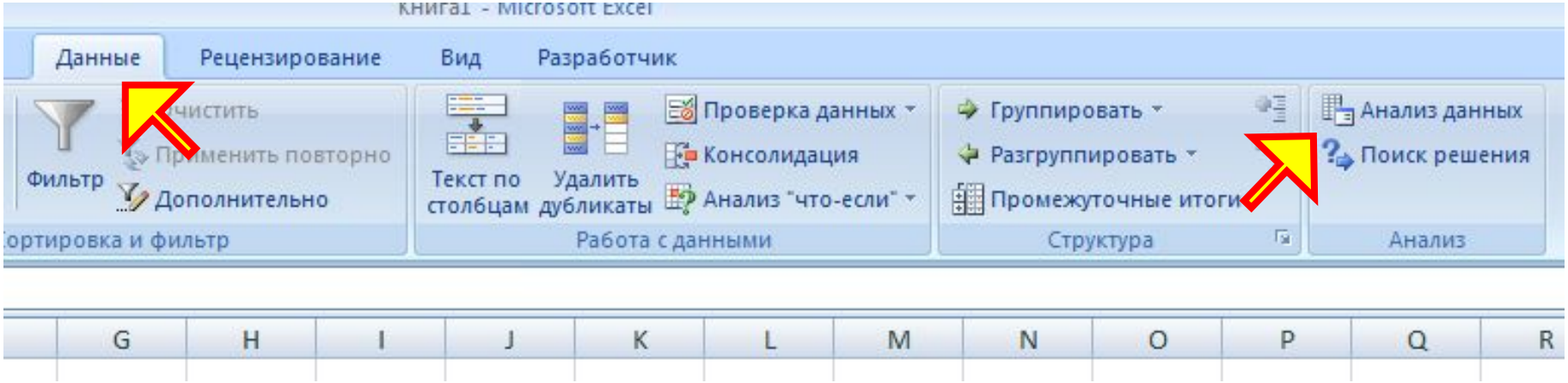
Русская Служба Оценки

Использование Microsoft Excel для построения регрессионных зависимостей

Установка «Пакета анализа» (Office 2007)



Если «Пакет анализа» и «Поиск решения» не установлен:

- Щелкните значок **Кнопка Microsoft Office** , затем **Параметры Excel**
- Выберите команду **Надстройки**
- В окне **Управление** выберите пункт **Надстройки Excel**
- Нажмите кнопку **Перейти**.
- Установите флажки **Пакет анализа** и **Поиск решения**
- Нажмите **ОК**

Регрессионный анализ

Регрессионный анализ - раздел математической статистики, объединяющий практические методы исследования регрессионной зависимости между величинами по статистическим данным. Цель Регрессионного анализа состоит в определении общего вида уравнения регрессии, построении оценок неизвестных параметров, входящих в уравнение регрессии, и проверке статистических гипотез о регрессии. …

БСЭ

$$Y = f(X) + \xi$$

Y – зависимая переменная (отклик)
X – независимые переменные (факторы, параметры, предикторы, признаки)
ξ – случайная величина (ошибка эксперимента)

$[y_i; x_{i1}; x_{i2}; \dots x_{1m}]$ – наблюдение (данные по i-му аналогу)
n – объем выборки (количество наблюдений)
m – число факторов

y1	x11	x12	…	x1m
y2	x21	x22	…	x2m
y3	x31	x32	…	x3m
…	…	…	…	…
yn	xn1	xn2	…	xnm

$$y = a_1 * x_1 + a_2 * x_2 + \dots + a_m * x_m + c$$

Независимые переменные

Предикторы

Количественные

*Значение количественной переменной выражается **числом***

- ✔ Площадь
- ✔ Мощность двигателя
- ✔ Объем емкости
- ✔ Масса станка
- ✔ Выручка компании
- ✔ И т.д.

Качественные

*Значение качественной переменной выражается **текстовым описанием**, рисунком или каким-либо другим поясняющим его смысл способом*

- ✔ Район города
- ✔ Тип двигателя
- ✔ Материал
- ✔ Уровень отделки
- ✔ Класс объекта
- ✔ И т.д.

Качественные** переменные могут «маскироваться» под **количественные:

Этаж расположения - 1. «первый», «последний», «средние этажи» или 2. «крайние этажи» и «средние этажи»

Оцифровка качественных параметров:

замена бинарными признаками

W вариантов значений

Качественный параметр
Класс A
Класс B+
Класс B-
Класс C
Класс D

Бинарные признаки			
Класс A	Класс B+	Класс B-	Класс C
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	0	0

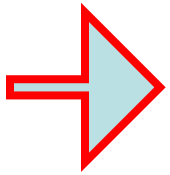
(w-1)
Бинарных признаков

«-» увеличение числа переменных

«+» нет необходимости в оптимизационных процедурах

Оцифровка качественных параметров: замена порядковыми переменными

Качественный параметр	Порядковый параметр
Класс A	4
Класс B+	3
Класс B-	2
Класс C	1
Класс D	0



«+» не увеличивает число переменных

«-» обычно требуется проведение оптимизационных процедурах

Оцифровка качественных параметров:
ранжирование по внешним данным

Качественный параметр	Параметр	Арендная ставка*, долл. / кв. м
Класс А	3,08	770
Класс В+	1,92	480
Класс В-	1,40	350
Класс С	1,20	300
Класс D	1,00	250

* - R-Way, №171 июнь 2009 г.

«+» не увеличивает число переменных

«-» необходимость использования (поиска) внешних данных

Взаимовлияние качественных параметров

Квартиры на первом этаже обычно ***дешевле*** аналогичных квартир на других этажах

Но: Квартиры на первом этаже в центральном районе могут быть ***дороже*** аналогичных квартир на других этажах

Варианты решения

Переменная ***этаж***:
- «первый этаж в периферийных районах» (1)
- «последний этаж» (2)
- «средние этажи» (3)
- «первый этаж в центральных районах» (4)

Переменная ***этаж***:
- «первый этаж» (1)
- «последний этаж» (2)
- «средние этажи» (3)

