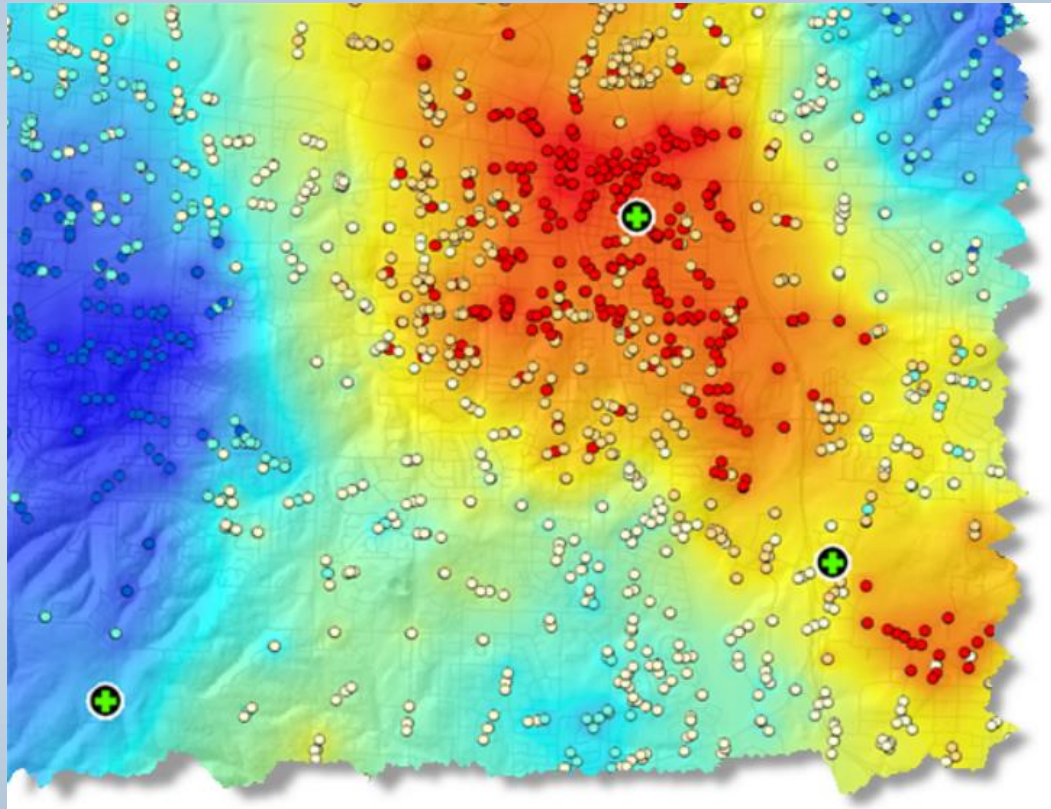


Регрессионный анализ.



