Анализ данных в системе Fitter

Алексей Померанцев

Институт химической физики РАН

Содержание

1. Введение
2. Данные
3. Модель
4. Оценка параметров
5. Пример BoxBod
6. Прогнозирование
7. Тестирование модели
8. Мультиколлинеарность
9. Прогноз срока службы ПВХ изоляции
10. Выводы

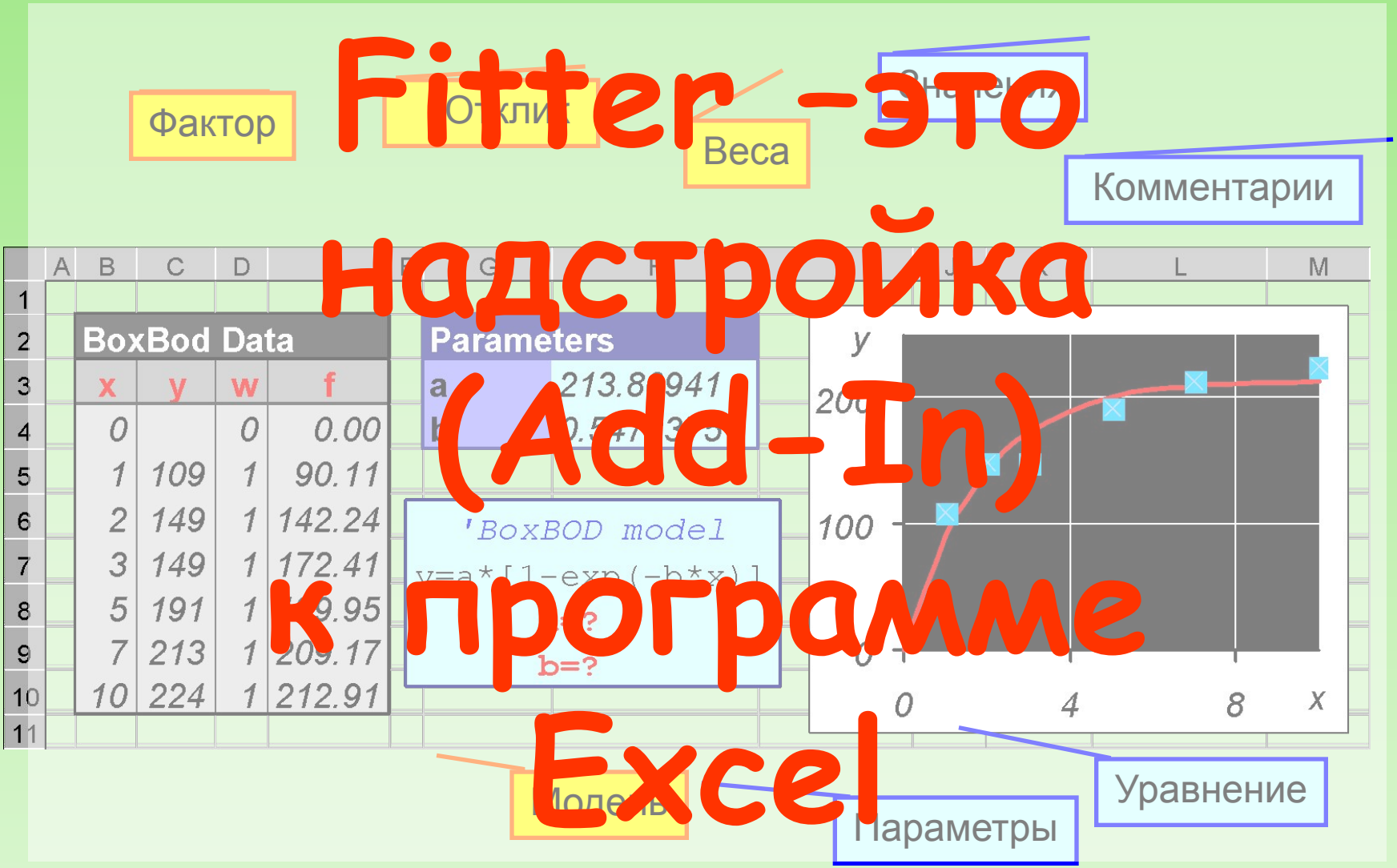
1. Введение

Содержательно (hard) или формально (soft)?

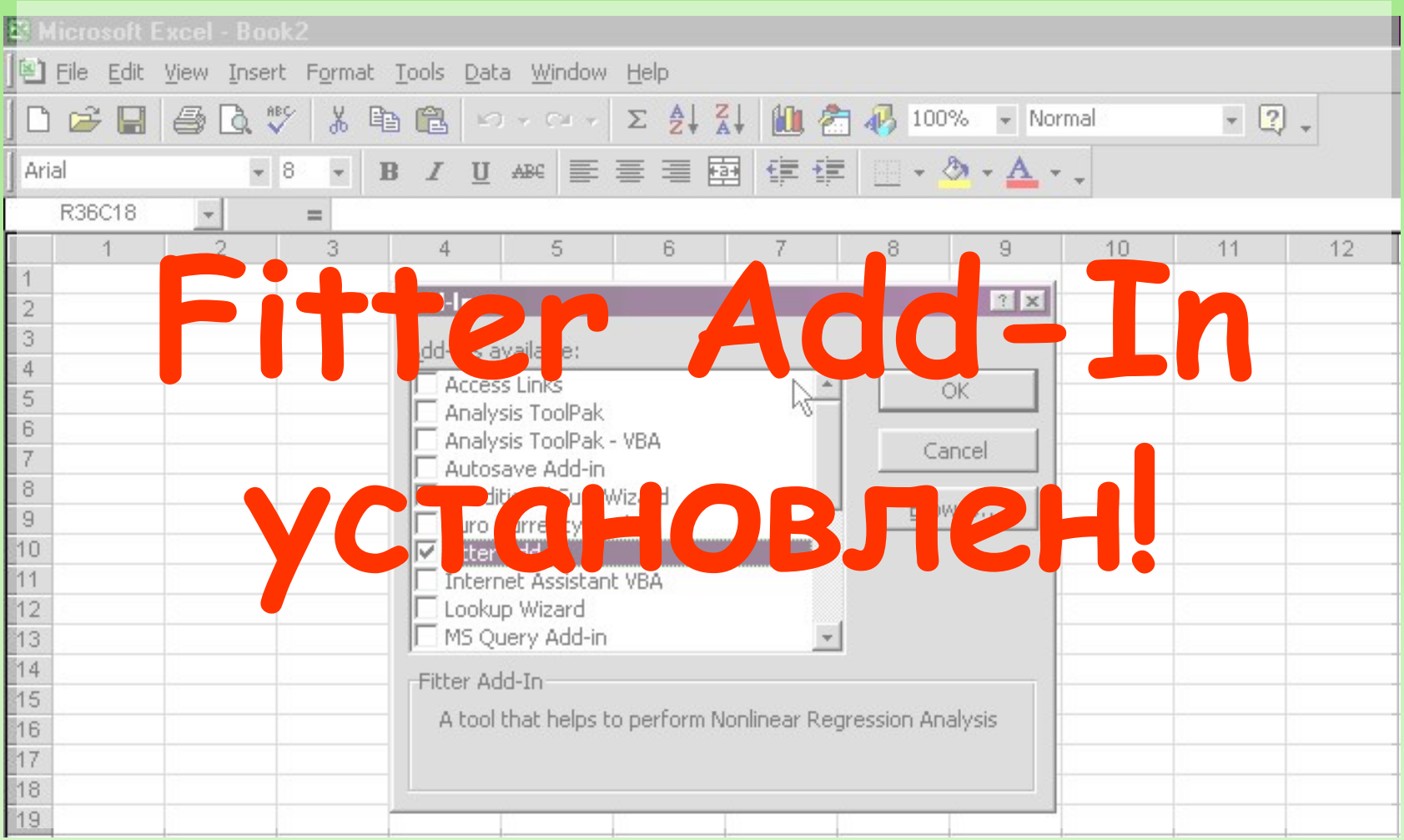
	Содержательно	Формально
Откуда	Физика, химия, ….	Математика, …
Формула	Трудно ?	Легко ?
Построение	Придумать модель	Получить данные
Проблемы	Экстраполяция	Интерполяция
Назначение	Fitter, Matlab, Origin, …	Unscrambler, Simca, SIC,…
Программы		

Разнообразие украшает жизнь!