+

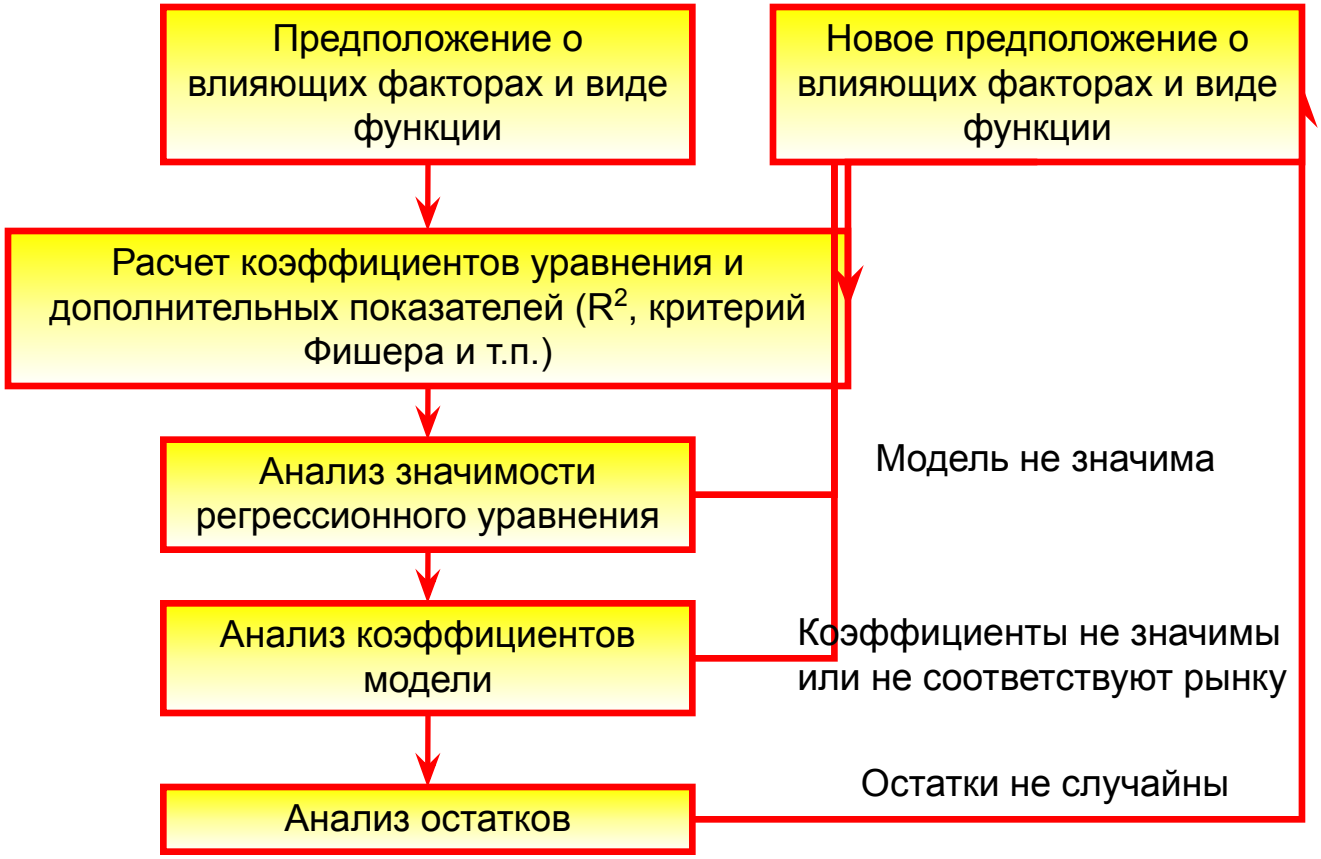
Переменная ***1-й этаж в центре***:
- «да» (1)
- «нет» (0)

Алгоритм действий

Пошаговый регрессионный анализ:

1. Последовательное исключение в модели незначительных переменных

2. Последовательное включение в модель переменных



Предположение о влияющих факторах и виде функции

- В качестве зависимой переменной лучше выбрать не «Стоимость объекта», а «Удельную стоимость»
- Корреляционная матрица поможет выбрать влияющие параметры (а также выделить взаимозависимые факторы)
- Графики $Y\text{-}X_i$ для количественных переменных могут помочь определить вид зависимости
- Переменные-агрегаты могут уменьшить число переменных и/или исключить мультиколлинеарность:

… Выбор единиц сравнения должен быть обоснован оценщиком… (ФСО-1, п. 22а)

Вместо «Площадь» и «Площадь ЗУ» – «Плотность застройки»

Вместо геометрических размеров – «Объем»

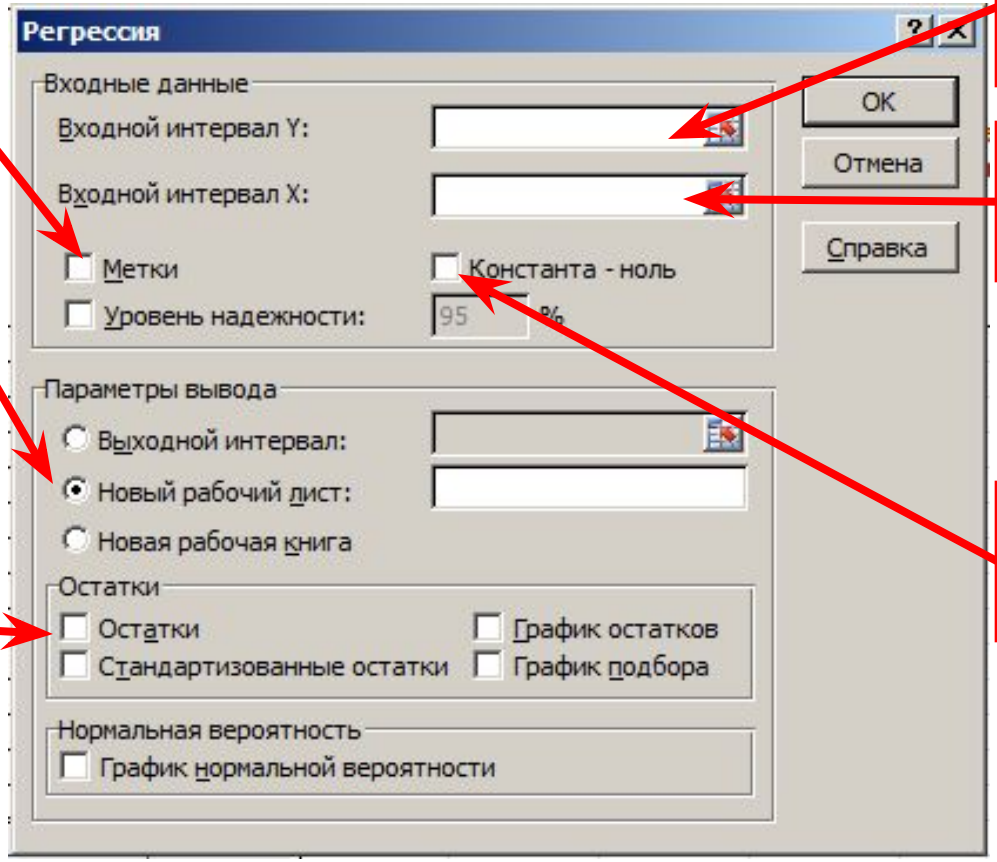
Вместо «Диаметр трубы», «Толщина стенки» и «Давление» – «Масса металла»