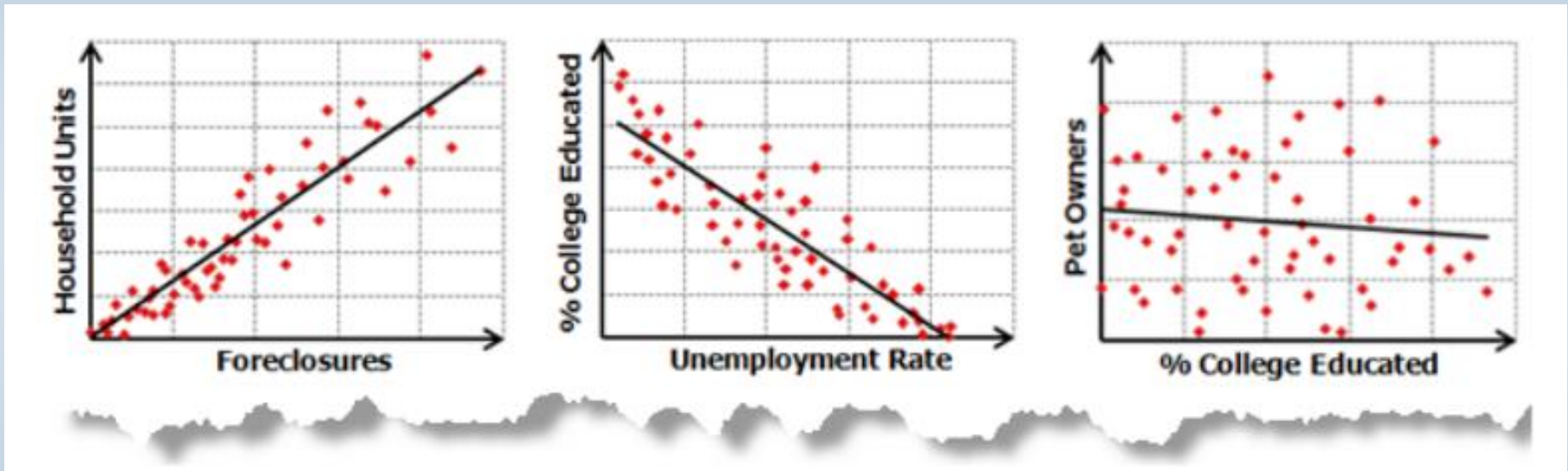
КОЛЬЦОВ С.Н

Примеры применение регрессионного анализ.

1. Моделирование числа поступивших в университет для лучшего понимания факторов, удерживающих детей в том же учебном заведении.
2. Моделирование потоков миграции в зависимости от таких факторов как средний уровень зарплат, наличие медицинских, школьных учреждений, географическое положение...
3. Моделирование дорожных аварий как функции скорости, дорожных условий, погоды и т.д.,
4. Моделирование потерь от пожаров как функции от таких переменных как количество пожарных станций, время обработки вызова, или цена собственности.

Суть регрессионного анализа заключается в нахождении наиболее важных факторов, которые влияют на зависимую переменную.

Примеры применение регрессионного анализ.



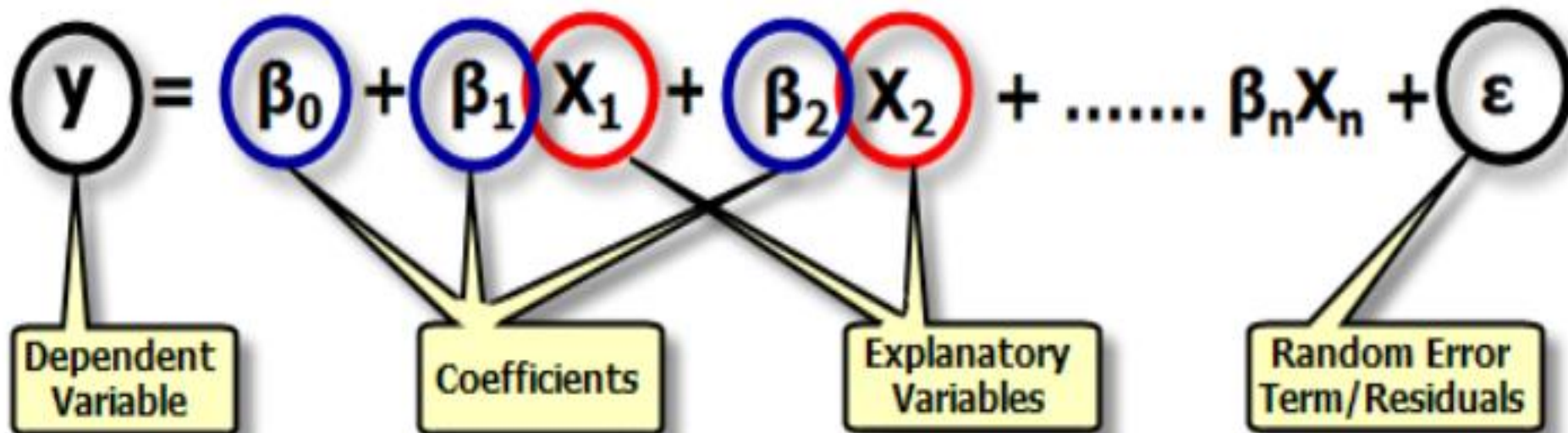
Связь между переменными может быть положительная, отрицательная или плохая

