	0,4696
-1,9494	6,9416	3,8002
3,8805	33,9877	15,0583
1,7136	4,6955	2,9364
-1,3870	4,1539	1,9238
Сумма	**385,3724**	**188,5214**

DW = 2,0442

Проведенное преобразование увеличило значение статистики DW с величины 1,098 до величины 2,0442.

Поскольку значение DW лежит в пределах от 1,5 до 2,5 значит **автокорреляция остатков отсутствует.**

8. Этап 8 – Формирование уравнения

$$(V_t - \rho V_{t-1}) = (a - \rho a) + b\,(C_t - \rho C_{t-1}) + (e_t - \rho e_{t-1}) =$$
$$18{,}037 - 0{,}167\,(C_t - \rho C_{t-1}) + (e_t - \rho e_{t-1}).$$

В нашем случае:

a – ρ*a = 18,037

b = – 0,167,

тогда a*(1- ρ) = 18,037 $a = \frac{18{,}037}{1-\rho} = \frac{18{,}037}{1-0{,}46} = 33{,}402$.

С учетом найденных значений уравнение связи между объемом продаж и ценой записывается в виде:

V = 33,042 – 0,167 C