Пакет Анализа: «Поехали…»

Поставить «X», если в первой строке диапазонов включены названия

Указать место, куда следует поместить результаты

Указать необходимость расчета остатков и других показателей (обязательно отметить остатки)



Ссылка на диапазон зависимых переменных

Ссылка на диапазон независимых переменных

диапазоны д.б. непрерывными!!!

Поставить «X», если не нужно учитывать константу

Регрессионная статистика и Дисперсионный анализ

Шкала Чеддока	
R^2	Характеристика силы связи
0,1-0,3	Слабая
0,3-0,5	Умеренная
0,5-0,7	Заметная
0,7-0,9	Высокая
0,9-0,99	Весьма высокая

	A	B	C	D	E	F
1	ВЫВОД ИТОГОВ					
2						
3	*Регрессионная статистика*					
4	Множественный R	0,9712				
5	R-квадрат	0,9432				
6	Нормированный R-квадрат	0,9170				
7	Стандартная ошибка	0,1951				
8	Наблюдения	20				
9						
10	Дисперсионный анализ					
11		*df*	*SS*	*MS*	*F*	*Значимость F*
12	Регрессия	6	8,2234	1,3706	35,9974	0,0000
13	Остаток	13	0,4950	0,0381		
14	Итого	19	8,7184			

Критерий Фишера или F-критерий

$F_{расч} > F_{крит}$

$F_{крит}$=FРАСПРОБР(α;m;n-m-1)

Вероятность признать влияние факторов значимым при отсутствии такового влияния. Должна быть меньше стандартных уровней доверительной вероятности (например, 0,05).

Несколько важных замечаний про R^2

Коэффициент детерминации R^2 - оценка качества ("объясняющей способности") уравнения регрессии, показывает долю объясненной дисперсии зависимой переменной y.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y)^2}$$

y_i - наблюдаемое значение зависимой переменной y,
$\hat{y}_i$ - значение зависимой переменной, предсказанное по уравнению регрессии,
y - среднее арифметическое зависимой переменной.

Высокое значение R^2 не свидетельствует о хорошем качестве модели.
Низкое значение R^2 может объясняться не включением в модель существенных факторов.

Показатели **R^2 в разных моделях** с разным числом переменных и/ или наблюдений **не сравнимы**

Коэффициент детерминации нормированный – скорректированный на число степеней свободы.
Скорректированный R^2 ограниченно сравним в разных моделях (с разным набором факторов и/или наблюдений)

$$R^2_{скор} = 1 - (1 - R^2) * \frac{n - 1}{n - m - 1}$$

R^2 - коэффициент детерминации;
m - число переменных, вошедших в модель
n - число наблюдений

Анализ коэффициентов модели

Искомые коэффициенты модели. Должны соответствовать «**рыночным реалиям**»
Проверяем знаки коэффициентов!!!

Распределение Стьюдента (t-статистика). $t_{расч} > t_{крит}$
$t_{крит}$ = СТЬЮДРАСПОБР(α;n-m-1)

Верхняя и нижняя границы доверительного интервала при заданном уровне вероятности.
Должны быть одного знака.

17		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
18	Y-пересечение	15,1788	0,5244	28,9438	0,0000	14,0459	16,3118	14,0459	16,3118
19	Количественный показатель 1	-0,0471	0,0242	-1,9459	0,0736	-0,0994	0,0052	-0,0994	0,0052
20	Местоположение (3 - Район 1, 2 - Район 2, 1 - Район 3)	0,3627	0,0588	6,1670	0,0000	0,2356	0,4897	0,2356	0,4897
21	Количественный показатель 2	-0,8785	0,1304	-6,7351	0,0000	-1,1603	-0,5967	-1,1603	-0,5967
22	Бинарный признак 1	0,3701	0,0989						
23	Бинарный признак 2	0,2868	0,1007						

Сравнивая коэффициент с его стандартной ошибкой можно судить о его значимости. Критических значений нет. Используется **t-статистика.**

Показывает вероятность того, что t-статистика может оказаться больше наблюдаемой.
Если P-Значение меньше α, то коэффициент значим на уровне α.
Должно быть меньше стандартных уровней доверительной вероятности (например, 0,05).

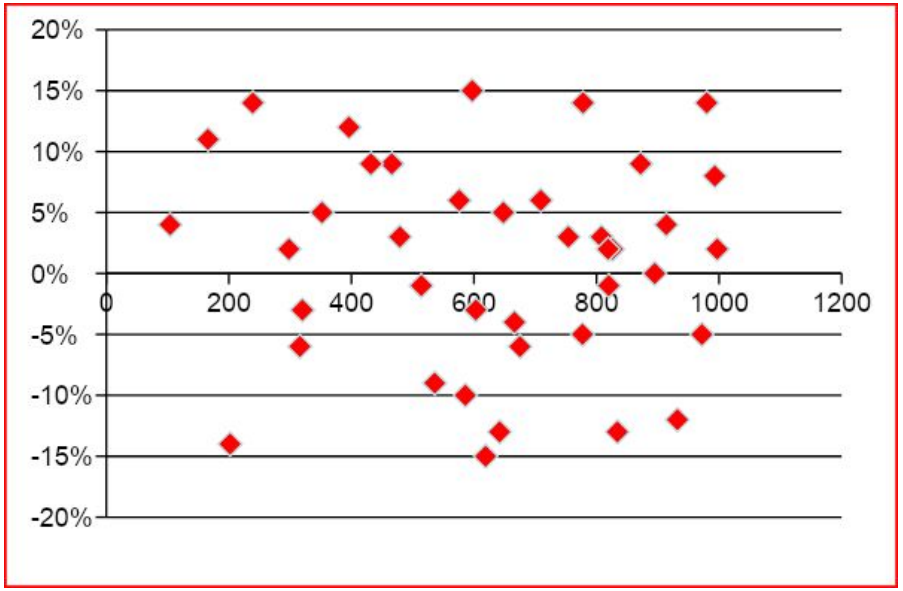
Анализ остатков

- Остатки имеют нулевое среднее
- Зависимая переменная не коррелированна с остатками
- Наблюдаемые значения остатков не коррелированны друг с другом
- Остатки имеют постоянную дисперсию
- Остатки распределены нормально