Термины и концепции регрессионного анализа

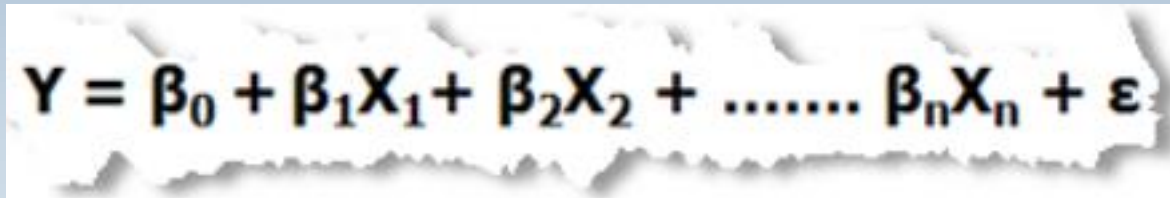
Уравнение регрессии. Это математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо смоделировать



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$



Термины и концепции регрессионного анализа


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$$

Зависимая переменная (Y) — это переменная, описывающая процесс, который мы пытаемся предсказать или понять.

Независимые переменные (X) это переменные, используемые для моделирования или прогнозирования значений зависимых переменных. В уравнении регрессии они располагаются справа от знака равенства и часто называются объяснительными переменными. Зависимая переменная - это функция независимых переменных.

Коэффициенты регрессии (β) — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

Невязки. Существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как **случайные ошибки ε** .

Термины и концепции регрессионного анализа



Создание регрессионной модели представляет собой итерационный процесс, направленный на поиск эффективных независимых переменных, чтобы объяснить зависимые переменные, которые мы пытаемся смоделировать или понять, запуская инструмент регрессии, чтобы определить, какие величины являются эффективными предсказателями. Затем пошаговое удаление и/или добавление переменных до тех пор, пока вы не найдете наилучшим образом подходящую регрессионную модель. Т.к. процесс создания модели часто исследовательский, он никогда не должен становиться простым "подгоном" данных. **Процесс построения регрессионной модели должен учитывать теоретические аспекты, мнение экспертов в этой области и **здоровый смысл.****

Виды регрессионного анализа

Различают линейные и нелинейные регрессии.

Линейная регрессия: $y = a + b \cdot x + \varepsilon$

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам.

Регрессии, нелинейные по объясняющим переменным:

полиномы разных степеней $y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + \varepsilon$

равносторонняя гиперболола $y = a + \frac{b}{x} + \varepsilon$.

Регрессии, нелинейные по оцениваемым параметрам:

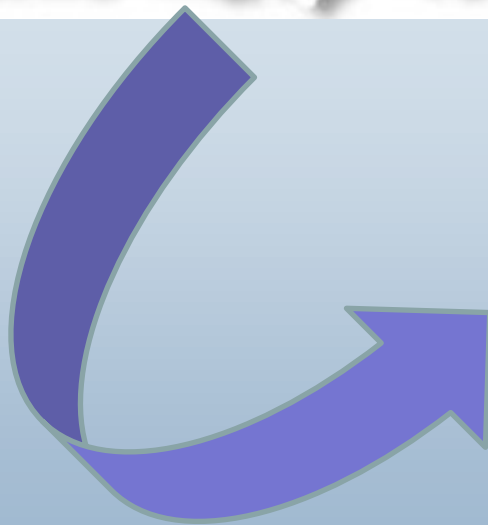
степенная $y = a \cdot x^b - \varepsilon$

показательная $y = a \cdot b^x - \varepsilon$

Линейный регрессионный анализ

Построение уравнения регрессии сводится к оценке ее параметров. Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических минимальна.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$



$$\begin{cases} Y^1 = \alpha + \beta_1 X_1^1 + \dots + \beta_k X_k^1 + u^1, \\ Y^2 = \alpha + \beta_1 X_1^2 + \dots + \beta_k X_k^2 + u^2, \\ \dots \dots \dots \\ Y^n = \alpha + \beta_1 X_1^n + \dots + \beta_k X_k^n + u^n, \end{cases}$$