Система Fitter

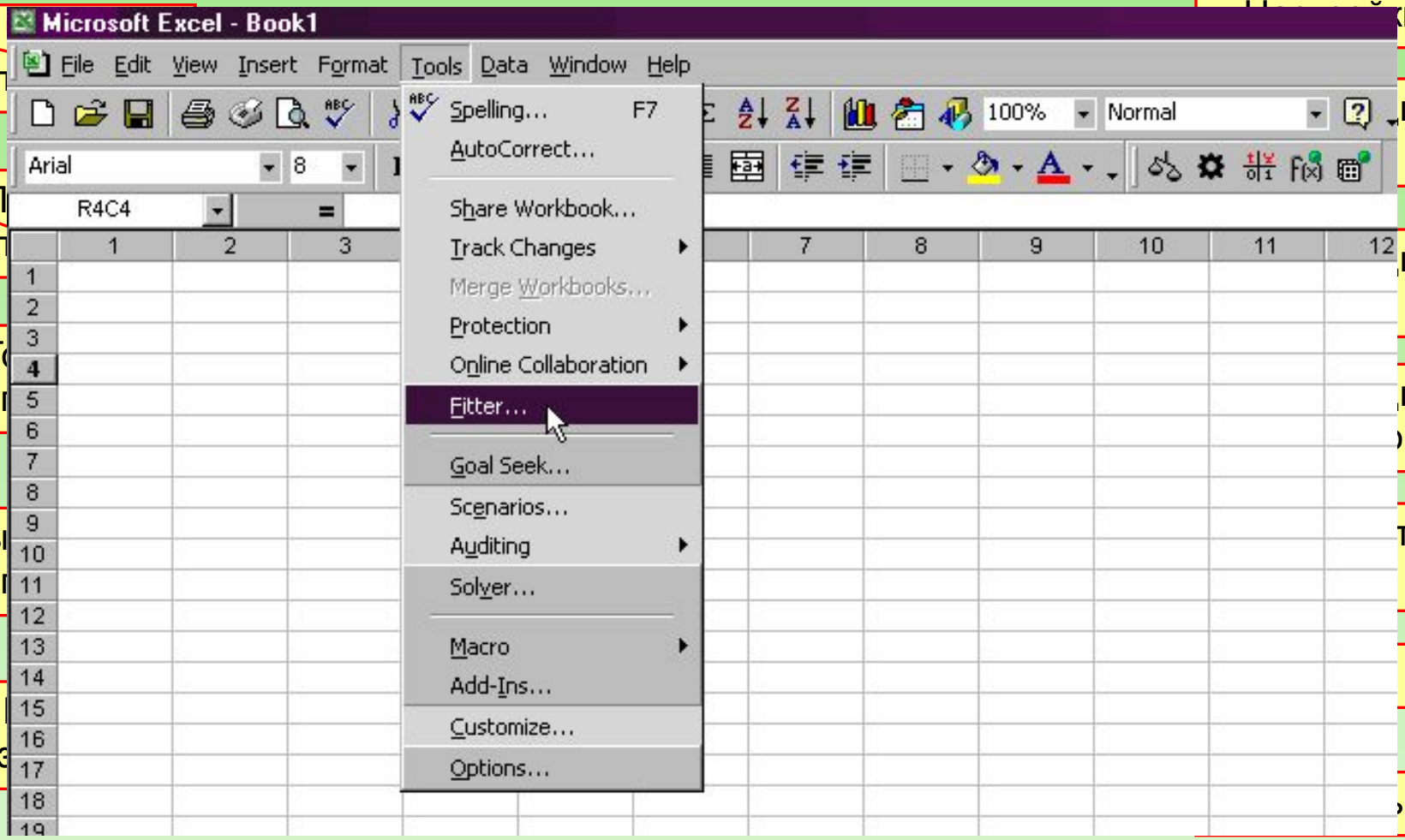


Установка Fitter Add-In



Главный диалог системы Fitter

Поиск



2. Данные

Данные и ошибки

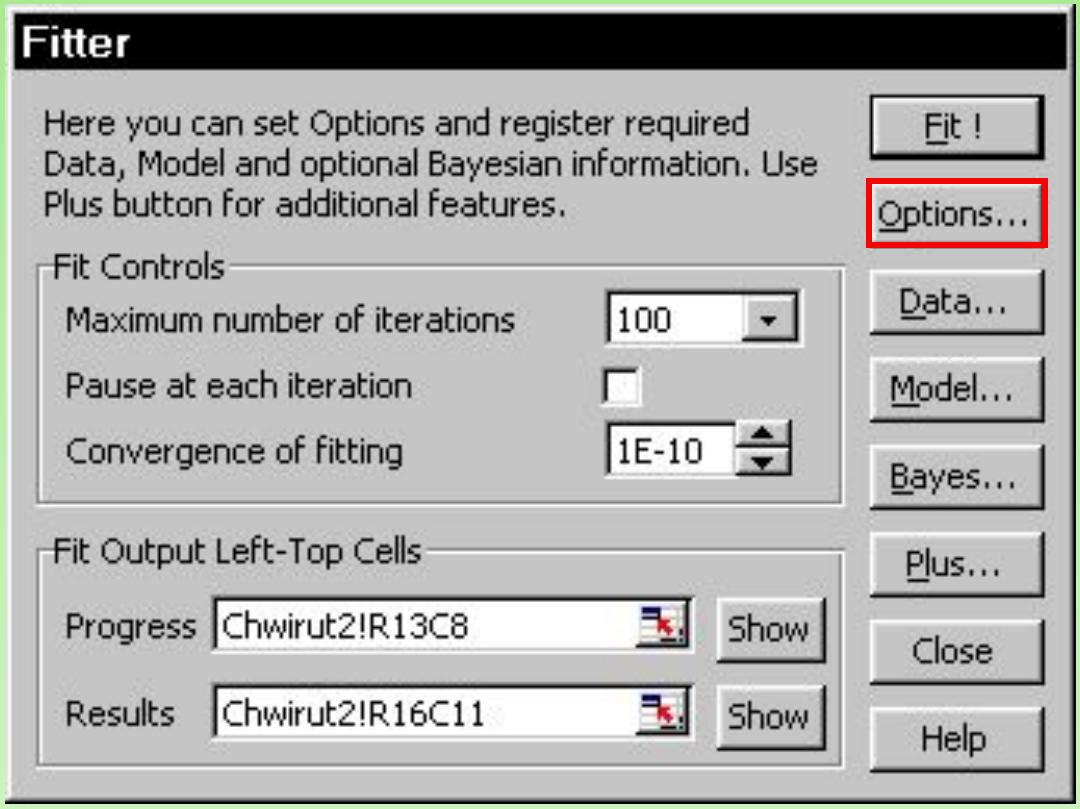
При содержательном подходе связь между данными нелинейная например

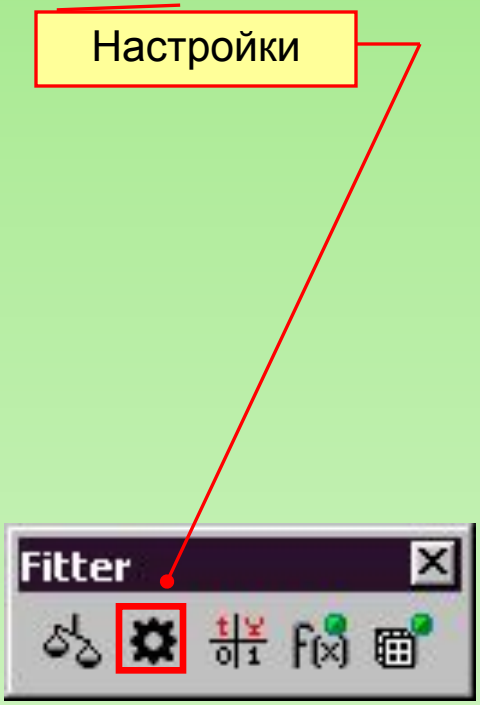
$y=a_1*\exp(a_2*x)+a_3*\exp(a_4*x)$

Отклик	Предикторы	Веса	Параметры	Модель
y_1 … y_N	x_{11} x_{12} … x_{1m} … x_{N1} x_{N2} … x_{Nm}	w_1 … w_N	a_1 a_2 … a_p	f_1 … f_N
y	X	w	a	f

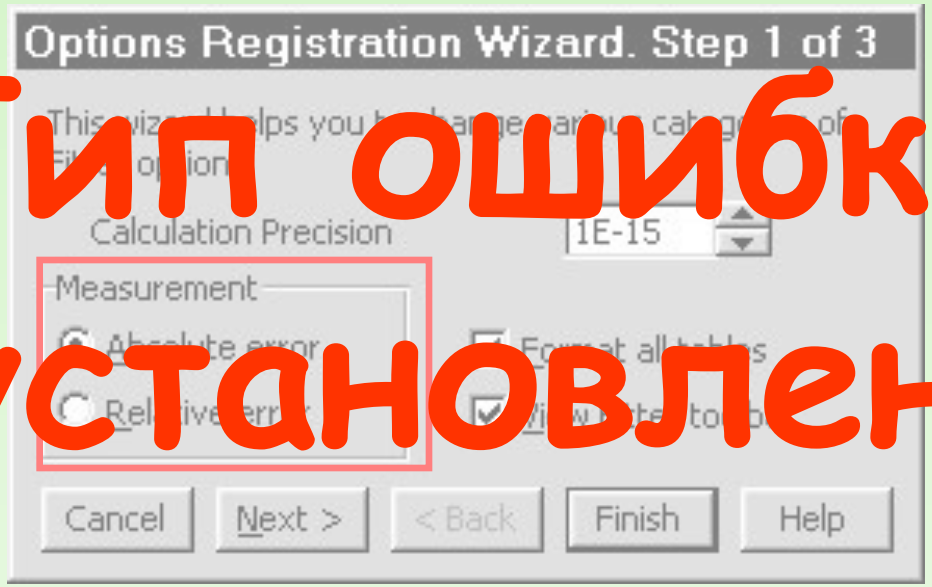
Абсолютная	$Y_i = f_i + \varepsilon_i$	Весы дисперсия
Относительная	$y_i = f_i(1+\varepsilon_i)$	$w_i^2 \operatorname{cov}(\varepsilon_i, \varepsilon_i) = \sigma_\varepsilon^2 = \text{Const}$

Регистрация типа ошибки





Регистрация типа ошибки . Шаг 1



Тип ошибки установлен!