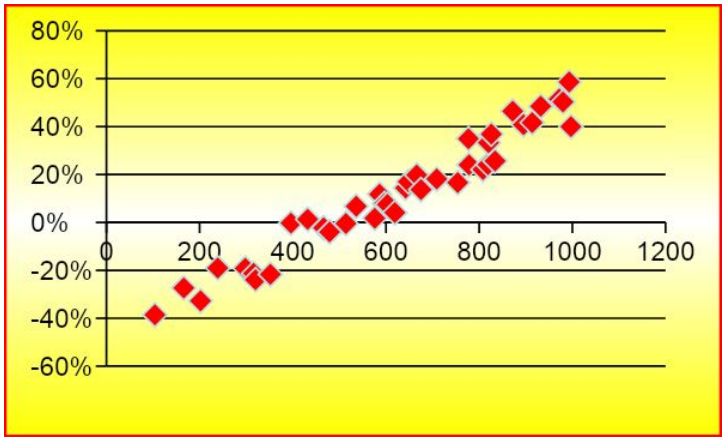
Строим график:

Ось абсцисс: $y_ф$
(фактическое значение)

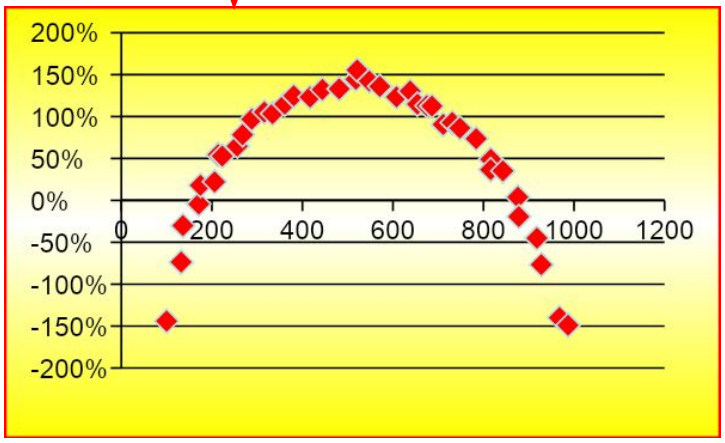
Ось ординат: $(y_{пр} – y_ф)/ y_ф$
(относительные остатки)



Анализ остатков

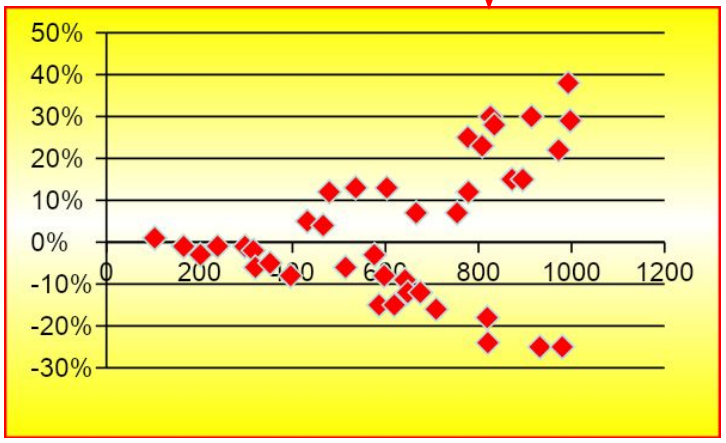


Зависимость не линейна по одному из параметров



Не учтена влияющая переменная

Рост дисперсии Гетероскедастичность



Использование функции ЛИНЕЙН()

Порядок использования:

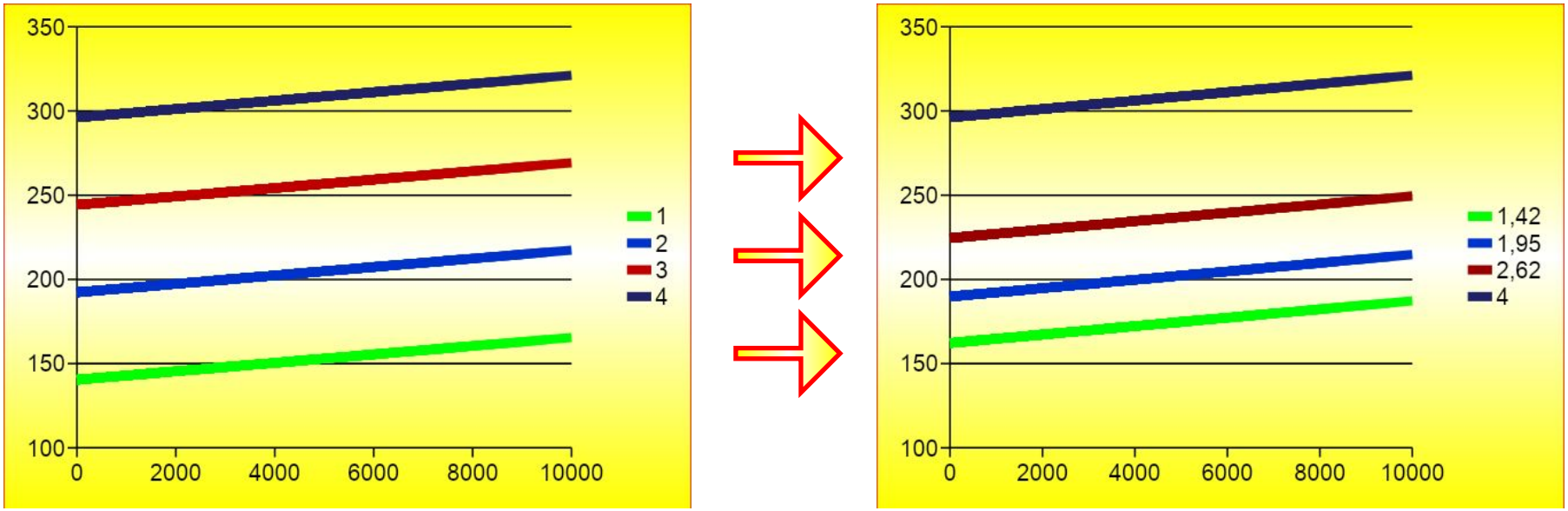
- Подготовить данные для расчетов;
- Выделить диапазон размером [5 строчек] X [m+1 колонка]
  (m – количество переменных);
- Нажать F2, ввести функцию;
- Нажать Ctrl+Shift+Enter