Элементы матричного анализа

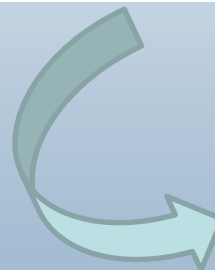
Пусть у нас есть такая таблица, мы предполагаем, что наша зависимая переменная y связана с переменными x_0 и x_1 линейным образом:

$$y = b_0 x_0 + b_1 x_1.$$

Номер опыта	x_0	x_1	y
1	+1	-2	0
2	+1	-1	1
3	+1	0	2
4	+1	+1	3
5	+1	+2	4

$$Y = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \quad X = \begin{bmatrix} +1 & -2 \\ +1 & -1 \\ +1 & 0 \\ +1 & +1 \\ +1 & +2 \end{bmatrix}$$

$$B = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$



$$\begin{aligned} 0 &= b_0 * 1 + b_1 * (-2), \\ 1 &= b_0 * 1 + b_1 * (-1), \\ 2 &= b_0 * 1 + b_1 * 0, \\ 3 &= b_0 * 1 + b_1 * (+1), \\ 4 &= b_0 * 1 + b_1 * (+2). \end{aligned}$$

Элементы матричного анализа и регрессионный анализ

Наша регрессионная модель может быть записана в матричном виде:

$$\begin{array}{l}
 0 = b_0 * 1 + b_1 * (-2), \\
 1 = b_0 * 1 + b_1 * (-1), \\
 2 = b_0 * 1 + b_1 * 0, \\
 3 = b_0 * 1 + b_1 * (+1), \\
 4 = b_0 * 1 + b_1 * (+2).
 \end{array}
 \quad \Rightarrow \quad
 \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}
 =
 \begin{bmatrix} +1 & -2 \\ +1 & -1 \\ +1 & 0 \\ +1 & +1 \\ +1 & +2 \end{bmatrix}
 *
 \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}
 \quad \Rightarrow \quad
 \mathbf{Y} = \mathbf{X}\mathbf{B}$$

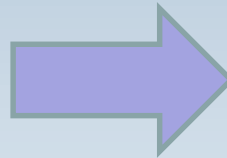
Определение обратной матрицы: Обратная матрица — такая матрица A^{-1} , при умножении на которую, исходная матрица A даёт в результате единичную матрицу E :

$$AA^{-1} = A^{-1}A = E$$

$$E = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Элементы матричного анализа и регрессионный анализ

$$\mathbf{YX}^{-1} = \mathbf{XX}^{-1} \mathbf{B}$$



$$\mathbf{YX}^{-1} = \mathbf{B}$$

$$\mathbf{XX}^{-1} = \mathbf{1}$$

Таким образом, мы получим в явной форме набор уравнений на компоненты вектора \mathbf{B} , то есть наше искомое решение регрессионной задачи.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \cdot \begin{bmatrix} x_{01} & x_{11} & \dots & x_{k1} \\ x_{02} & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \dots & \vdots \\ x_{0N} & x_{1N} & \dots & x_{kN} \end{bmatrix}^{-1} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$

Смысл коэффициента регрессии

В общем случае коэффициент регрессии k показывает, как в среднем изменится *результативный признак* (Y), если *факторный признак* (X) увеличится на единицу .
 $Y = 87610 + 2984 X$; X – число рабочих, Y – объем годового производства (руб.).

Пример интерпретации коэффициента регрессии

- В уравнении $Y = 87610 + 2984 X$; коэффициент регрессии равен +2984.

Что это означает?

• В данном случае смысл коэффициента регрессии состоит в том, что увеличение *числа рабочих* на 1 чел. приводит в среднем к увеличению объема годового *производства* на 2984 руб.

Свойства коэффициента регрессии

- Коэффициент регрессии может принимать любые значения.
- Коэффициент регрессии *не симметричен* , т.е. изменяется, если X и Y поменять местами.
- *Единицей измерения* коэффициента регрессии является отношение единицы измерения Y к единице измерения X : ($[Y] / [X]$).
- Коэффициент регрессии *изменяется при изменении единиц измерения* X и Y .
- Поскольку результативный признак Y измеряется в рублях, а факторный признак X в количестве рабочих (чел.), то коэффициент регрессии измеряется *в рублях на человека* (руб. / чел.)