Точность вычислений

Абсолютная ошибка

Формат таблиц

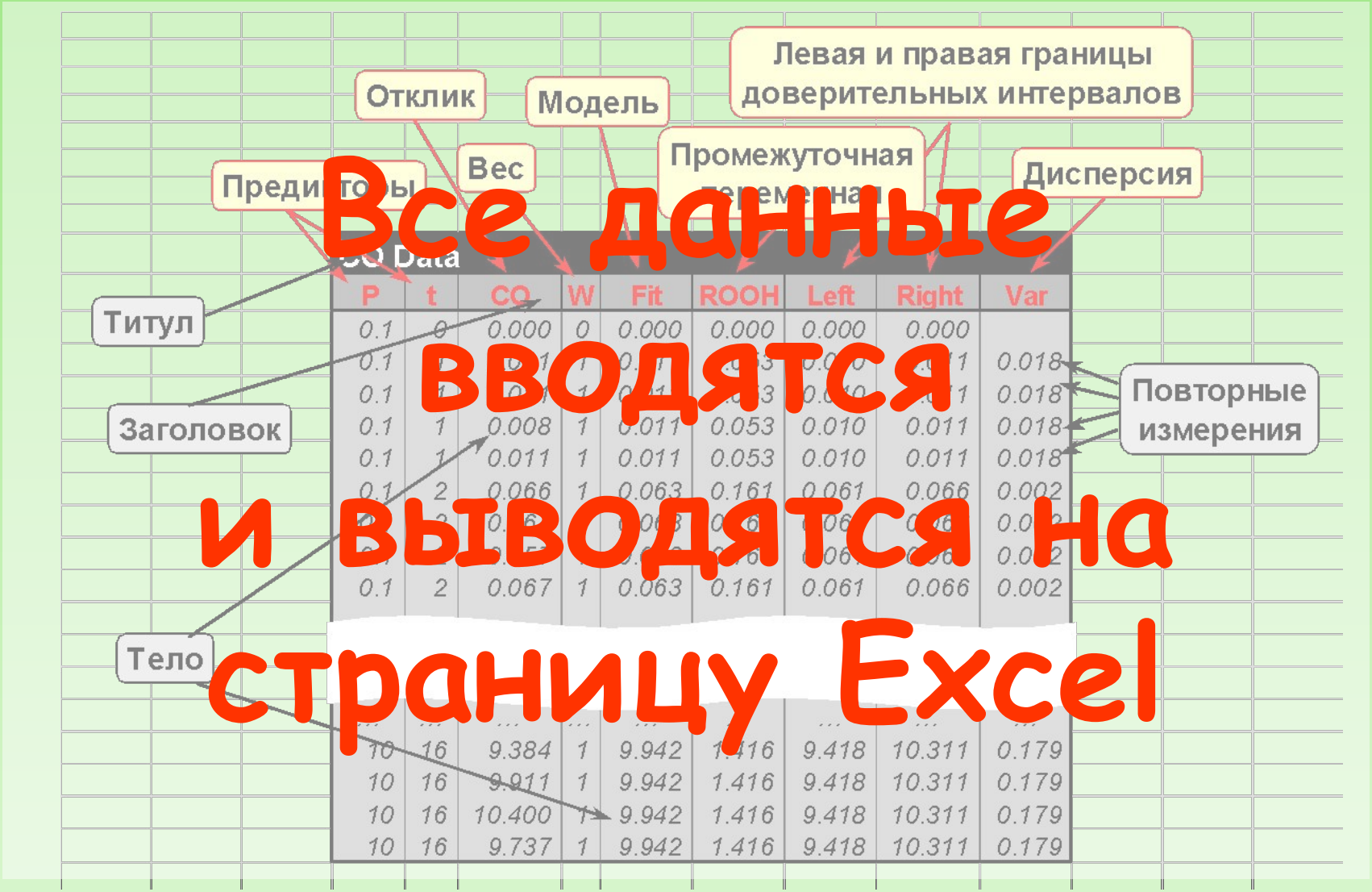
Панель инструментов

Относительная ошибка

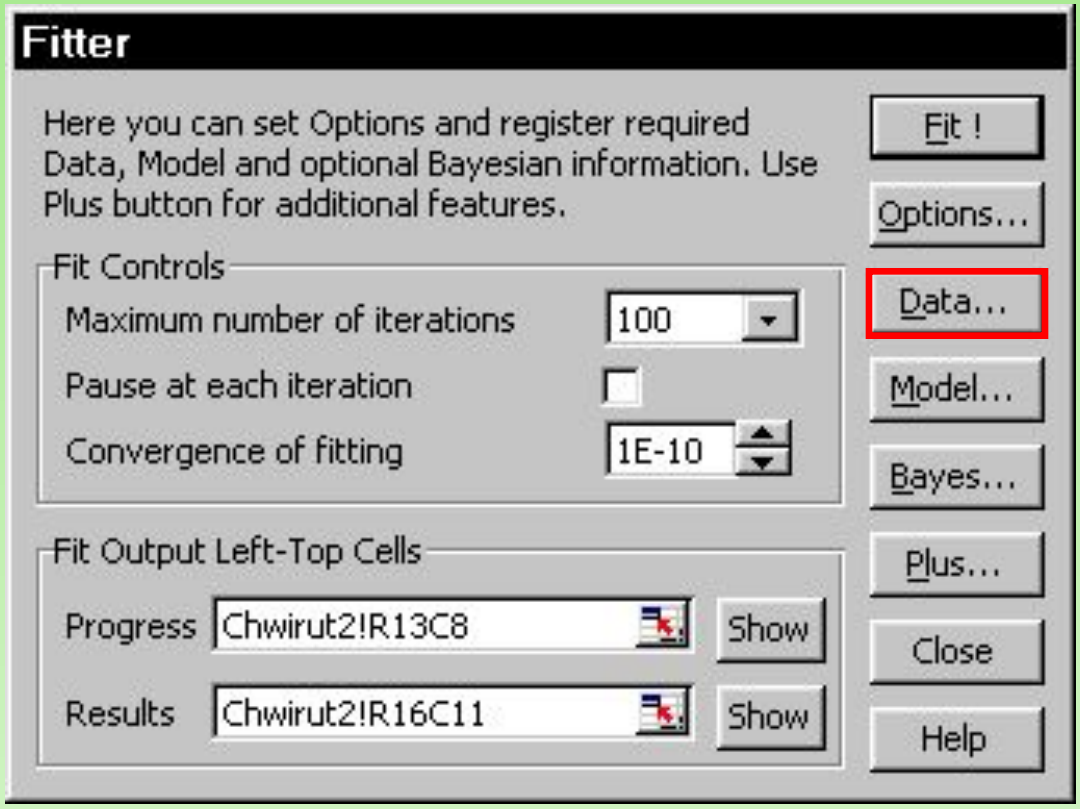
Продолжить

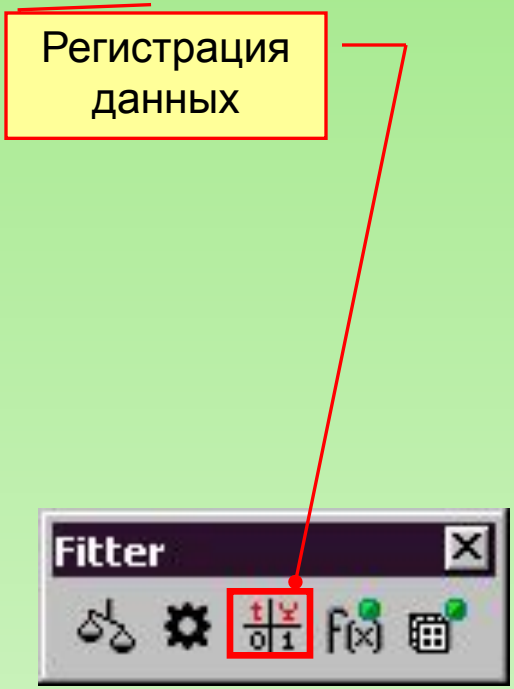
Закончить

Представление данных в системе Fitter

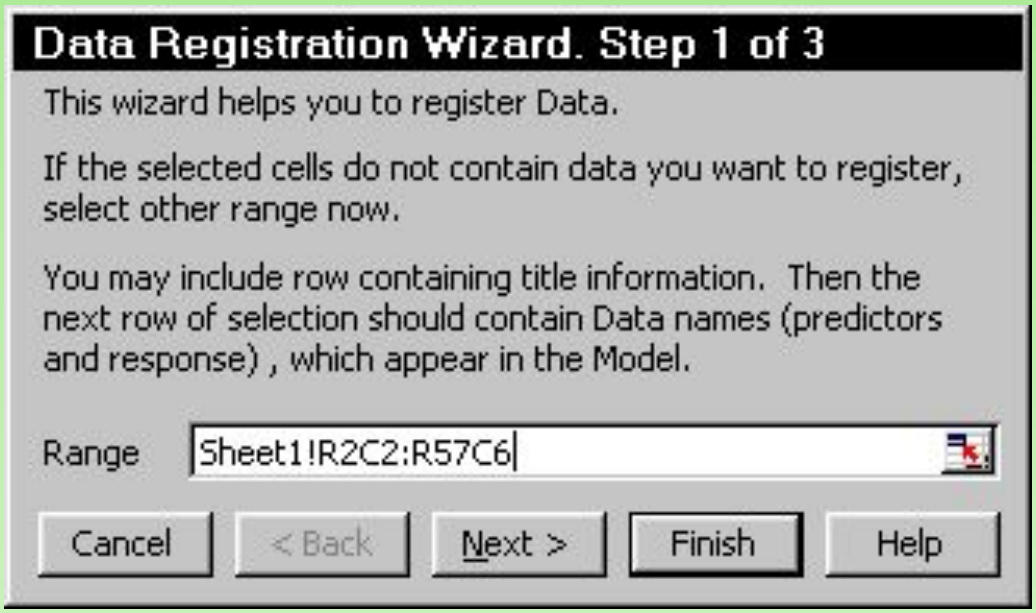


Регистрация данных





Регистрация данных. Шаг 1



Адрес таблицы данных

Закончить

Продолжить

Регистрация данных. Шаг 2

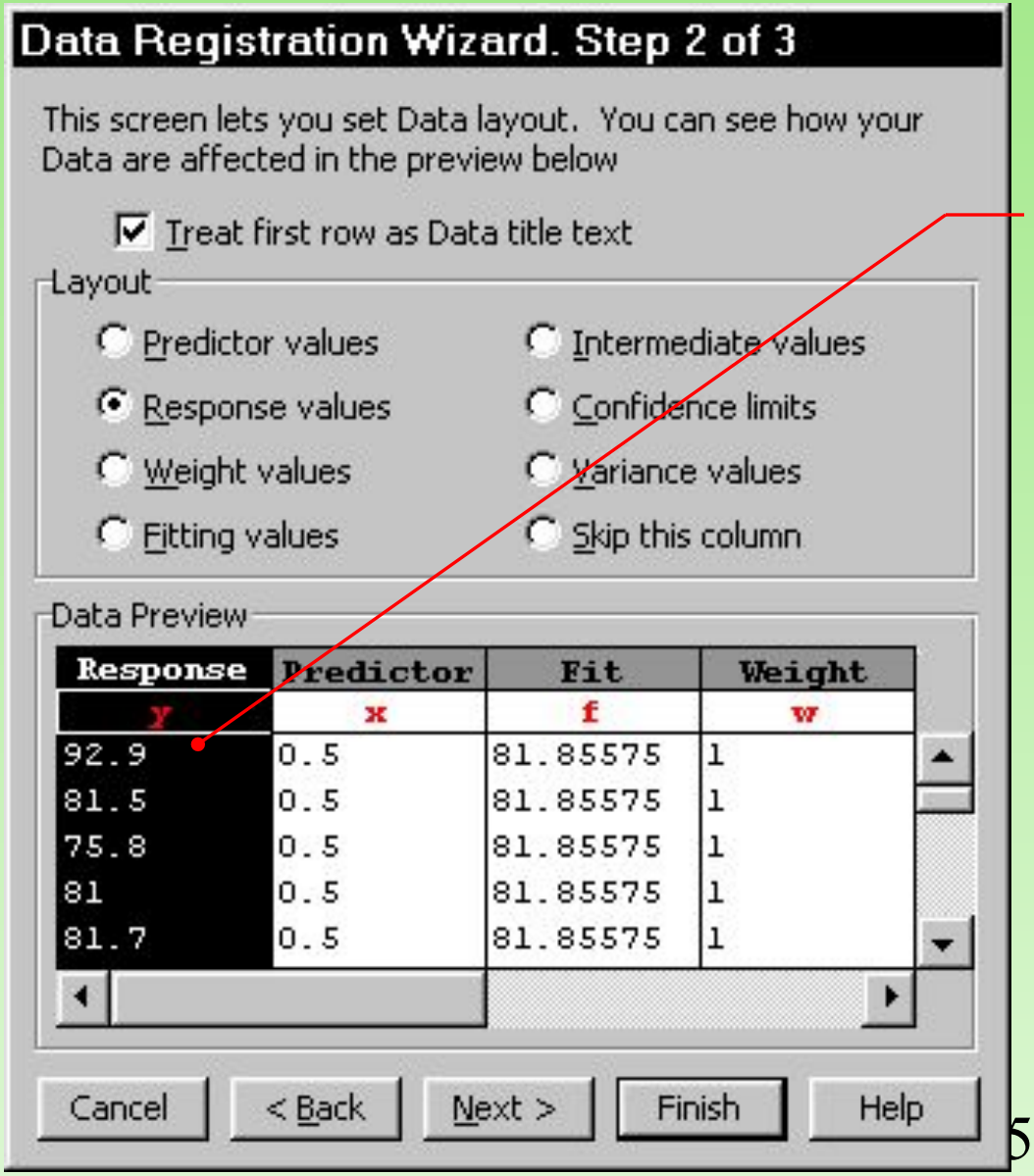
Первый ряд таблицы – это титул

Интерпретация столбца таблицы данных

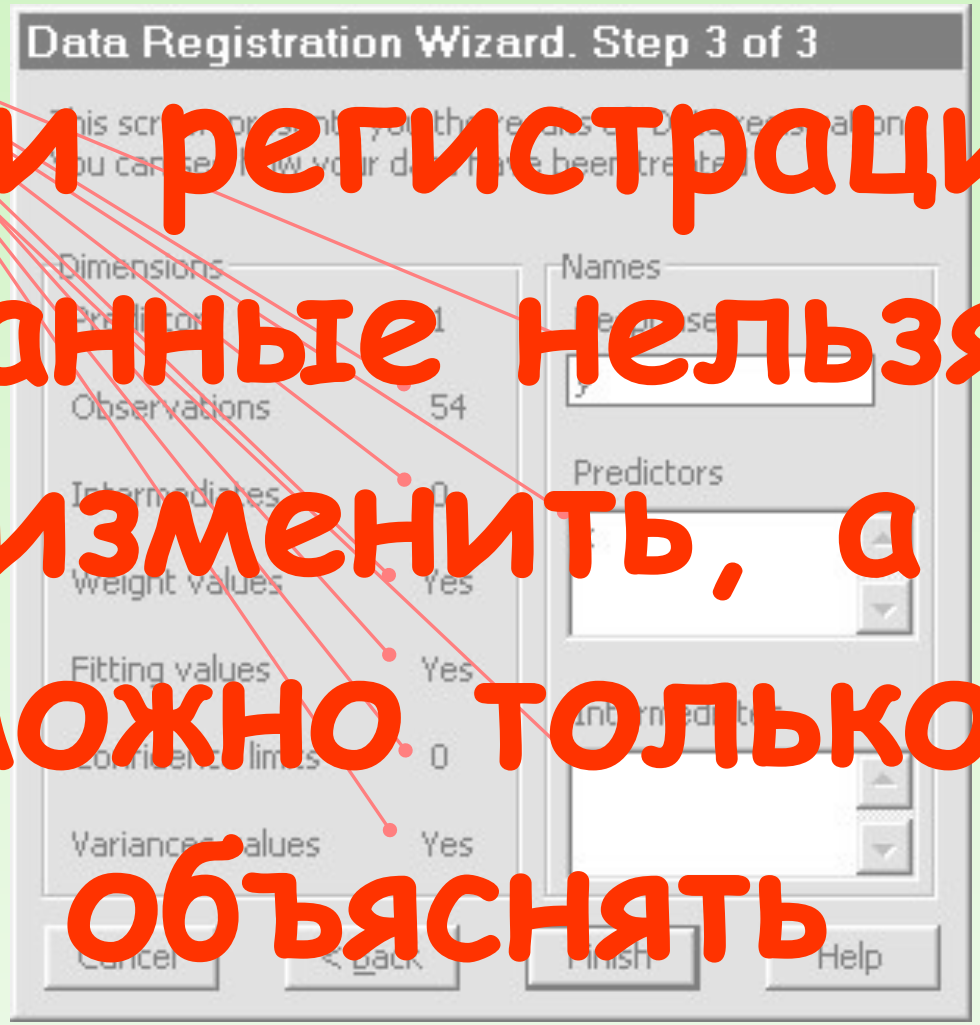
Продолжить

Вернуться на шаг назад

Закончить



Регистрация данных. Шаг 3



Результаты регистрации таблицы данных

При регистрации данные нельзя изменить, а можно только объяснять

Вернуться на шаг назад

Закончить

3. Модель

Данные и модель

$$y = f(x_1, x_2, \ldots, a_1, a_2, \ldots)$$

отклик — предикторы — параметры

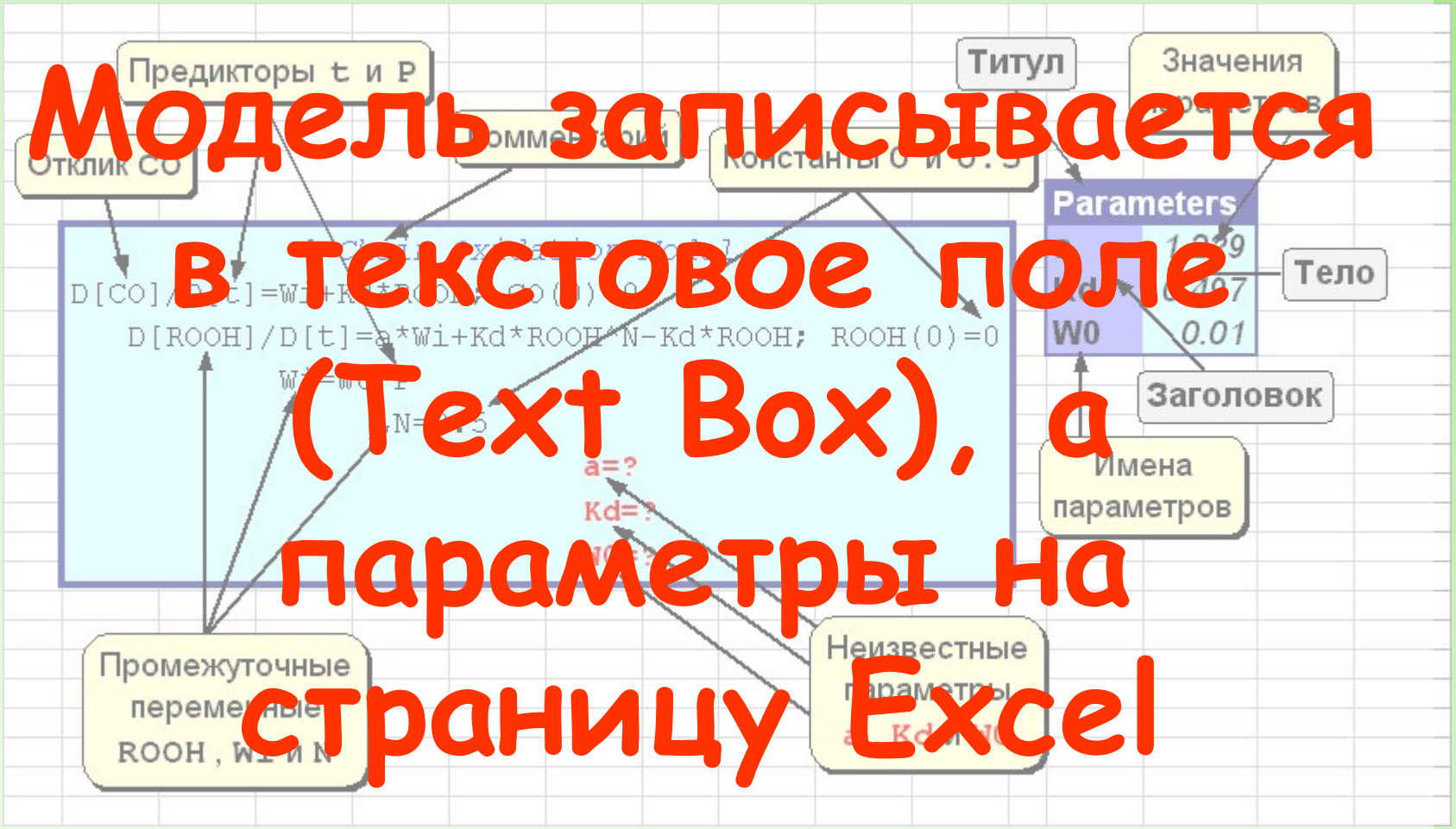
известны — неизвестны

Мера согласия данных и модели – это сумма квадратов отклонений

Мера согласия

$$S(a) = \sum_{i=1}^{N} (y_i - f_i)^2$$

Представление модели в системе Fitter



Модель записывается в текстовое поле (Text Box), а параметры на страницу Excel

Модель $f(x$…

Различные … ели

Явная

Неявная

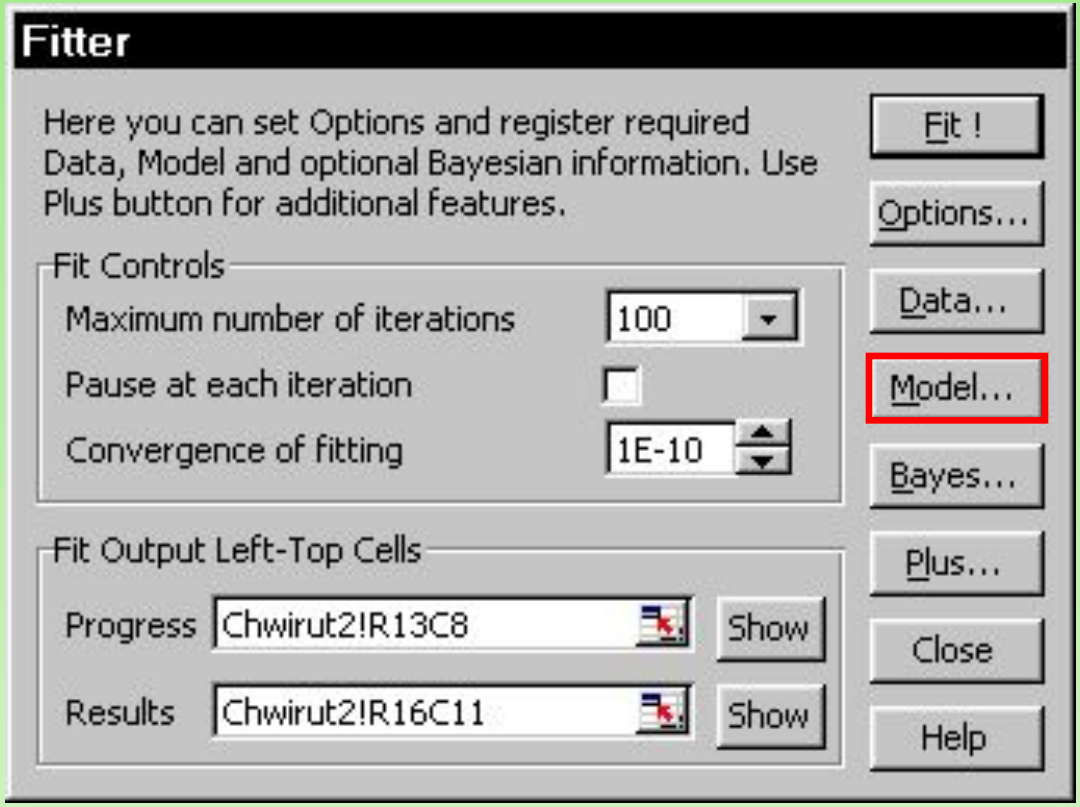
Дифф. уравне… b

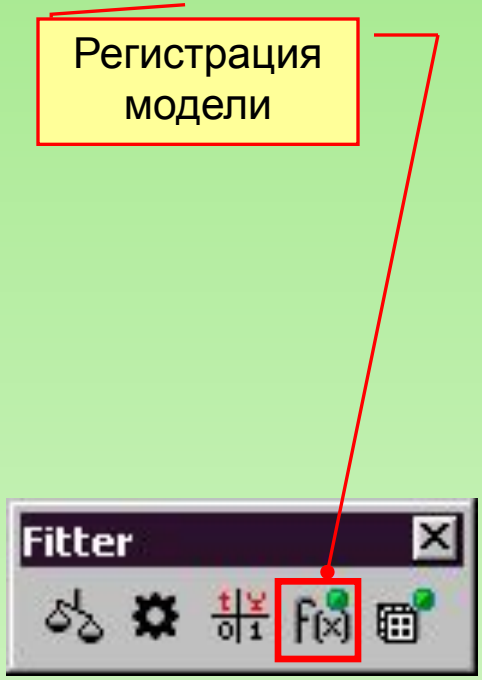
```
'  Explicit…
y=a+(b-a)*e…                                    …(0)=b
            a=…
            b=…
            c=…
```

```
                 'Цикл "увлажнение-сушка"
M=Sor*hev(t1-t)+Des*[hev(t-t1)+imp(t-t1)]
'Кинетика "увлажнения"
    Sor=Sor1*hev(USESor1)+Sor2*[hev(-USESor1)+imp(-USESor1)]
'Кинетика "сушки"
