

Модели статистического прогнозирования (11класс)



1 2
4 5

Рассмотрим способ нахождения зависимости частоты заболеваемости жителей города бронхиальной астмой от качества воздуха.

0011 011

Качество воздуха в городе(чем хуже воздух, тем больше больных астмой)

Качественное заключение. Его не достаточно для того чтобы управлять уровнем загрязнённости воздуха

Определим характер зависимости

Частота легочных заболеваний

Нужно установить, какие именно примеси сильнее всего влияют на здоровье людей, как связана концентрация этих примесей в воздухе с числом заболеваний. Такую зависимость можно установить только экспериментальным путём: посредством сбора многочисленных данных, их анализа и обобщения.

Статистика- наука о сборе, измерении и анализе массовых количественных данных

Виды статистики:

- медицинская статистика
- математический аппарат статистики разрабатывает наука под названием **математическая статистика**
- экономическая статистика
- социальная статистика ...

Зависимости устанавливаются экспериментальным путем:

-сбор данных;

- анализ;

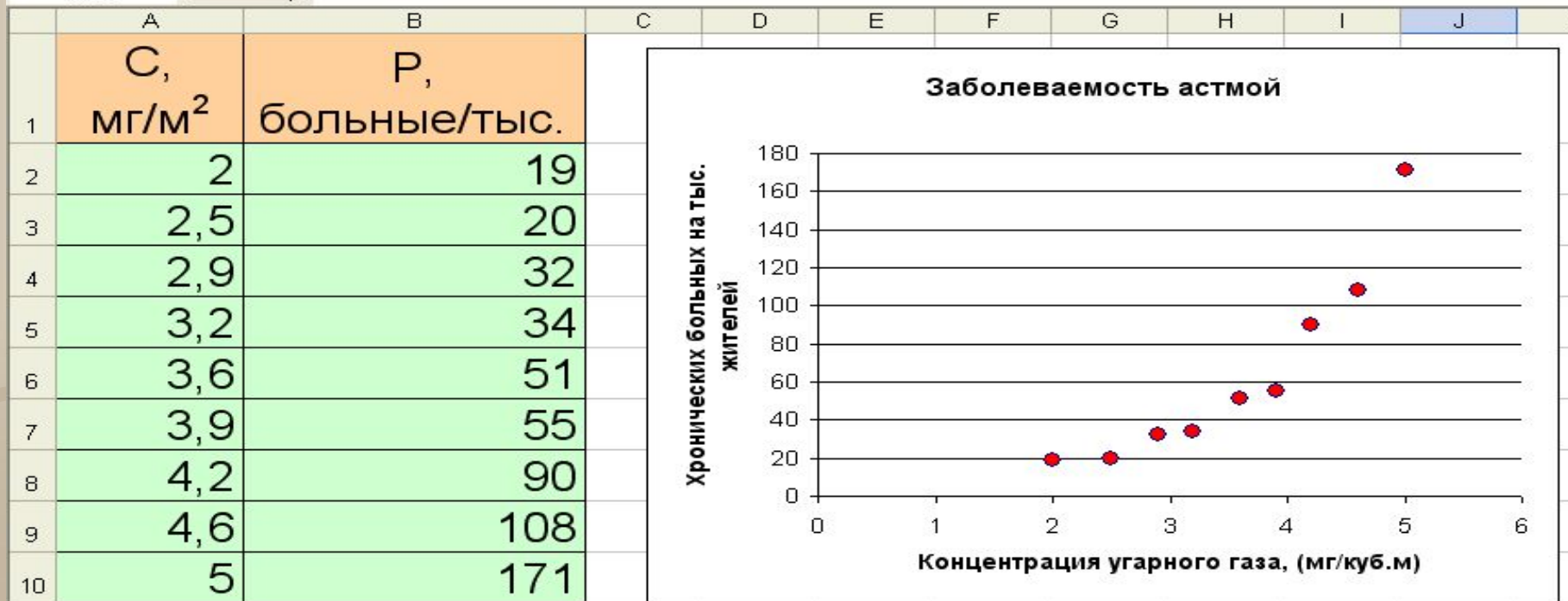
- обобщение.