Расчет линейной регрессии в Экселе

Пакет 'Анализ данных'

города_регрессия - Microsoft Excel + Analyse-it®

Рецензирование Вид Настройки Load Test Acrobat Analytic Solver Platform XLMiner Platform

Сортировка и фильтр | Фильтр | Очистить | Повторить | Дополнительно | Текст по столбцам | Удалить дубликаты | Проверка данных | Консолидация | Анализ "что если" | Группировать | Разгруппировать | Промежуточный итог | Анализ данных | Поиск решения | min z, x ≤ y, x = 2 | Show/Hide Model | Quick Solve | OpenSolver

	5	6	7
м на 1ч	среднемесячная заработная плата руб	кол. Преступлений на 100 чел.	кол. Образов. Учржд. На 100 чел.

Средства для анализа данных
Средства для анализа финансовых и научных данных.

Анализ данных

Инструменты анализа

- Ковариация
- Описательная статистика
- Экспоненциальное сглаживание
- Двухвыборочный F-тест для дисперсии
- Анализ Фурье
- Гистограмма
- Скользящее среднее
- Генерация случайных чисел
- Ранг и перцентиль
- Регрессия**

OK
Отмена
Справка

Исходные данные

города	Население т.ч	Плотность 1кв. Км на 1ч	среднемесячная заработная плата руб	кол. Преступлений на 100 чел.	кол. Образов. Учржд. На 100 чел.
Москва	11514.00	10588.00	38410.00	16.00	68.00
Санкт-Петербург	4848.00	3480.00	27189.00	13.00	72.00
Новосибирск	1473.00	2947.00	23374.00	29.00	83.00
Екатеринбург	1350.00	2489.00	23216.00	33.00	147.00
Н. Новгород	1250.00	3153.00	21821.00	35.00	84.00
Самара	1164.00	2152.00	20690.00	27.00	82.00
Омск	1154.00	1923.00	19317.00	17.00	86.00
Казань	1143.00	1865.00	19410.00	20.00	88.00
Челябинск	1130.00	2258.00	20510.00	26.00	87.00
Ростов на дону	1064.00	3127.00	21053.00	19.00	78.00
Уфа	1089.00	1518.00	22089.00	21.00	93.00
Волгоград	1021.00	1791.00	18294.00	17.00	81.00
Пермь	991.00	1239.00	22678.00	29.00	93.00
Красноярск	973.00	2754.00	25159.00	29.00	86.00
Воронеж	890.00	1633.00	18178.00	14.00	87.00
Саратов	837.00	2192.00	18107.00	18.00	155.00
Краснодар	744.00	991.00	22587.00	18.00	103.00

В рамках данного примера, в качестве **зависимой переменной (Y)** возьмем переменную ‘население’, в качестве независимых переменных (X) будем использовать все остальные переменные.

Население= $A_0 + a_1 * \text{плотн.} + A_2 * \text{сред. мес. зароб} + A_3 * \text{преступ.} + A_4 * \text{образ учрежд.}$

Параметры регрессии

Регрессия

Входные данные

Входной интервал Y:

населения!R2C3:R18C3



Входной интервал X:

R2C4:R18C7



Метки

Константа - ноль

Уровень надежности:

95 %

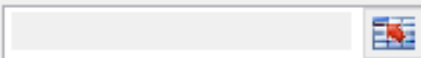
OK

Отмена

Справка

Параметры вывода

Выходной интервал:



Новый рабочий лист:



Новая рабочая книга

Остатки

Остатки

График остатков

Стандартизованные остатки

График подбора

Нормальная вероятность

График нормальной вероятности

Входной интервал Y – данные в колонке ‘население’.

Входной интервал X – все остальные данные.

Анализ результатов

<i>Регрессионная статистика</i>	
Множественный R	0.983413603
R-квадрат	0.967102314
Нормированный R-квадрат	0.956136419
Стандартная ошибка	553.0910588
Наблюдения	17
<i>Дисперсионный анализ</i>	
	<i>df</i>
Регрессия	4
Остаток	12
Итого	16
<i>Коэффициенты</i>	
Y-пересечение	-3375.125239
Переменная X 1	0.665945913
Переменная X 2	0.233697372
Переменная X 3	-80.87652333
Переменная X 4	0.530463184

R^2 - коэффициент детерминации, показывающий что на 74,5% расчетные параметры модели, то есть сама модель, объясняют зависимость и изменения изучаемого параметра - Y от исследуемых факторов - X . Можно сказать что, это показатель качества модели и чем он выше тем лучше. Понятное дело, что он не может быть больше 1 и считается неплохо, когда R^2 выше 0,8, а если меньше 0,5, то смысл такой модели можно смело ставить под большой вопрос.