    Des=Des1*hev(USEDes1)+Des2*[hev(-USEDes1)+imp(-USEDes1)]
'Условие применимости асимптотик
    USESor1=Sor2-Sor1
    USEDes1=Des1-Des2
'константы и промежуточные величины
    t3=(t-t1)*hev(t-t1)
    t4=t*hev(t1-t)+t1*[hev(t-t1)+imp(t-t1)]
    P2=PI*PI
    P12=(PI)^(-0.5)
    R=r*(M1-M0)*exp(-r*t4)
    K=M1+(M0-M1)*exp(-r*t4)
    V0=M0-C0
    V1=M1-C0
'асимптотик[illegible]бции при 0<t<tau
    Sor1=C[illegible]4*P1[illegible](d*t)^0.5*[M0-C0+(M1-M0)*beta]
       bet[illegible]=1-ex[illegible])
       x=[illegible]t
       z=[illegible]*x*x+b3*x*x*x)
          a1[illegible]666539[illegible]
          a2=0.012105101749
          a3=0.009922532242 8
          b1=0.084800623251 9
          b2=0.024663459122 3
          b3=0.001754994795 8
'ки[illegible]
    [illegible]or2=K-[illegible]S1
       [illegible]
          n0=P2
          U01=[(V0*n0*d-V1*r)*exp(-n0*d*t4)+R]/(n0*d-r)
          n1=P2*9
          U11=[(V0*n1*d-V1*r)*exp(-n1*d*t4)+R]/(n1*d-r)
          n2=P2*25
          U21=[illegible]+R][illegible](n2*d-r)
          [illegible]
          [illegible][(V[illegible]exp[illegible]R][illegible](n3*d-r)
          [illegible]
          U41=[(V0*[illegible]exp(-n4*d*t4)+R]/(n4*d-r)
'асимптотика десорбции при t1<t<t1+tau
    Des1=K*[1-4*P12*(d*t3)^0.5]-8*S1
'кинетика десорбции при t1+tau<t
    Des2=8*S2
       S2=U02/n0+U12/n1+U22/n2+U32/n3+U42/n4
          U02=(K-U01)*exp(-n0*d*t3)
          U12=(K-U11)*exp(-n1*d*t3)
          U22=(K-U21)*exp(-n2*d*t3)
          U32=(K-U31)*exp(-n3*d*t3)
          U42=(K-U41)*exp(-n4*d*t3)
'неизвестные параметры
             d=?
             M0=?
             M1=?
             C0=?
             r=?
             t1=?
```

Очень сложная модель!