Синтаксис функции:

=ЛИНЕЙН(изв.y; изв.x; конст.; статистика)

изв.y - ссылка на диапазон с известными Y;
изв.x - ссылка на диапазон с известными X;
конст. - логическое значение: ИСТИНА (1) – учитывать константу обычным образом; ЛОЖЬ (0) – константа равна нулю;
статистика - логическое значение: ИСТИНА (1) – рассчитывается дополнительная статистика; ЛОЖЬ (0) – рассчитываются только коэффициенты и константа.

Оптимизация



Алгоритм:

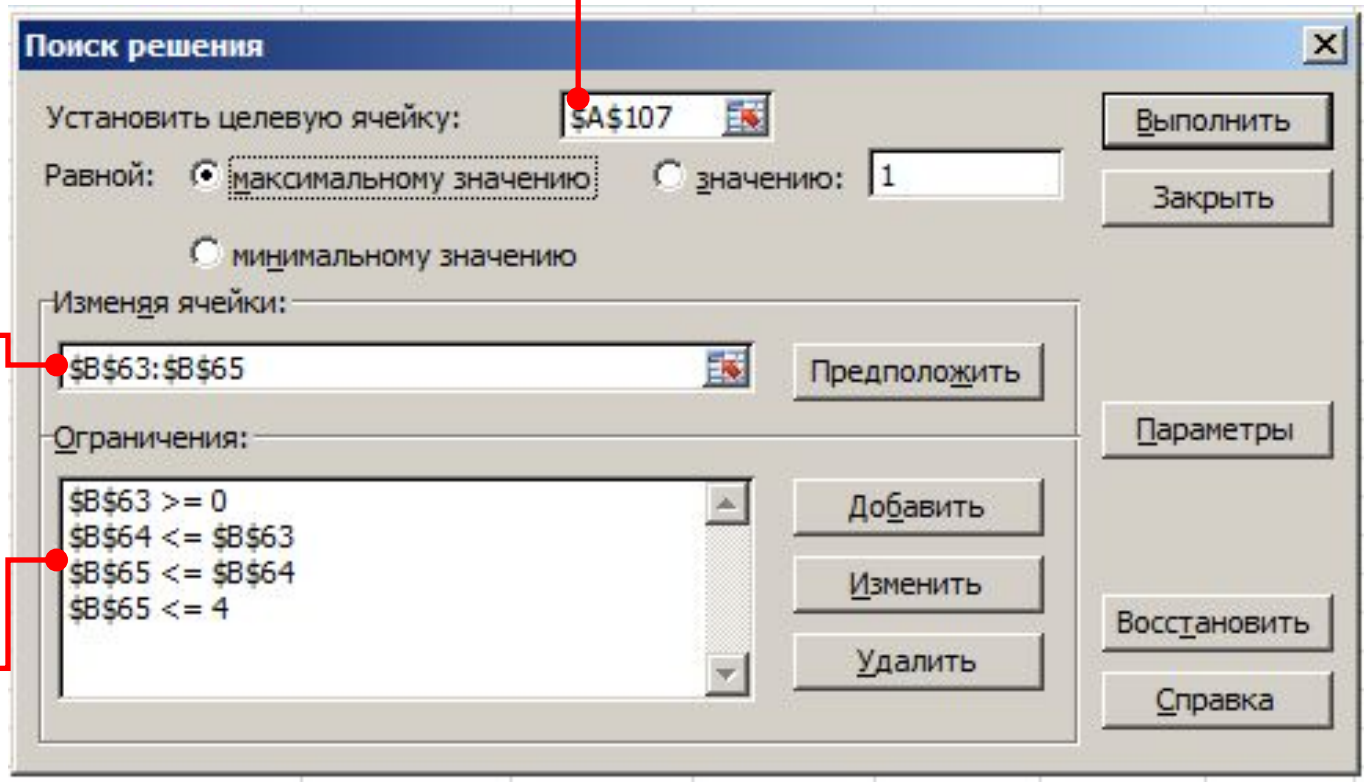
- Оцифровку качественных параметров оформить в виде ссылок на «диапазон меток»;
- Рассчитать коэффициенты и статистику при помощи функции ЛИНЕЙН;
- При помощи надстройки Excel «Поиск решения» подобрать метки, максимизируя R^2.

Оптимизация

Ссылка на коэффициент детерминации R^2

Ссылка на «диапазон меток»