Анализ результатов

<i>Регрессионная статистика</i>	
Множественный R	0.983413603
R-квадрат	0.967102314
Нормированный R-квадрат	0.956136419
Стандартная ошибка	553.0910588
Наблюдения	17
<i>Дисперсионный анализ</i>	
	<i>df</i>
Регрессия	4
Остаток	12
Итого	16
<i>Коэффициенты</i>	
Y-пересечение	-3375.125239
Переменная X 1	0.665945913
Переменная X 2	0.233697372
Переменная X 3	-80.87652333
Переменная X 4	0.530463184

Y пересечение - коэффициент который показывает какой будет Y в случае, если все используемые в модели факторы будут равны 0, подразумевается что это зависимость от других неописанных в модели факторов;

В данной модели использованы четыре переменных, соответственно, зеленым цветом выделены 4 коэффициента, которые характеризуют степень влияния независимых переменных на зависимую переменную Y.

$$A_0 = -3375.125239$$

$$A_1 = 0.665945913$$

$$A_2 = 0.665945913$$

$$A_3 = -80.87652333$$

Анализ результатов

ВЫВОД ОСТАТКА		
<i>Наблюдение</i>	<i>Предсказанное Y</i>	<i>Остатки</i>
1	11394.27327	119.7267305
2	4283.162931	564.8370695
3	1748.469007	-275.4690066
4	1116.985144	233.0148559
5	1037.993169	212.0068307
6	753.0208428	410.9791572
7	1090.539823	63.46017703
8	832.080172	310.919828
9	865.0744219	264.9255781
10	2132.040588	-1068.040588
11	1148.847992	-59.84799213
12	760.910235	260.089765
13	453.6846478	537.3153522
14	2038.682644	-1065.682644
15	874.3942347	15.60576534
16	942.63089	-105.63089
17	1162.209989	-418.2099889

Предсказанное Y – величины, которые получились в результате предсказания. Остатки – это разница между реальными данными и предсказанными, то есть

Остатки = Y – Y(пред)

Коэффициент детерминации R²

Коэффициент детерминации рассматривают, как правило, в качестве основного показателя, отражающего меру качества регрессионной модели, описывающей связь между зависимой и независимыми переменными модели. Коэффициент детерминации показывает, какая доля вариации объясняемой переменной y учтена в модели и обусловлена влиянием на нее факторов, включенных в модель:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

где – y_i значения наблюдаемой переменной, \bar{y} – среднее значение по наблюдаемым данным, \hat{y}_i – модельные значения, построенные по оцененным параметрам.

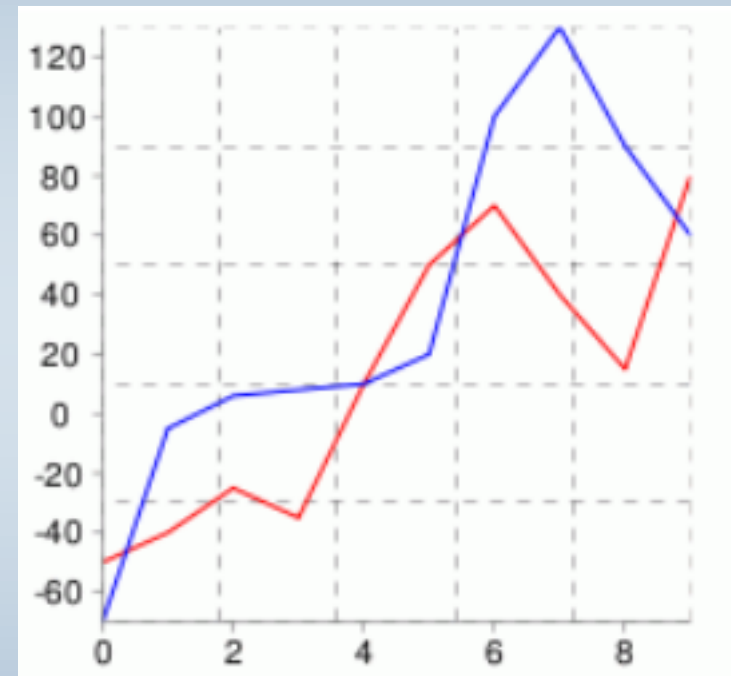
Достоинства и недостатки регрессионных моделей

Достоинства

1. Простота вычислительных алгоритмов.
2. Наглядность и интерпретируемость результатов (для линейной модели).