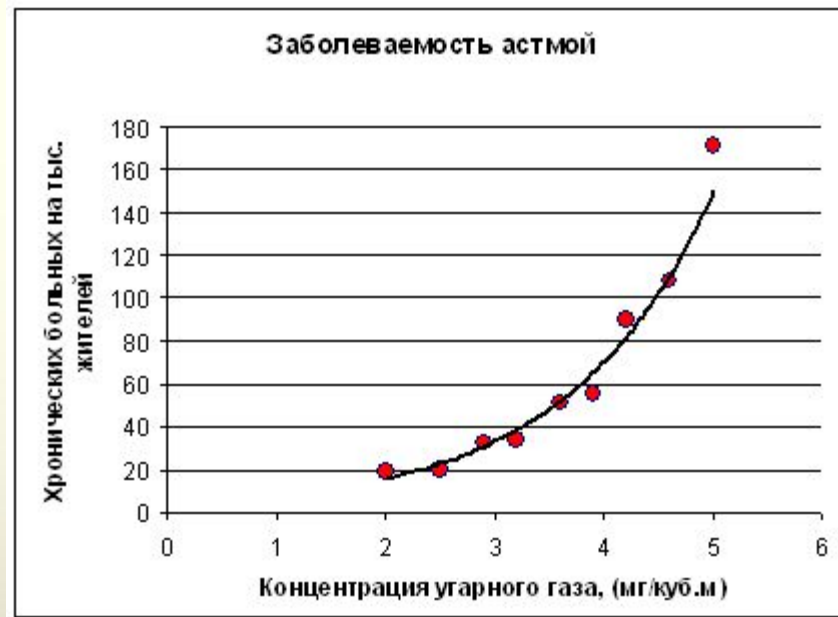
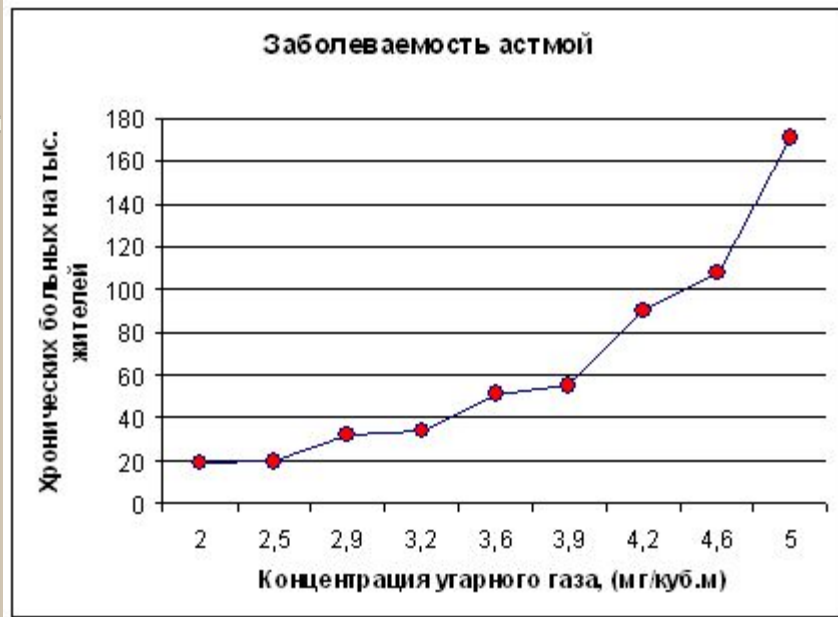
Рассмотрим пример из области медицинской статистики:

Известно, что наиболее сильное влияние на бронхиально - легочные заболевания оказывает угарный газ –оксид углерода. Специалисты по медицинской статистике проводят сбор данных.

Сведения о средней концентрации угарного газа в атмосфере (С) и о заболеваемости астмой (число хронических больных на 1000 жителей (Р) можно свести в таблицу и представить в виде точечной диаграммы.



2 варианта построения графической зависимости по экспериментальным данным



Основные требования к искомой функции:

- она должна быть **достаточно простой** для использования ее в дальнейших вычислениях;
- график функции должен проходить **вблизи экспериментальных точек** так, чтобы отклонения этих точек от графика были минимальны и равномерны.

Полученную таким образом функцию называют в статистике **регрессионной моделью**.

Получение регрессивной модели происходит в два этапа:

1) подбор вида функции:

$y = ax + b$ - линейная функция;

$y = ax^2 + bx + c$ - квадратичная функция (полиномиальная);

$y = a \ln(x) + b$ - логарифмическая функция;

$y = ae^{bx}$ - экспоненциальная функция;

$y = ax^b$ - степенная функция.