Регистрация модели





Регистрация модели. Шаг 1

Выбор текстового поля с моделью

Адрес первой клетки таблицы параметров

Закончить

Продолжить

Регистрация модели. Шаг 2

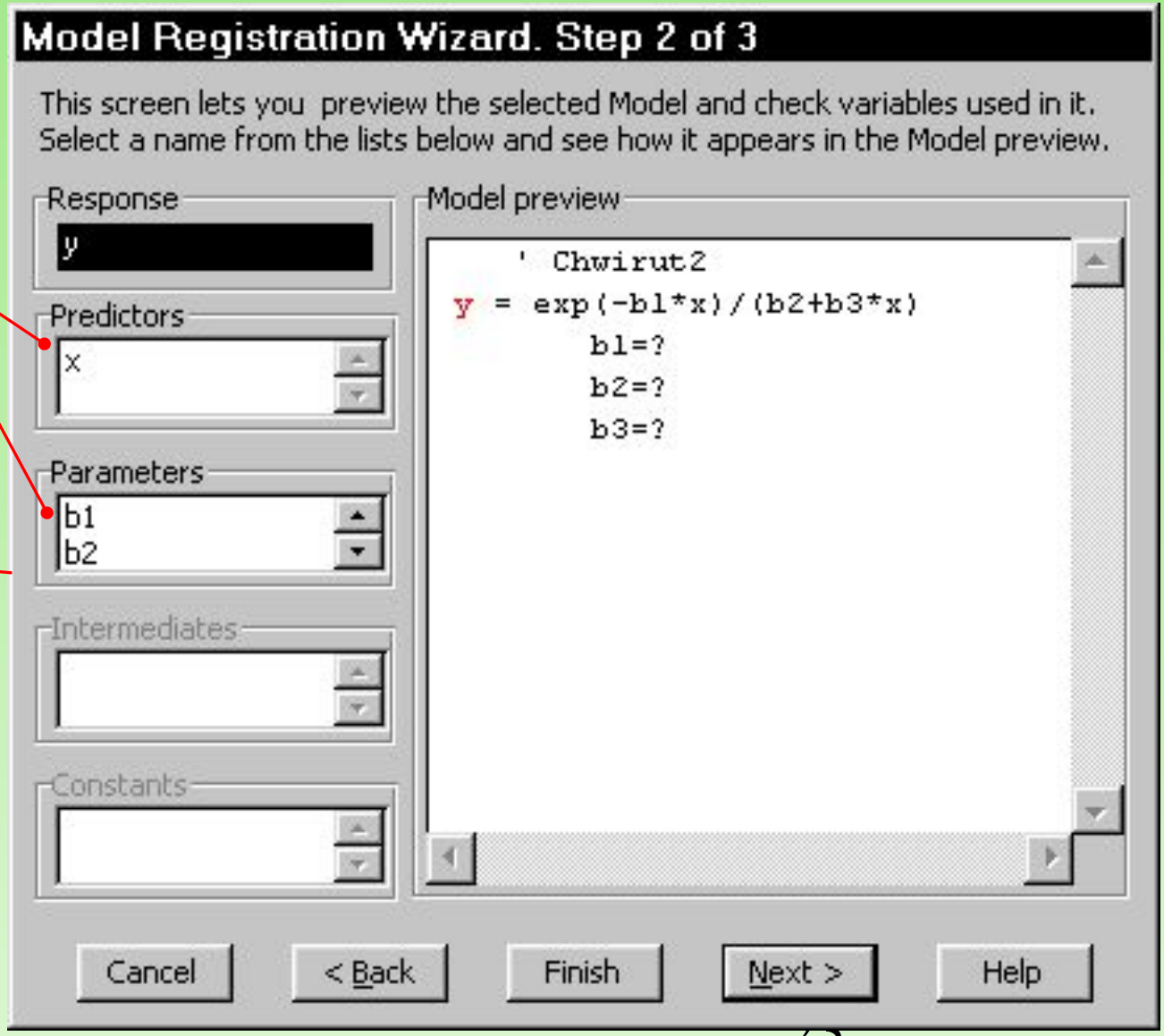
Вид модели

Переменные модели

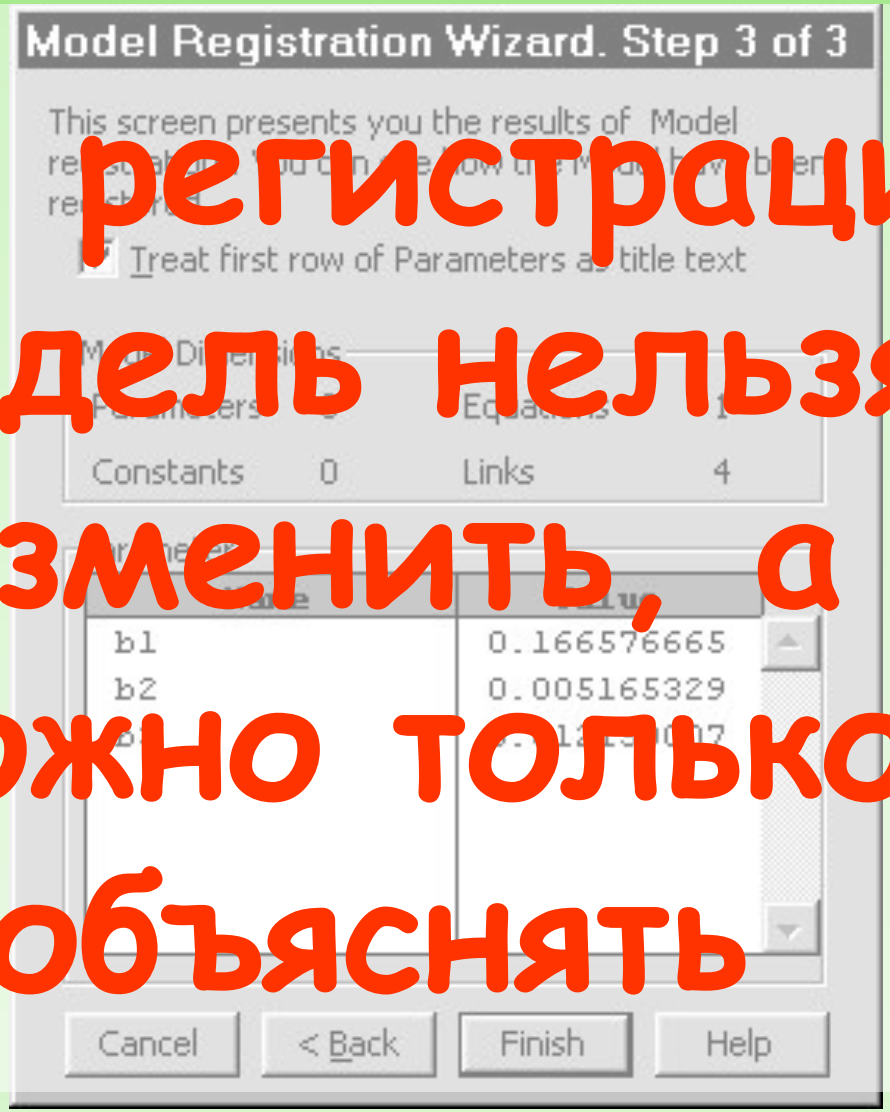
Продолжить

Закончить

Вернуться на шаг назад



Регистрация модели. Шаг 3



Первый ряд таблицы параметров – это титул

Начальные значения параметров

Вернуться на шаг назад

Закончить

При регистрации модель нельзя изменить, а можно только объяснять

4. Оценка параметров

Целевая функция $Q(a)$

Целевая функция Q – это сумма квадратов и, может быть, что-то еще…

Сумма квадратов	$S(a) = \sum_{i=1}^{N} w_i^2 (y_i - f_i)^2 g_i^2$
Байесовский член	$B(a) = s^2 \left[N_a + (a-b)^t H (a-b) \right]$
Целевая функция	$Q(a) = S(a) + B(a)$
Оценки параметров	$\hat{a} = \arg\min_a Q(a)$

Градиентный метод поиска оценок

$$Q(a) \approx Q(a_n) + b_n^t (a - a_n) + \frac{1}{2}(a - a_n)^t A (a - a_n)$$

$$a_{n+1} = a_n + A^+ b_n$$

$$A^+ A = I$$

Проблемы поиска

Исходная точка a_0

$$\det(A) \approx 0$$

Локальный минимум

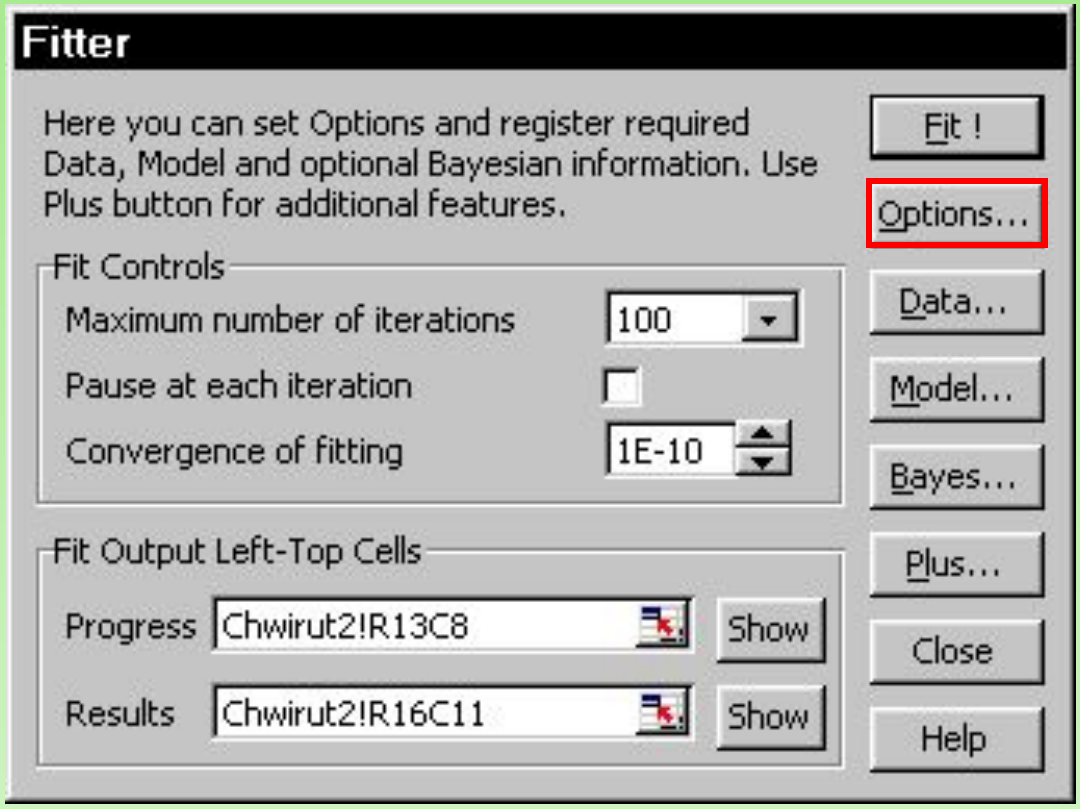
Матрица A – это ключ к поиску!

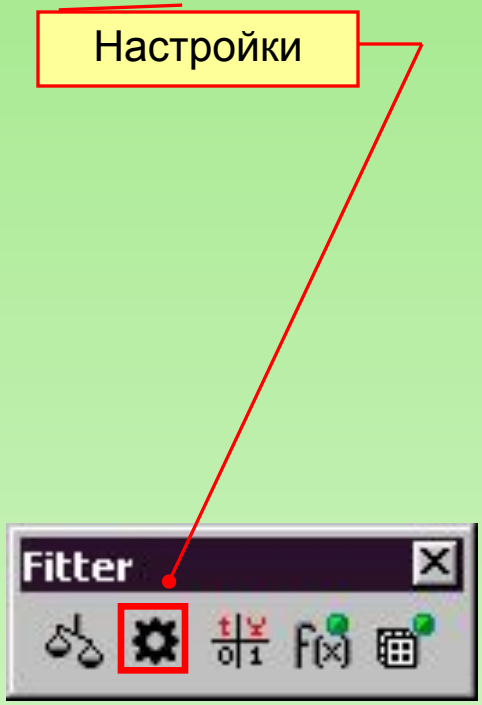
Точность оценивания (статистики)

Матрица ковариаций	$\operatorname{cov}(\hat{a},\hat{a}) = C = s^2 A^{-1}$
Средне-квадратичные отклонения оценок	$\operatorname{dev}(\hat{a}_\alpha) = \sqrt{C_{\alpha\alpha}}$
Матрица корреляций	$\operatorname{cor}(\hat{a}_\alpha,\hat{a}_\beta) = \dfrac{C_{\alpha\beta}}{\sqrt{C_{\alpha\alpha}C_{\beta\beta}}}$
F-матрица (Фишера)	$F = s^{-2}A = C^{-1}$
Оценка дисперсии и число степеней свободы	$s^2 = \dfrac{S(\hat{a})}{N_f} \qquad N_f = N - p$

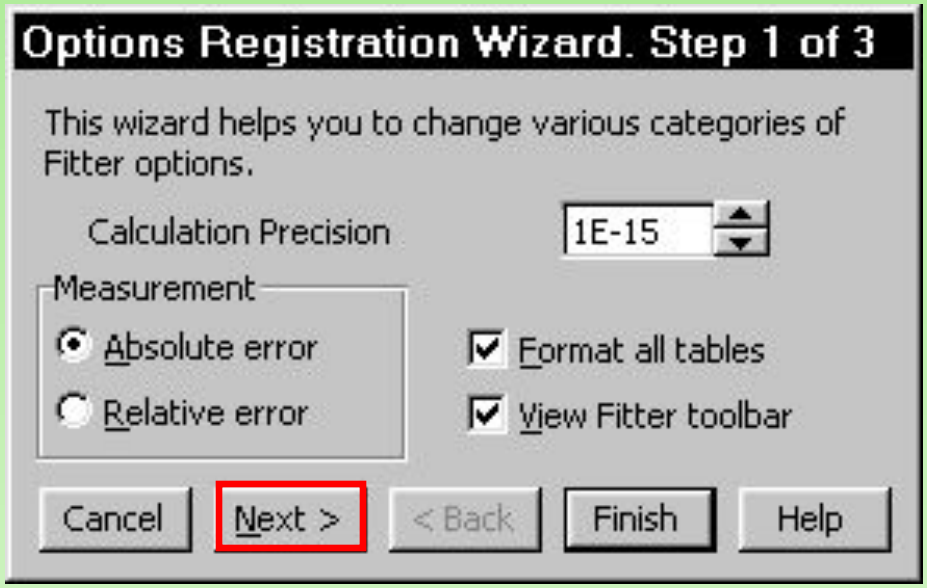
Матрица A – это мера качества оценивания!