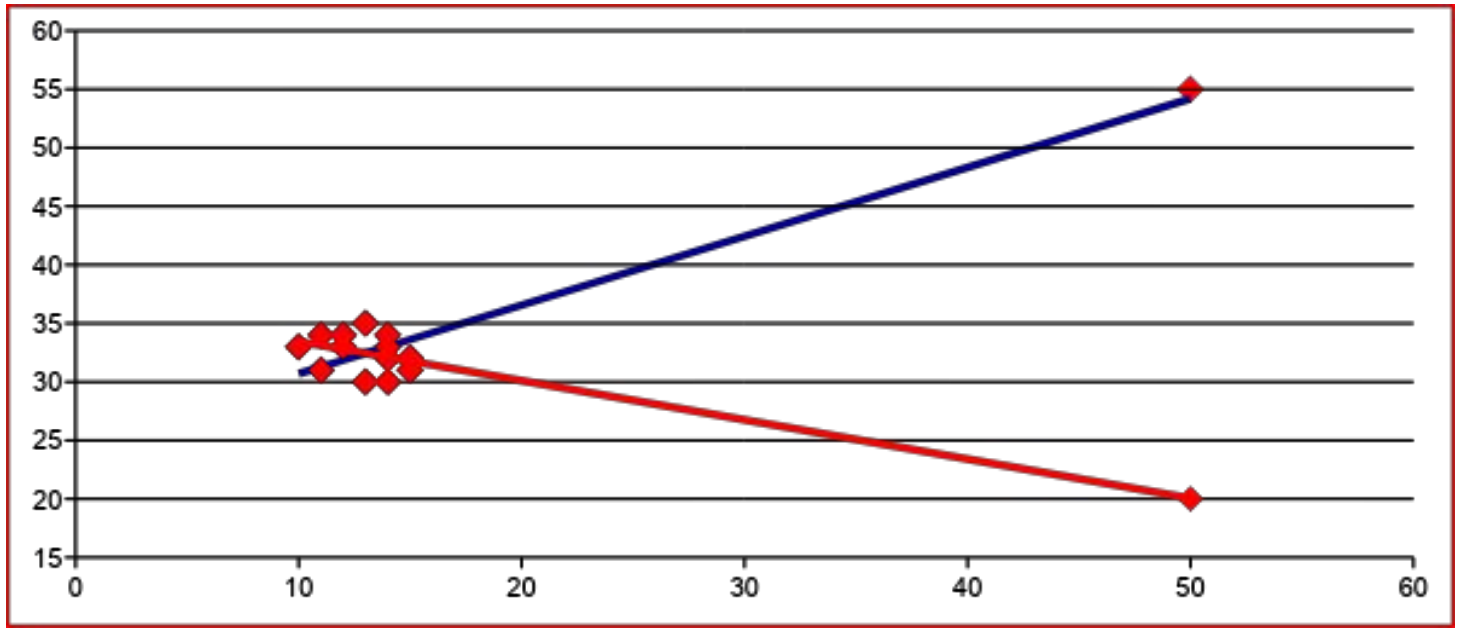
Необходимые предположения



Балансировка модели

X	Y
10	33
11	34
11	31
12	33
12	34
13	30
13	35
14	32
14	34
14	30
14	33
15	32
15	32
15	31
50	**55**

50	**20**

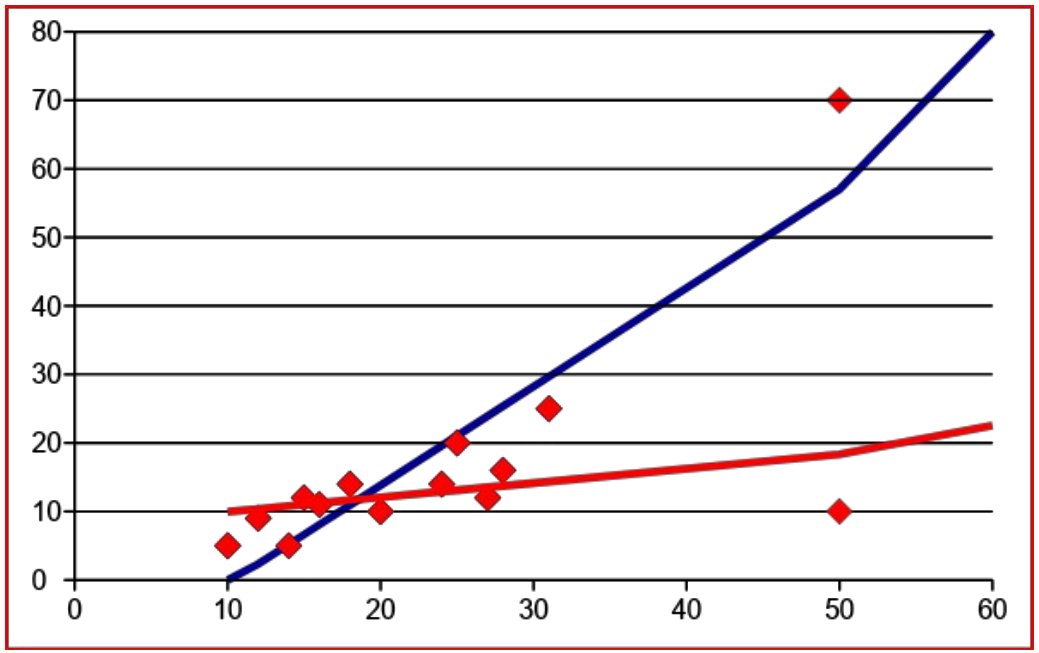


Набор аналогов	Уравнение	R^2
Синий [50;55]	**y = 0,587x + 24,80**	0,891
Красный [50;20]	**y = -0,334x + 36,8**	0,834

Балансировка модели

X	Y
10	5
12	9
14	5
15	12
16	11
18	14
20	10
24	14
25	20
27	12
28	16
31	25
50	**70**

50	**10**



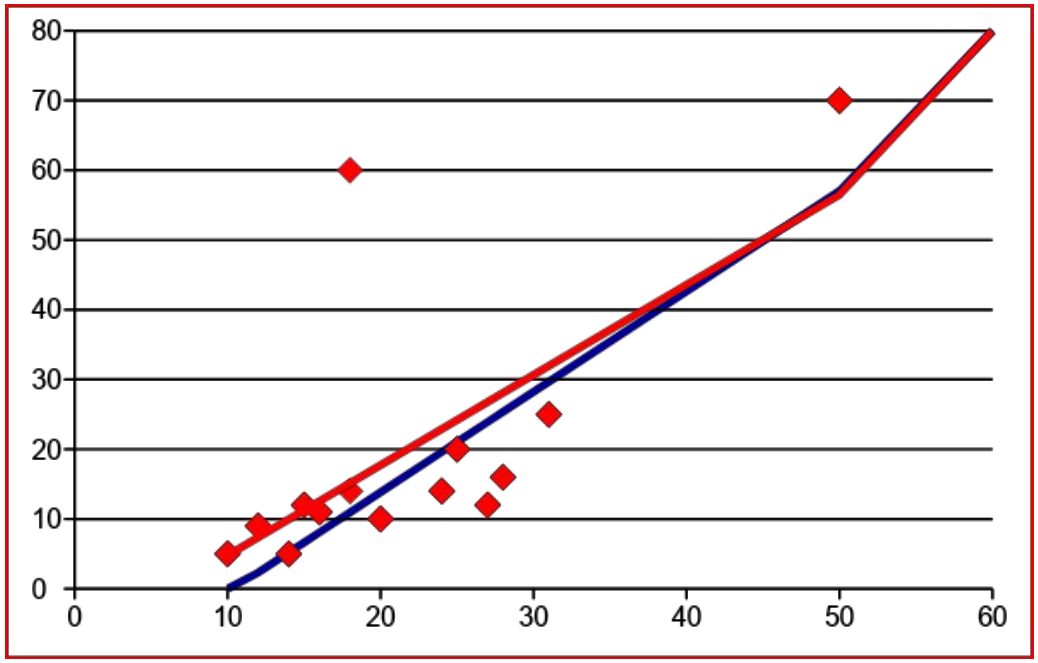
Набор аналогов	Уравнение	R^2
Синий [50;70]	$y = 0{,}203x^{1{,}385}$	0,802
Красный [50;10]	$y = 1{,}693x^{0{,}633}$	0,360