Во всех этих формулах x - аргумент, y - значение функции, a, b, c, d - параметры функции, $\ln(x)$ - натуральный логарифм, e - константа, основание логарифма.

2) вычисление параметров функции:

метод наименьших квадратов (МНК) - сумма квадратов отклонений y -координат всех экспериментальных точек от y -координат графика функции должна быть минимальной.

Метод наименьших квадратов (МНК) был предложен в XVIII веке немецким учёным математиком К.Гауссом.

Карл Гаусс (1777-1855)

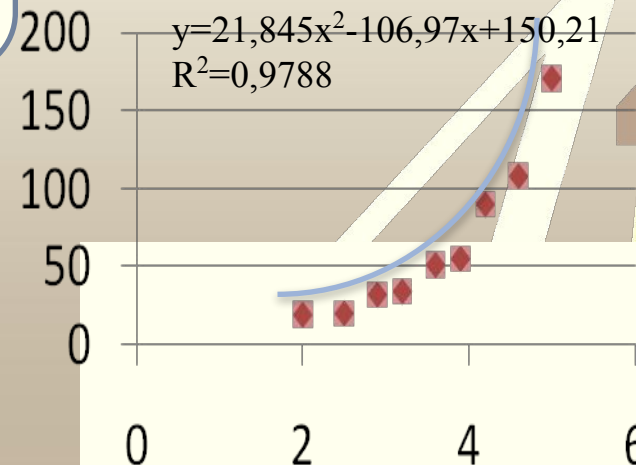
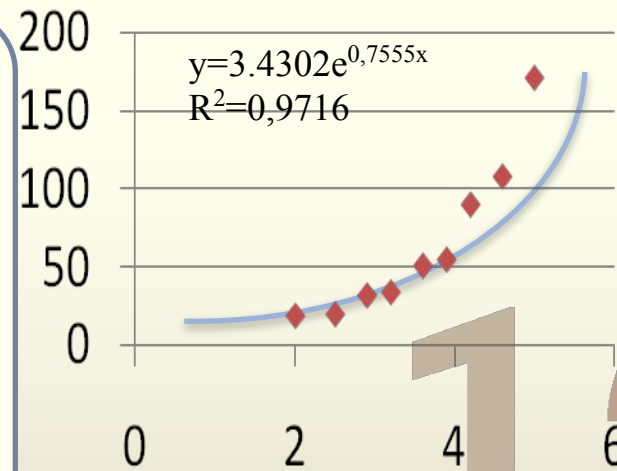
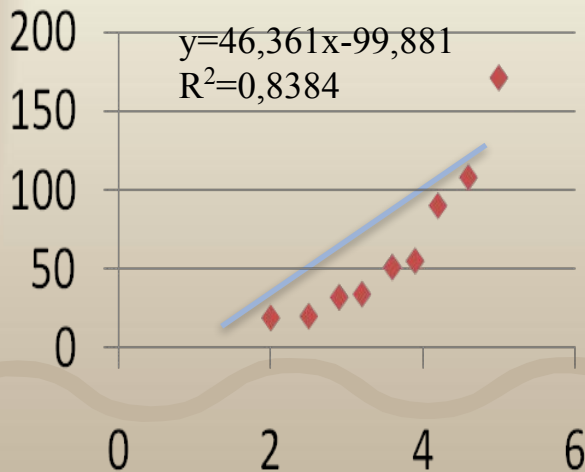


- Его математическое дарование проявилось уже в детстве. Рассказывают, что в 3-ёхлетнем возрасте он удивил окружающих, поправив расчёты своего отца с каменщиками. Однажды в школе (Гауссу в то время было 10 лет) учитель предложил классу сложить все числа от одного до ста. Пока он диктовал задание, у Гаусса уже был готов ответ. На его грифельной доске было написано: $101 * 50 = 5050$