Регистрация настроек статистик





Регистрация настроек . Шаг 1



Продолжить

Регистрация настроек статистик. Шаг 2

Общие данные

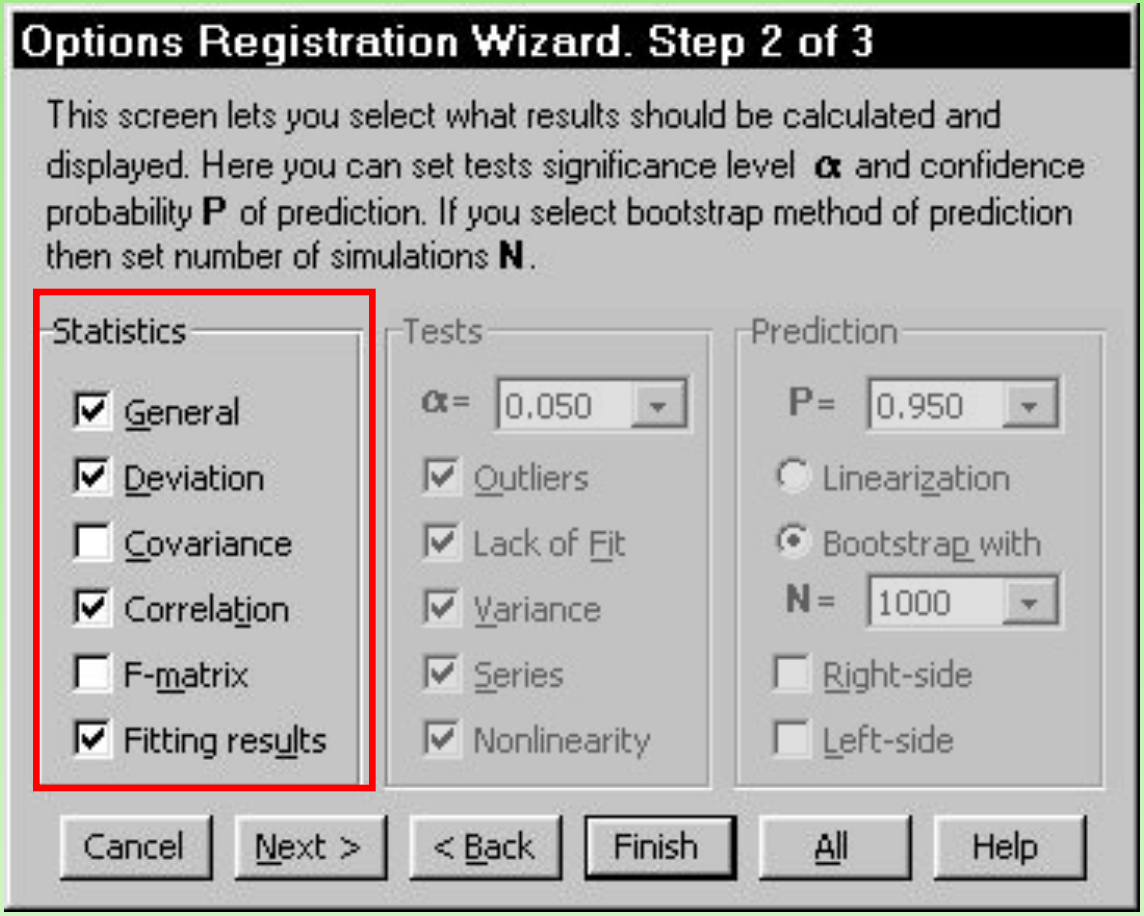
СКО

Ковариации

Корреляции

F-матрица

Итоги поиска



Закончить

5. Пример BoxBod (NIST)

Данные

x	y	f
1	109	90.11
2	149	142.24
3	149	172.44
5	191	199.95
7	213	209.17
10	224	212.1

Модель

$$y = a_1\left(1 - e^{-a_2 x}\right)$$

Параметры

	Оценки	СКО
a_1	213.809	12.3545
a_2	0.547237	0.104560

Посмотрим, как работает Fitter

Априорная (байесовская) информация

Bayesian Information

Name	Value	Matrix	Exclude

Переключаемся в Excel

BoxBod Example (NIST Data Base)

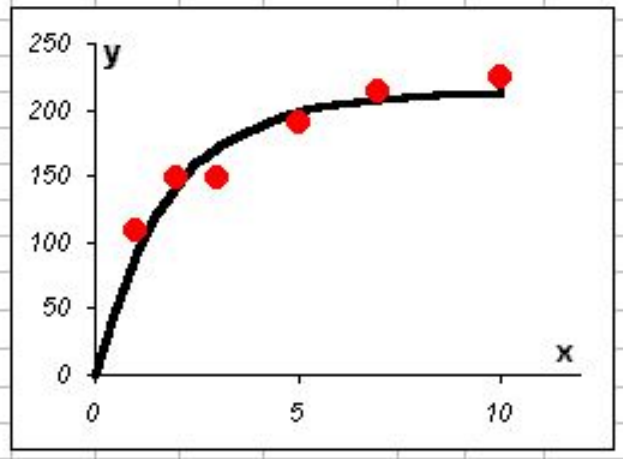
BoxBod Data

x	y	w	f
0		0	0
1	109	1	90.11
2	149	1	142.24
3	149	1	172.41
5	191	1	199.95
7	213	1	209.17
10	224	1	212.91

Parameters

a1	213.809439
a2	0.547237136

```
' BoxBOD
y=a1*(1-exp(-a2*x))
a1=?
a2=?
```



Search Progress

Objective value	1168.01
Completeness	1
Objective change	4.6E-11
Iteration	31

General

Date	21.01.03 18:08
Data	R2C2:R10C5
Model	BoxBod
Parameters	R2C7
Bayes	
Precision	1E-10
Convergence	0.00001
Error type	Absolute
Significance	0.05
Confidence	0.75
Prediction	Linearization