Балансировка модели

X	Y
10	5
12	9
14	5
15	12
16	11
18	**14**
20	10
24	14
25	20
27	12
28	16
31	25
50	70

→

18	**60**



Набор аналогов	Уравнение	R^2
Синий [18;14]	$y = 0{,}203x^{1{,}385}$	0,802
Красный [18;60]	$y = 0{,}287x^{1{,}309}$	0,519

Балансировка модели

Расстояние Кука - это мера влияния соответствующего наблюдения на уравнение регрессии, показывает разницу между вычисленными коэффициентами и значениями, которые получились бы при исключении соответствующего наблюдения. В адекватной модели все расстояния Кука должны быть примерно одинаковыми; если это не так, то имеются основания считать, что соответствующее наблюдение (или наблюдения) смещает оценки коэффициентов регрессии.

$[a_1; a_2; \dots a_n; c]$

y1	x11	x12	…	x1n
y2	x21	x22	…	x2n
y3	x31	x32	…	x3n
…	…	…	…	…
yk	xk1	xk2	…	xkn

$[a_{11}; a_{12}; \dots a_{1n}; c_1]$ $\rho_1 = \sqrt{(a_1 - a_{11})^2 + (a_2 - a_{12})^2 + \dots + (c - c_1)^2}$

$[a_{21}; a_{22}; \dots a_{2n}; c_2]$ $\rho_2 = \sqrt{(a_1 - a_{21})^2 + (a_2 - a_{22})^2 + \dots + (c - c_2)^2}$

$[a_{31}; a_{32}; \dots a_{3n}; c_3]$ $\rho_2 = \sqrt{(a_1 - a_{31})^2 + (a_2 - a_{32})^2 + \dots + (c - c_3)^2}$

$[a_{k1}; a_{k2}; \dots a_{kn}; c_k]$ $\rho_k = \sqrt{(a_1 - a_{k1})^2 + (a_2 - a_{k2})^2 + \dots + (c - c_k)^2}$

Логарифмирование

Исходные данные	Модель
$y; x_1; x_2; x_3$	$y = A_1*x_1 + A_2*x_2 + A_3*x_3 + C$
$\ln(y); x_1; x_2; x_3$	$\ln(y) = A_1*x_1 + A_2*x_2 + A_3*x_3 + C$
$\ln(y); \ln(x_1); \ln(x_2); \ln(x_3)$	$\ln(y) = A_1*\ln(x_1) + A_2*\ln(x_2) + A_3*\ln(x_3) + C$
$\ln(y); \ln(x_1); x_2; x_3$	$\ln(y) = A_1*\ln(x_1) + A_2* x_2 + A_3* x_3 + C$
$y; \ln(x_1); \ln(x_2); \ln(x_3)$	$y = A_1*\ln(x_1) + A_2*\ln(x_2) + A_3*\ln(x_3) + C$

$$y = A_1 \cdot x_1 + A_2 \cdot x_2 + A_3 \cdot x_3 + C$$

$$y = e^{A_1 \cdot x_1} \cdot e^{A_2 \cdot x_2} \cdot e^{A_3 \cdot x_3} \cdot e^{C}$$

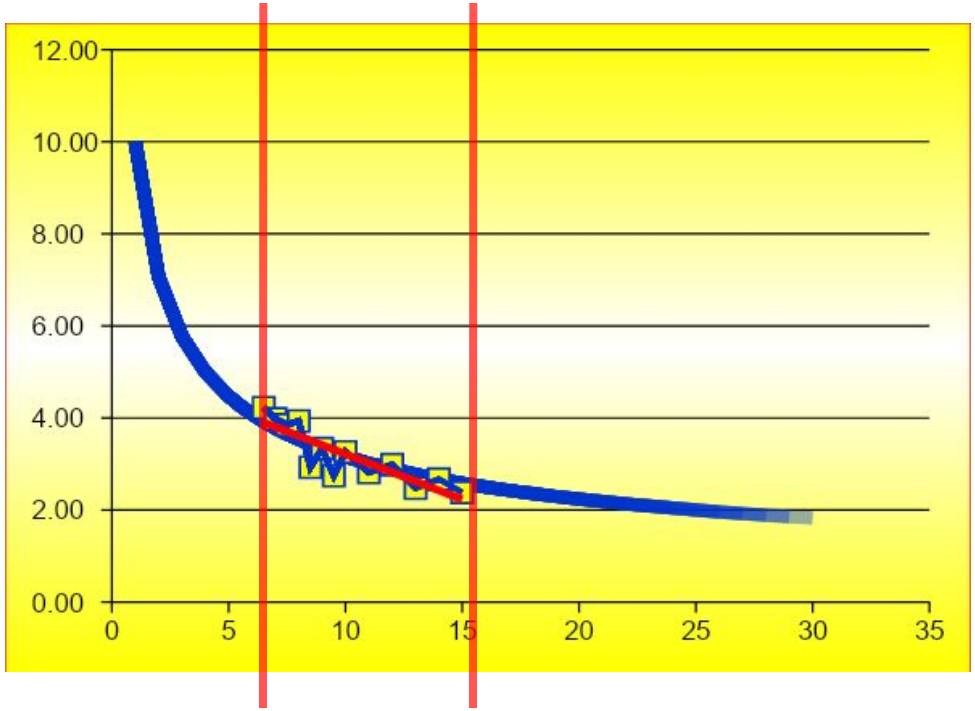
$$y = x_1^{A_1} \cdot x_2^{A_2} \cdot x_3^{A_3} \cdot e^{C}$$

$$y = x_1^{A_1} \cdot e^{A_2 \cdot x_2} \cdot e^{A_3 \cdot x_3} \cdot e^{C}$$

$$y = A_1 \cdot \ln(x_1) + A_2 \cdot \ln(x_2) + A_3 \cdot \ln(x_3) + C$$

Границы применимости

- Модель применима внутри диапазона варьирования признаков объектов-аналогов;
- Возможность применения модели за пределами диапазона варьирования признаков в каждом случае решается индивидуально, на основании анализа рынка (или сопоставления с опытом предыдущего моделирования);
- Экстраполяция по качественным признакам не возможна!!! (нельзя спрогнозировать стоимость в районе Б на основании аналогов из района А).