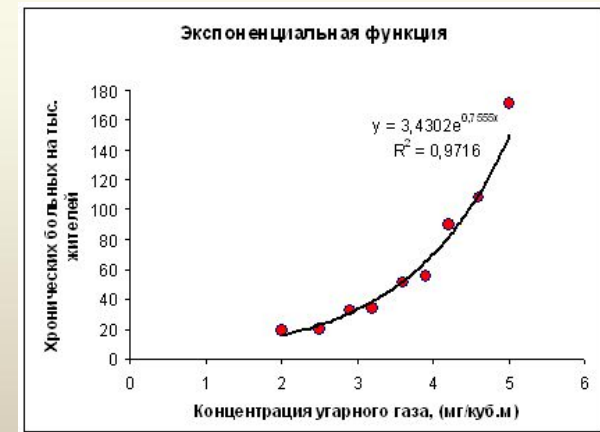
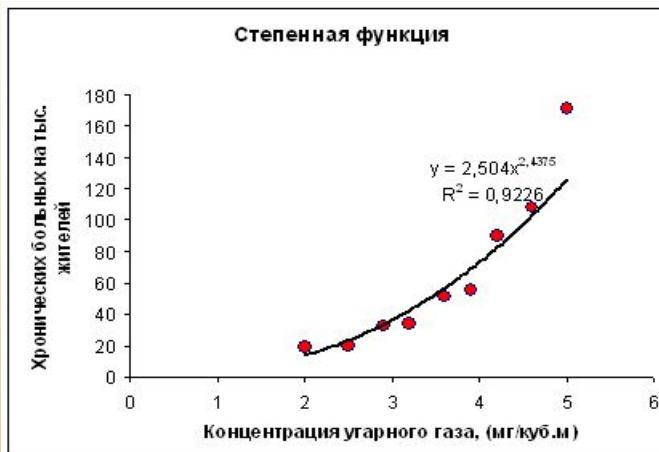
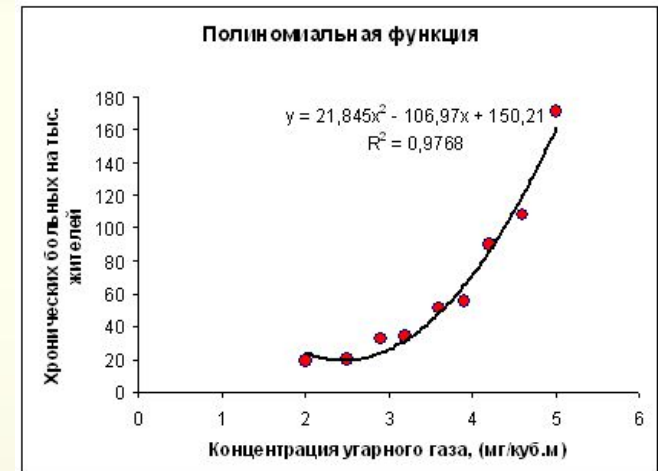
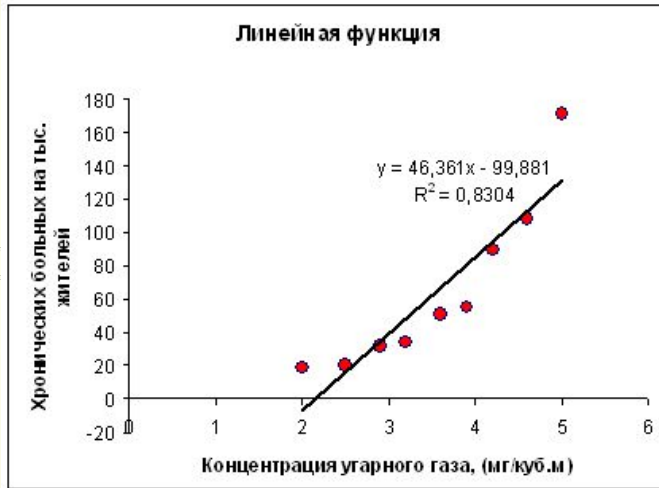
Метод наименьших квадратов

- $y=ax+b$ – линейная функция;
- $y=ax^2+bx+c$ – квадратичная функция;
- $y=a \ln(x)+b$ – логарифмическая функция;
- $y=ae^{bx}$ – экспоненциальная функция;
- $y=ax^b$ – степенная функция;
- $y=ax^3+bx^2+cx+d$ – полином 3 степени.

График регрессивной модели называется **ТРЕНДОМ** (англ. “trend”) – общее направление или тенденция



Графики функций, построенные по МНК, - тренды



Обратим внимание на подписи, присутствующие на графиках.