Недостатки

1. Невысокая точность прогноза
(в основном - интерполяция данных).
1. Субъективный характер выбора вида конкретной зависимости (формальная подгонка модели под эмпирический материал).
2. Отсутствие объяснительной функции (невозможность объяснения причинно-следственной связи).



Авторегрессия

Авто регрессионная модель описывает связь между переменной Y от самого себя, вернее от того каков был Y в прошлом периоде (день, месяц, год и т.п.)

Общий вид модели авторегрессии: $Y_i = a_0 + \sum a_i * Y_{i-1} + \varepsilon_i$

где a_0 — постоянная - коэффициент показывает каким будет итог модели в случае, когда все влияющие факторы равны нулю;

a_i — коэффициенты, которые описывают степень зависимости итогового Y от влияющих факторов, в данном случае, от того каким был Y в прошлом периоде регрессии;

Y_{i-1} — влияющие факторы, которые в данном случае и есть итоговый Y , но тот, каким он был раньше.

ε_i — случайная компонента или как еще ее принято называть погрешность модели (по сути, это разница между расчетным значением модели за известные периоды и между самими известными значениями, то есть $Y_{расч.} - Y$).

AR I - Авторегрессия первого порядка

$$Y_i = a_0 + a_i * Y_{i-1} + \varepsilon_i$$

Линейная модель авторегрессии первого порядка состоит только из одного влияющего фактора, а именно из Y_{i-1} , то есть изучается наиболее тесная зависимость только от того каким был итоговый показатель периодом с одним шагом назад.

Регрессия

Входные данные

Входной интервал Y: Лист1!\$E\$4:\$E\$10

Входной интервал X: Лист1!\$F\$4:\$F\$11

Метки

Константа - ноль

Уровень надежности: 95 %

Параметры вывода

Выходной интервал:

Новый рабочий лист:

Новая рабочая книга

Остатки

Остатки

Стандартизованные остатки

График остатков

График подбора

Нормальная вероятность

График нормальной вероятности

OK

Отмена

Справка

AR I - Авторегрессия первого порядка

Регрессионная статистика

Множественный R	0.971302904
R-квадрат	0.943429332
Нормированный R-квадрат	0.934000887
Стандартная ошибка	90751.79091
Наблюдения	8