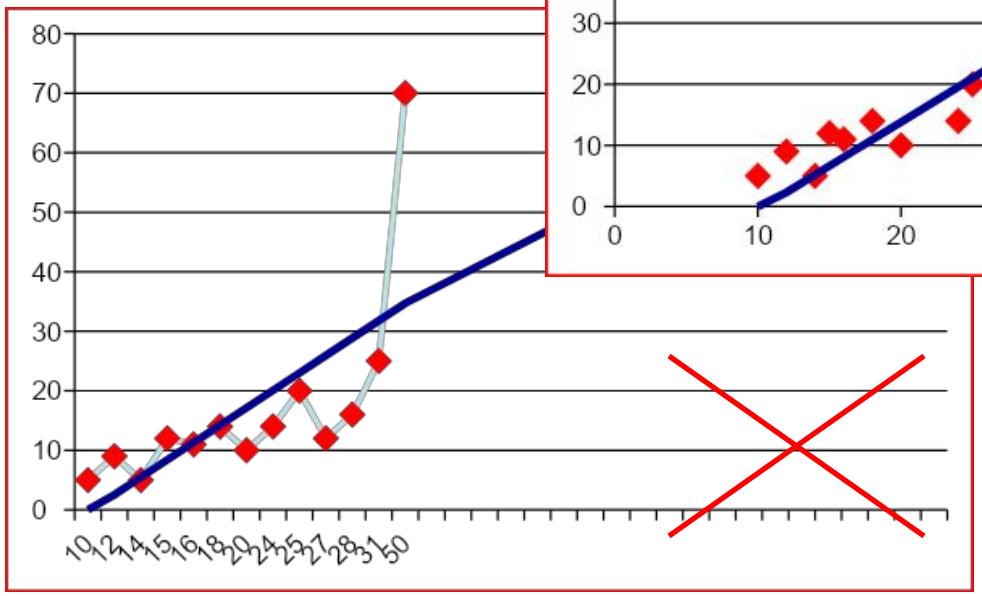


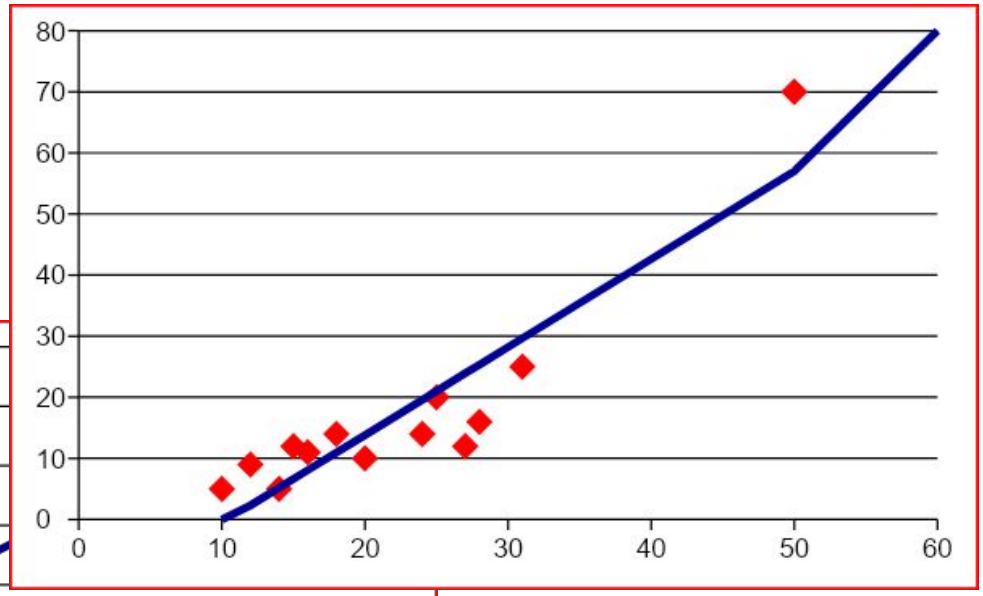
С экстраполяцией надо быть осторожными, т.к. применимость любой регрессионной модели ограничена, особенно, за пределами экспериментальной области.

Графики бывают разные…

X	Y
10	5
12	9
14	5
15	12
16	11
18	14
20	10
24	14
25	20
27	12
28	16
31	25
50	70

«График»





«Точечная»

Несколько полезных источников

- **Ю.Н. Тюрин, А.А. Макаров Анализ данных на компьютере** / Под. ред. В.Э.Фигурнова. - 3-е изд., перераб. и доп. – М.:ИНФРА-М, 2003
- **С.В. Пупенцова Модели и инструменты в экономической оценке инвестиций.** – СПб.: Изд-во «МКС», 2007
- **Электронный учебник StatSoft**: http://www.statsoft.ru/home/textbook/
- **Грибовский С.В., Баринов Н.П., Анисимова И.Н.**

  Учет разнотипных ценообразующих факторов в многомерных регрессионных моделях оценки недвижимости (http://www.appraiser.ru/default.aspx?SectionId=41&Id=1575)
- **Грибовский С.В., Баринов Н.П., Анисимова И.Н.**

  О требованиях к количеству сопоставимых объектов при оценке недвижимости сравнительным подходом (http://www.appraiser.ru/default.aspx?SectionId=41&Id=1577)
- **Грибовский С.В., Баринов Н.П., Анисимова И.Н.**

  О повышении достоверности оценки рыночной стоимости методом сравнительного анализа (http://www.appraiser.ru/default.aspx?SectionId=41&Id=1578)
- **Анисимова И.Н. Отчет по НИР «Применение регрессионных методов в задачах индивидуальной оценки объектов недвижимости при сравнительном подходе»** (http://www.appraiser.ru/default.aspx?SectionId=41&Id=1579)
- **В.Г. Мисовец материалы лекции «Применение регрессионного анализа в оценке»** http://appraiser.ru/default.aspx?SectionId=73&ProductID=334

Спасибо за внимание!

Андрей Марчук
тел. +7 495 648 95 99
E-mail info@rusvs.ru
www.rusvs.ru