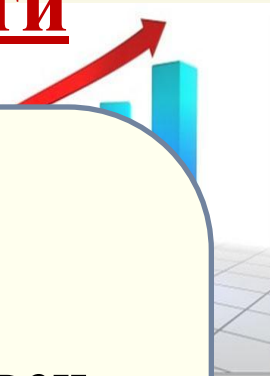
Во-первых, это записанные в явном виде искомые функции –регрессивные модели. На графиках присутствует ещё одна величина, полученная в результате построения трендов. Она обозначена как R^2 . В статистике эта величина называется

Коэффициентом детерминированности, который всегда заключён в диапазоне от 0 до 1.

Коэффициент детерминированности

R^2 – коэффициент детерминированности (определяет, насколько удачной является полученная регрессионная модель). Если он равен 1, то функция точно проходит через табличные значения, если 0, то выбранный вид регрессивной модели предельно неудачен.

Чем R^2 ближе к 1, тем удачнее регрессивная модель.



45

Алгоритм построения регрессионной модели по МНК с помощью MS Excel (линейный тренд)

- Ввести табличные данные зависимости заболеваемости **P** от концентрации угарного газа **C**.
- Построить точечную диаграмму. (В качестве подписи к оси OX выбрать название тренда - «Линейный», остальные надписи и легенду можно игнорировать).
- Щелкнуть мышью по полю диаграммы; выполнить команду **Диаграмма – Добавить линию тренда**;
- В открывшемся окне на вкладке **Тип** выбрать **Линейный** тренд;
- Перейти на вкладку **Параметры** и установить галочки на флажках **показывать уравнения на диаграмме** и **поместить на диаграмме величину достоверности аппроксимации R^2**
- щелкнуть ОК.

Построение регрессионной модели по МНК с помощью MS Excel 2007 (линейный тренд)

The screenshot displays the Microsoft Excel 2007 interface. The 'Макет' (Layout) ribbon is active, with the 'Линия тренда' (Trendline) button circled in red. The 'Формат линии тренда' (Format Trendline) task pane is open, showing the 'Линейная' (Linear) trendline option selected. The chart, titled 'Заболеваемость астмой', plots 'Хронических больных на тыс. жителей' (Chronic patients per thousand residents) on the y-axis against 'концентрация угарного газа (мг/куб.м)' (carbon monoxide concentration in mg/cubic meter) on the x-axis. A linear regression line is shown with the equation $y = 46,36x - 99,88$ and $R^2 = 0,830$. The data points are as follows:

С, мг/м3	Р, бол./тыс.
2	19
2,5	20
2,9	32
3,2	34
3,6	51
3,9	55
4,2	90
4,6	108
5	171

Практическая работа 3.1

Получение регрессионных моделей в MS Excel

0011 0010 1010 1101 0001 0100 1011

Цель работы: освоение способов построения по экспериментальным данным регрессионной модели и графического тренда средствами табличного процессора MS Excel.

Семакин И.Г. Практикум. Информатика и ИКТ 11 кл., стр.209



Имея регрессивную модель , легко прогнозировать, производя расчёты с помощью электронных таблиц

0011 0010 1010 1101 0001 0100 1011

12
42
5

Имея регрессивную модель , легко прогнозировать, производя расчёты с помощью электронных таблиц

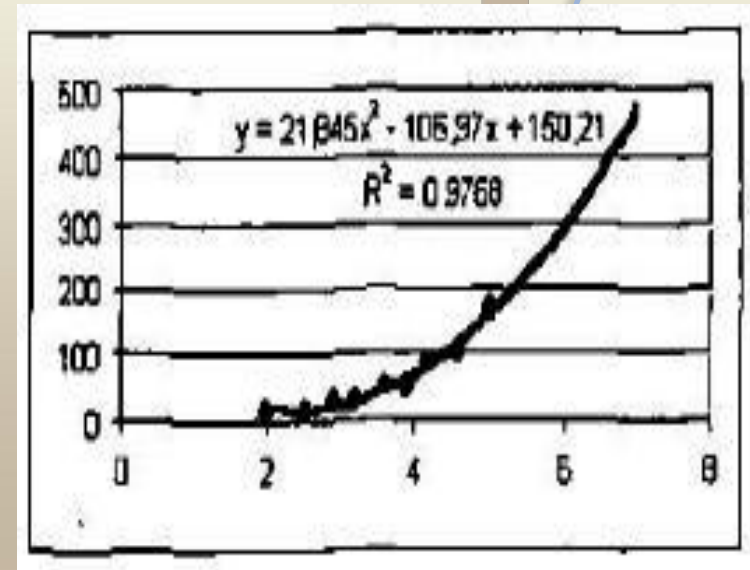
	A	B
1	Концентрация угарного газа (мг/куб.м)	Число больных астмой на 1 тыс. жителей
2		$=21,845*A2^2-106,97*A2+150,21$
3		