Дисперсионный анализ

	<i>df</i>
Регрессия	1
Остаток	6
Итого	7

Коэффициенты

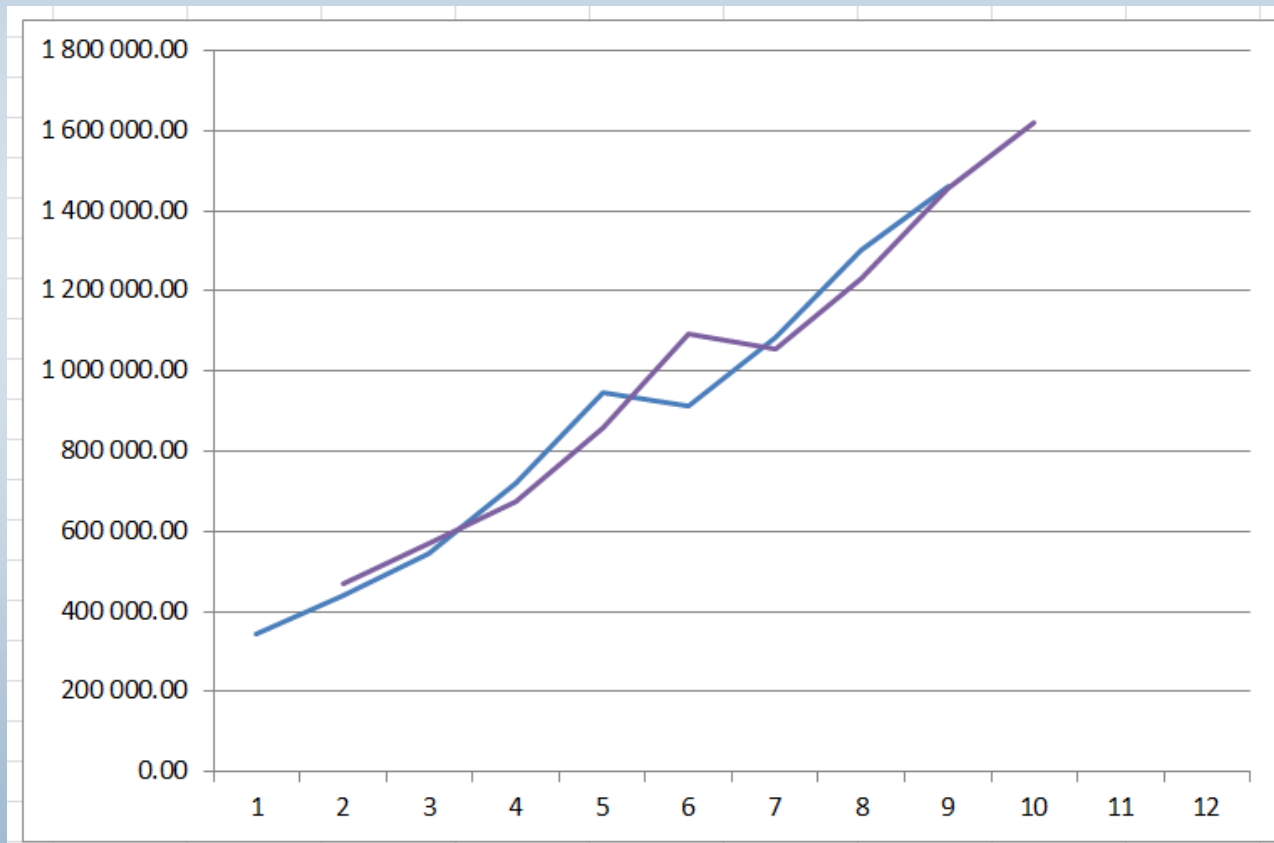
Y-пересечение	113436.6764
Переменная X 1	1.032789147

То есть у нас получилась модель:

$$Y = 113436,67 + 1,033 * Y_{i-1}$$

Прогноз на основе авторегрессии первого порядка

$$Y = 113436,67 + 1,033*Y_{i-1}$$



Согласно прогнозу ВВП растет.....

AR II - Авторегрессия второго порядка

$$Y_i = a_0 + a_1 * Y_{i-1} + a_2 * Y_{i-2} + \varepsilon_i$$

Модель авторегрессии второго порядка отличается от первой тем, что она включает в себя еще один влияющий фактор Y_{i-2} , то есть показывается зависимость от того каким был Y не только один период назад, но и от того каким он был два периода назад.



$$Y = 151395,987 + 0,724 * Y_{i-1} + 0,32 * Y_{i-2}$$

AR II - Авторегрессия второго порядка

<i>Регрессионная статистика</i>	
Множественный R	0.963118984
R-квадрат	0.927598177
Нормированный R-квадрат	0.891397266
Стандартная ошибка	104620.4471
Наблюдения	7
<i>Дисперсионный анализ</i>	
	<i>df</i>
Регрессия	2
Остаток	4
Итого	6
<i>Коэффициенты</i>	
Y-пересечение	151395.987
Переменная X 1	0.724077333
Переменная X 2	0.320347509

Какая модель лучше, регрессия первого порядка или второго порядка?

С чем связано изменение качества модели?

Достоинства и недостатки авторегрессионных моделей

ПЛЮСЫ:

1. Получение высококачественной модели с адекватным прогнозом при минимуме временных затрат и требований к исходным данным.

МИНУСЫ:

1. Прогноз по исходным данным возможен только на один период вперед. Если нужно сделать прогноз на более длительный срок, то в качестве влияющих факторов для расчета придется брать не реально существующий Y , а тот который рассчитан по модели, что в итоге даст прогноз на прогнозе, а значит адекватность такого прогноза, как минимум, в два раза меньше.

2. С увеличением разрядности авторегрессии возникает необходимость расширять диапазон исходных данных.



**INTERNET
STUDIES LAB**



HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
SAINT PETERSBURG