Parameters estimation

Name	Initial	Final	Deviation
a1	1	213.809	12.35452
a2	1	0.54724	0.10456

6. Прогноз

Достоверный прогноз

Оценка отклика

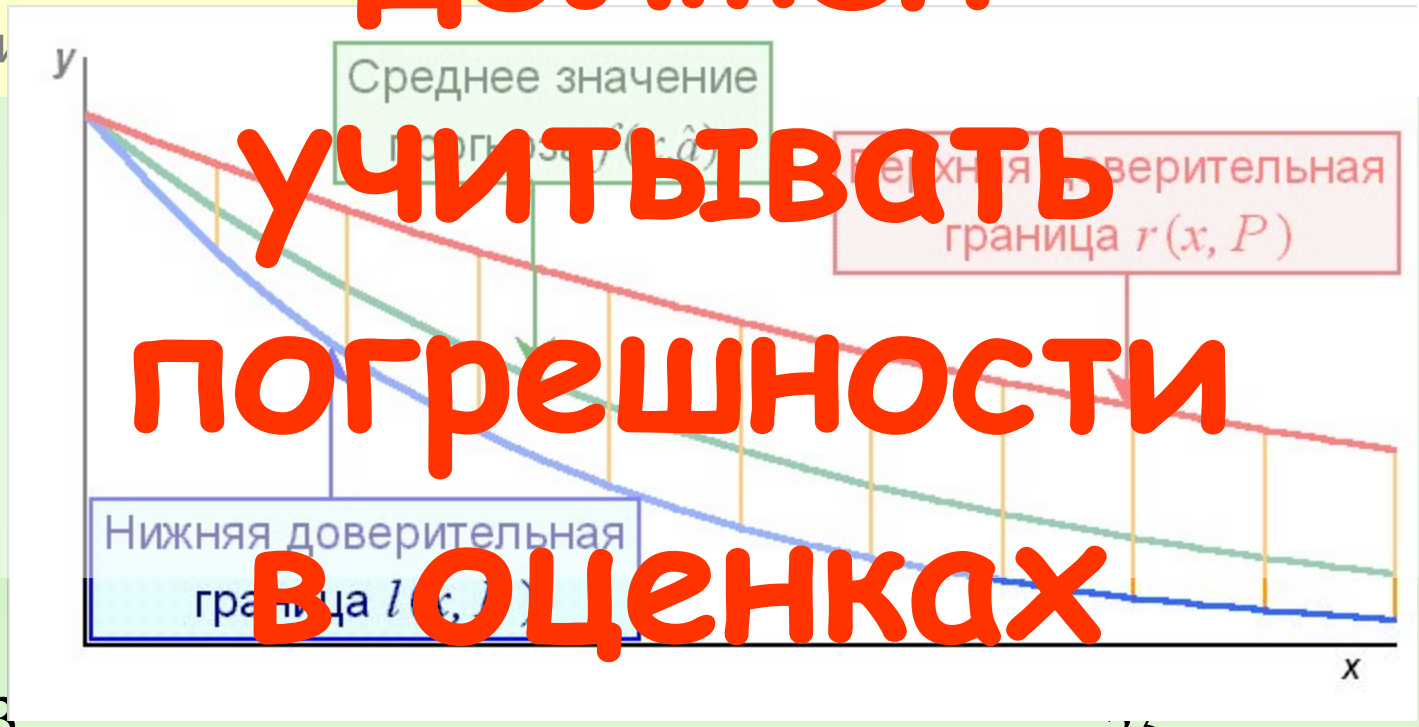
$$\hat{y} = f(\hat{x}, \hat{a})$$

Доверительные границы

$$\mathrm{Prob}\{l(x,P) < f(x,a) < r(x,P)\} \geq P$$

$$f(x,\hat{a}) \pm g(P)\sqrt{v^t C v}$$

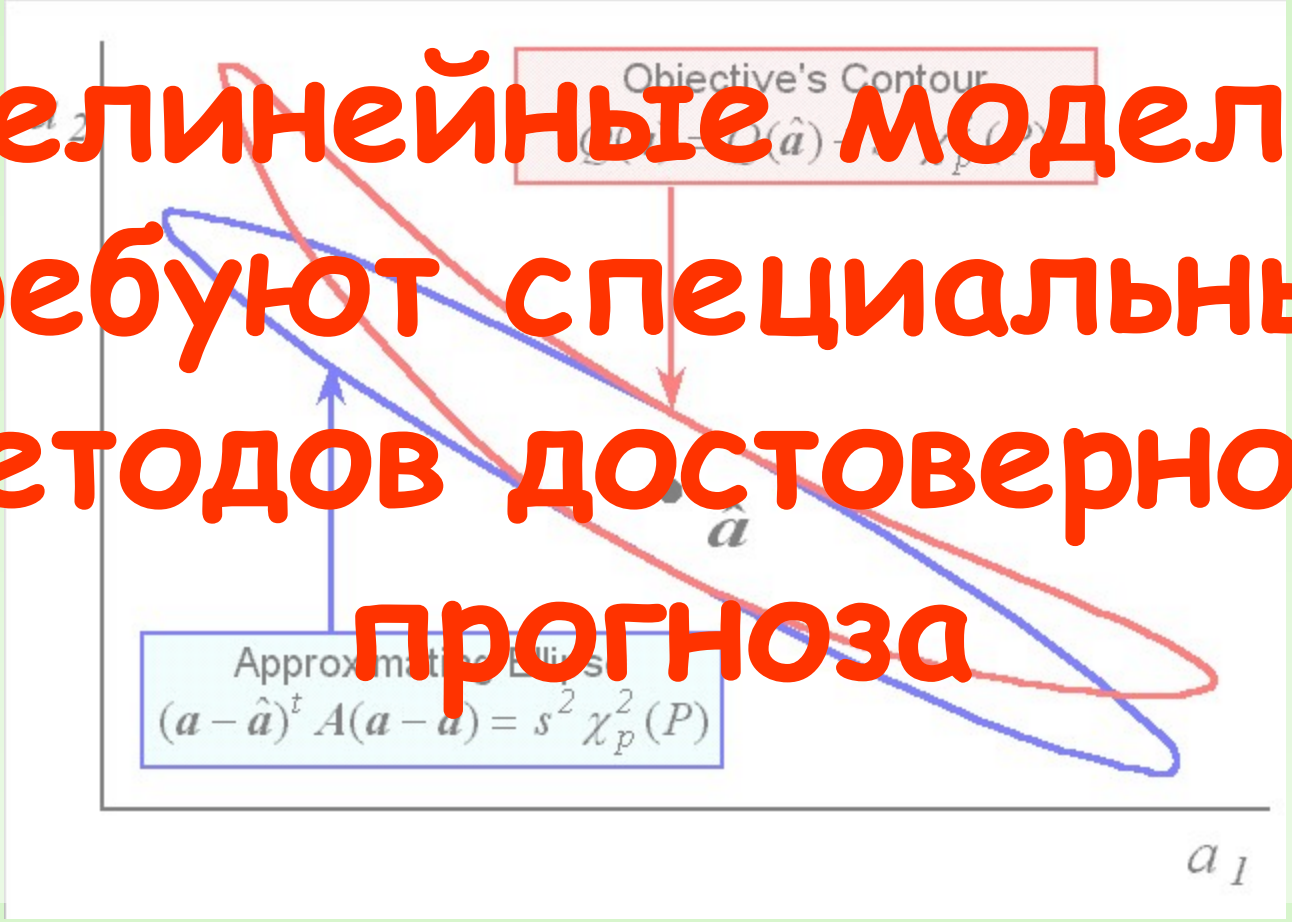
Ли



Достоверный прогноз должен учитывать погрешности в оценках

Имитационное прогнозирование

$$a^* \sim N(\hat{a}, C) \rightarrow a_1^*, \ldots, a_M^* \rightarrow f_1^*, \ldots, f_M^* \rightarrow r(P, x)$$



Нелинейные модели требуют специальных методов достоверного прогноза

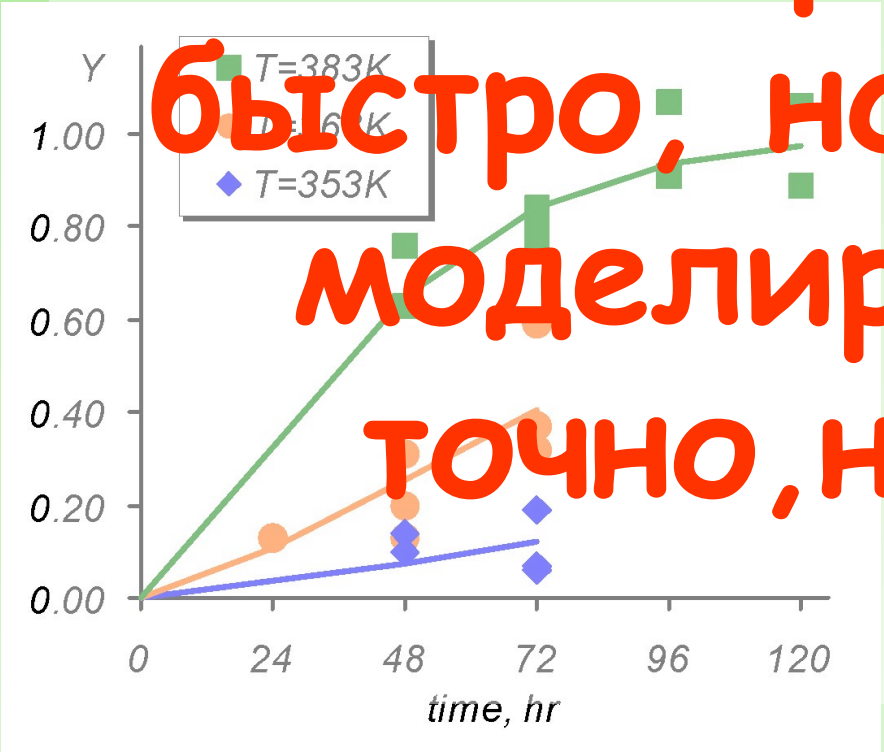
Пример разных методов прогноза

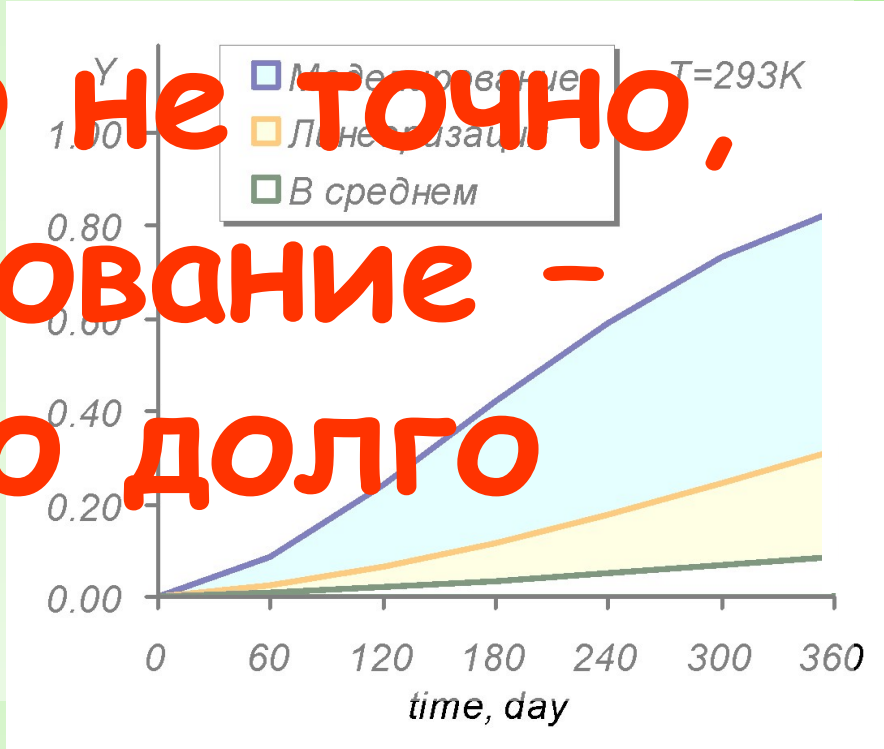
Модель старения резины

$$Y = 1 - e^{-(kt)^a}, \qquad k = e^{k_0 - \frac{E}{RT}}$$

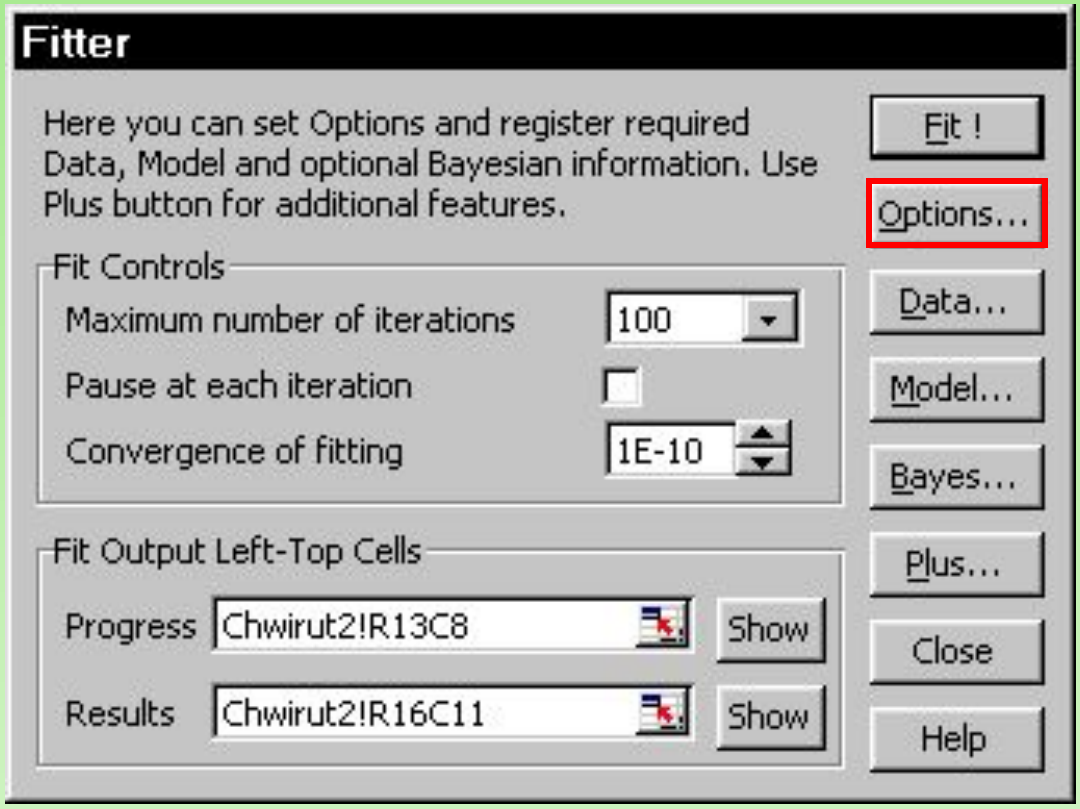
Результат ускоренных испытаний – Выходная предельная граница

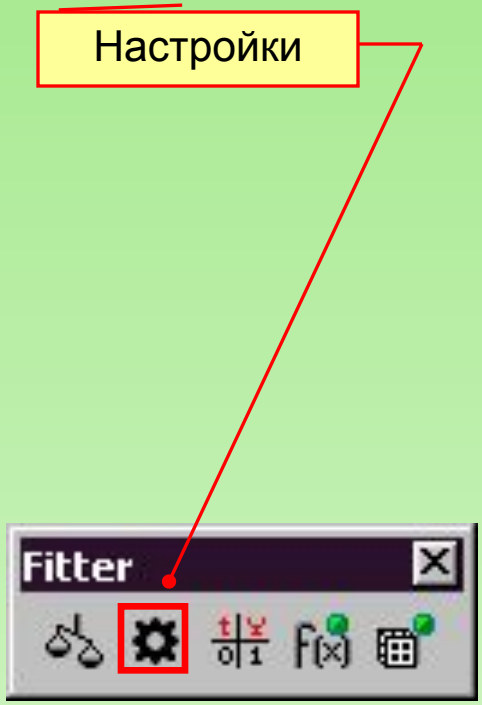
Линеаризация – быстро, но не точно, моделирование – точно, но долго