	A	B
1	Концентрация угарного газа (мг/куб.м)	Число больных астмой на 1 тыс. жителей
2	3	26
3		

	A	B
1	Концентрация угарного газа (мг/куб.м)	Число больных астмой на 1 тыс. жителей
2	7	472
3		



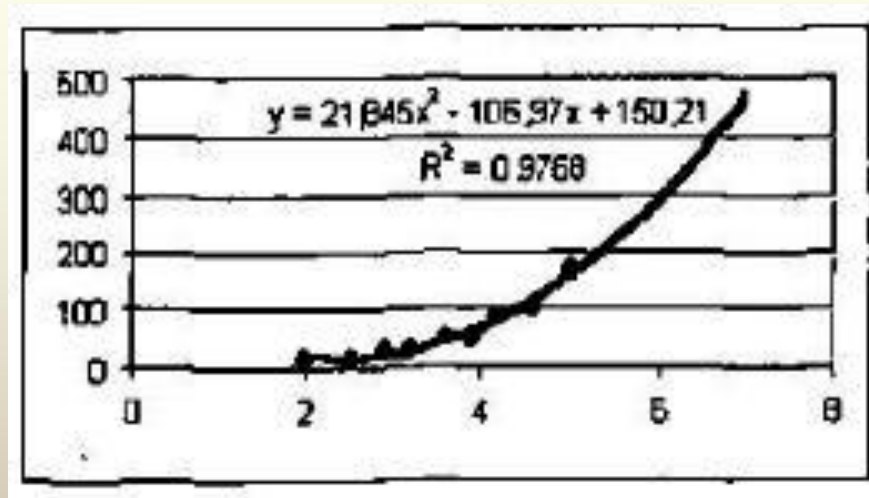
Табличный процессор даёт возможность производить экстраполяцию графическим способом, продолжая тренд за пределы экспериментальных данных. Как это выглядит при использовании квадратичного тренда для $S=7$ показано на графике.



ПРОГНОЗИРОВАНИЕ ПО РЕГРЕССИВНОЙ МОДЕЛИ

Существует два способа прогнозирования по регрессивной модели:

- **Восстановление значений** – прогноз в пределах экспериментальных значений независимой переменной.



- **Экстраполяция** – прогнозирование за пределами экспериментальных данных



Ограничения при экстраполяции !

Применимость всякой регрессионной модели ограничена, особенно за пределами экспериментальной области т.к. экстраполяция строится на гипотезе.

Вывод: применять экстраполяцию можно только в областях данных, близких к экспериментальной.



Практическая работа 3.2

Прогнозирование в MS Excel

0011 0010 1010 1101 0001 0100 1011

Цель работы: освоение приемов прогнозирования количественных характеристик системы по регрессионной модели путем восстановления значений и экстраполяции

Семакин И.Г. Практикум. Информатика и ИКТ 11 кл., **стр.211**



Типичные проблемы на этапе стратегического прогнозирования

1. Некритическая экстраполяция текущей ситуации и тенденций на долгосрочную перспективу, в частности – формальный трендовый подход.
2. Несоответствие прогнозов выявленным на этапе анализа ограничениям и тенденциям
3. Отказ учитывать в прогнозе существенные факторы в связи с тем, что их влияние трудно выразить количественно.
4. Отказ от «плохих» прогнозов. Например, если получается, что мы становимся убыточными, значит прогноз неправильный и его надо «улучшить»
5. Внутреннее несоответствие между отдельными прогнозами. Например, прогнозируемый рост доходов не сопровождается ростом расходов (инвестиций).
6. Некритичное использование статистики, прогнозных данных из госпрограмм и «заинтересованных» источников. Отсутствие «встречных проверок».
7. Полученные результаты прогнозирования интерпретируются формально, не делается важных для развития компании выводов и предположений.

Используемая литература

- И.Г.Семакин и др. Информатика 11. Практикум, М.: Бином. Лаборатория знаний, 2014
- И.Г.Семакин и др. Информатика 11. Базовый уровень, М.: Бином. Лаборатория знаний, 2014



0011 0010 1010 1101 0001 0100 1011

Спасибо за внимание!

1
2
4
5