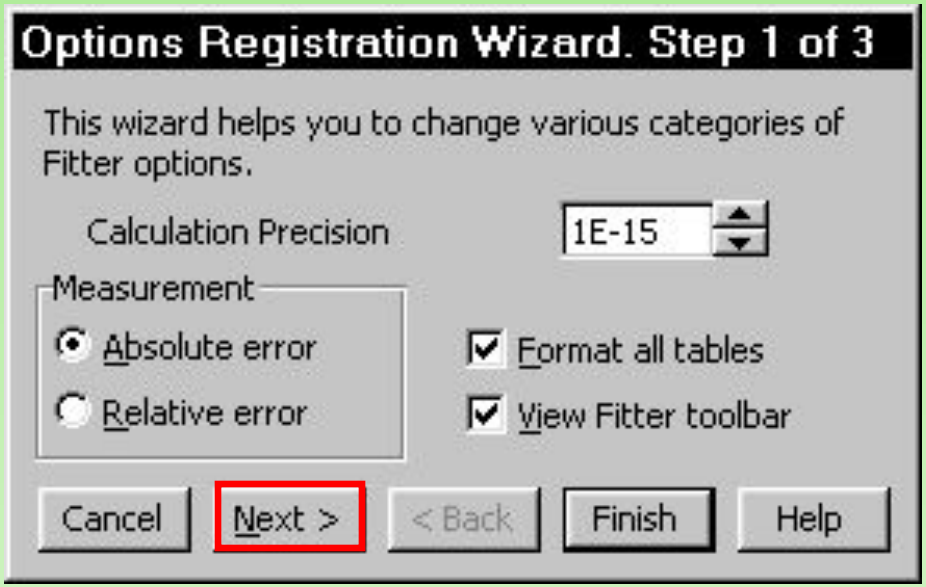


Регистрация настроек прогноза



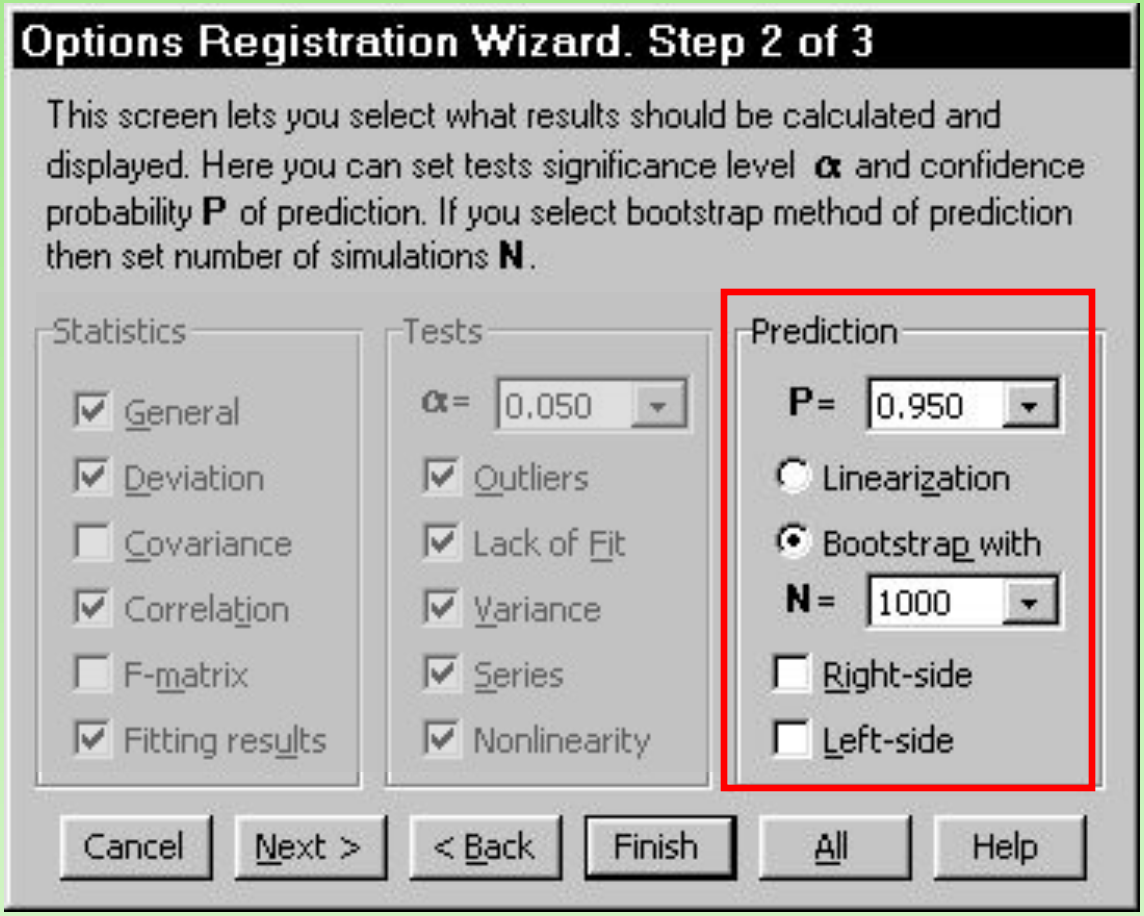


Регистрация настроек . Шаг 1



Продолжить

Регистрация настроек прогноза. Шаг 2



Достоверность прогноза

Метод - линеаризация

Метод - моделирование

Число повторов

Правая граница

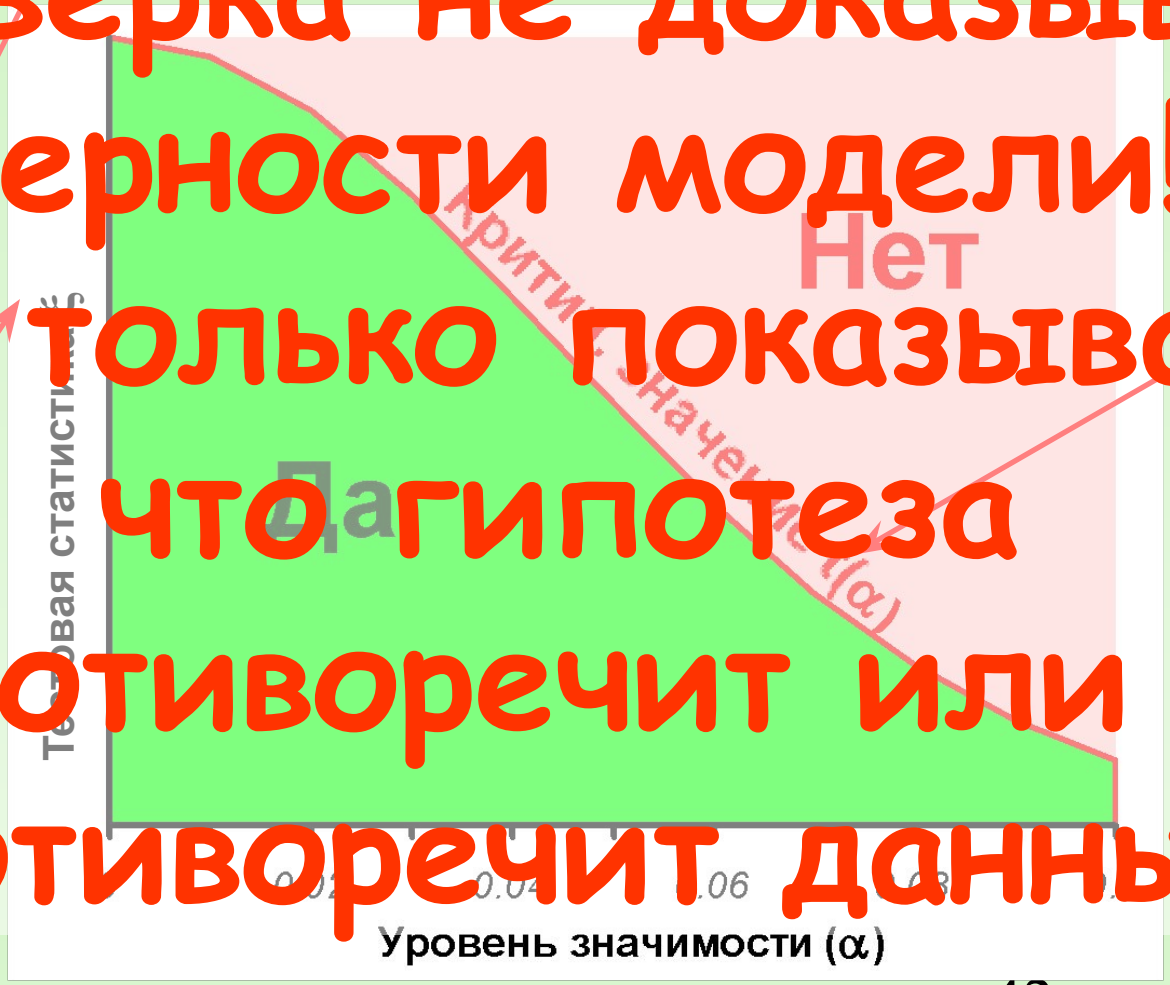
Левая граница

Закончить

7. Тестирование модели

Проверка гипотез

Статистика ξ сравнивается с критическим значением $t(\alpha)$



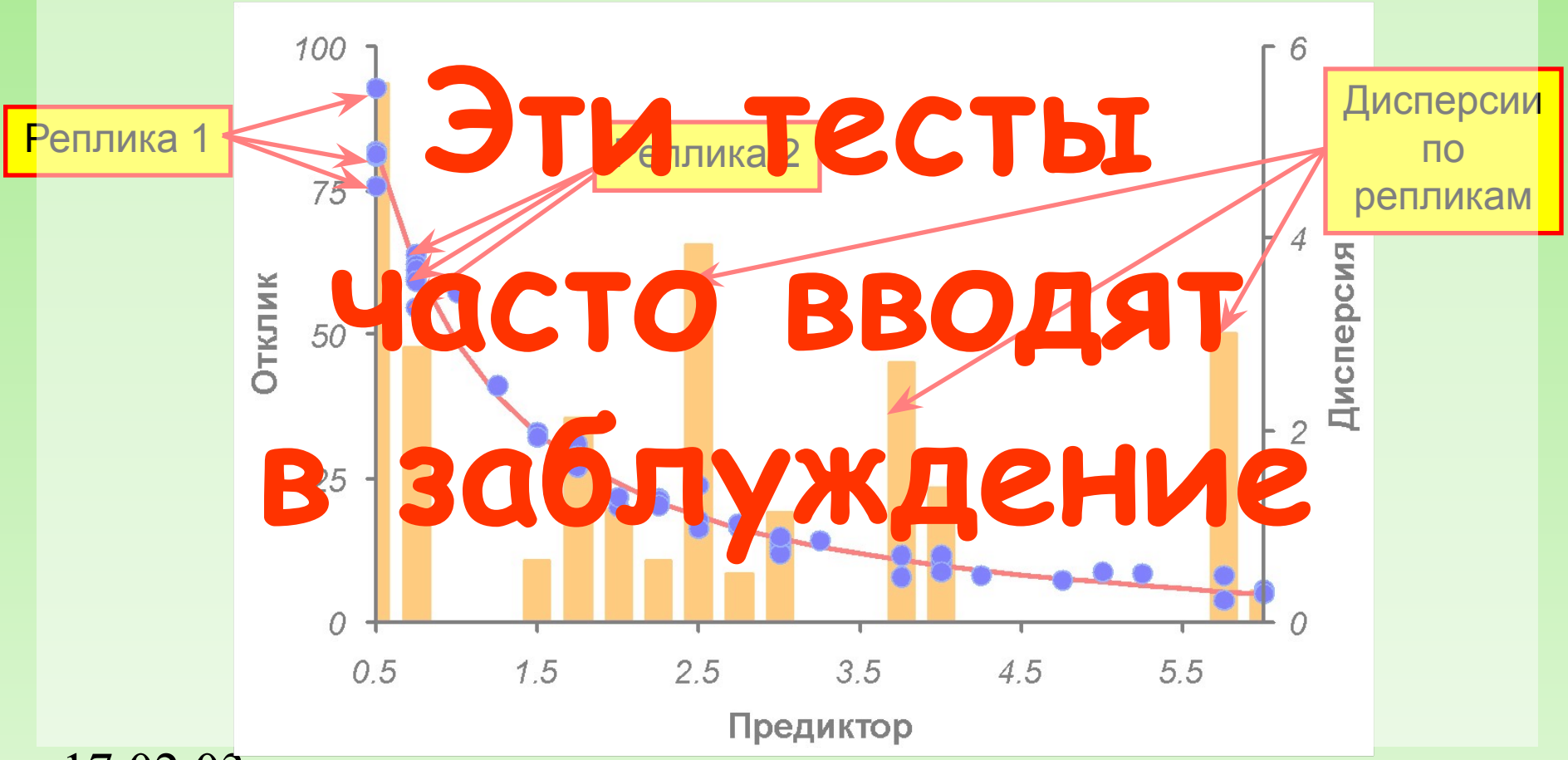
Из эксперимента

Из теории

Проверка не доказывает верности модели! Она только показывает, что гипотеза противоречит или не противоречит данным!

Адекватность и гомоскедастичность

Эти тесты используют выборочные дисперсии и они не могут быть проверены без **реплик**!

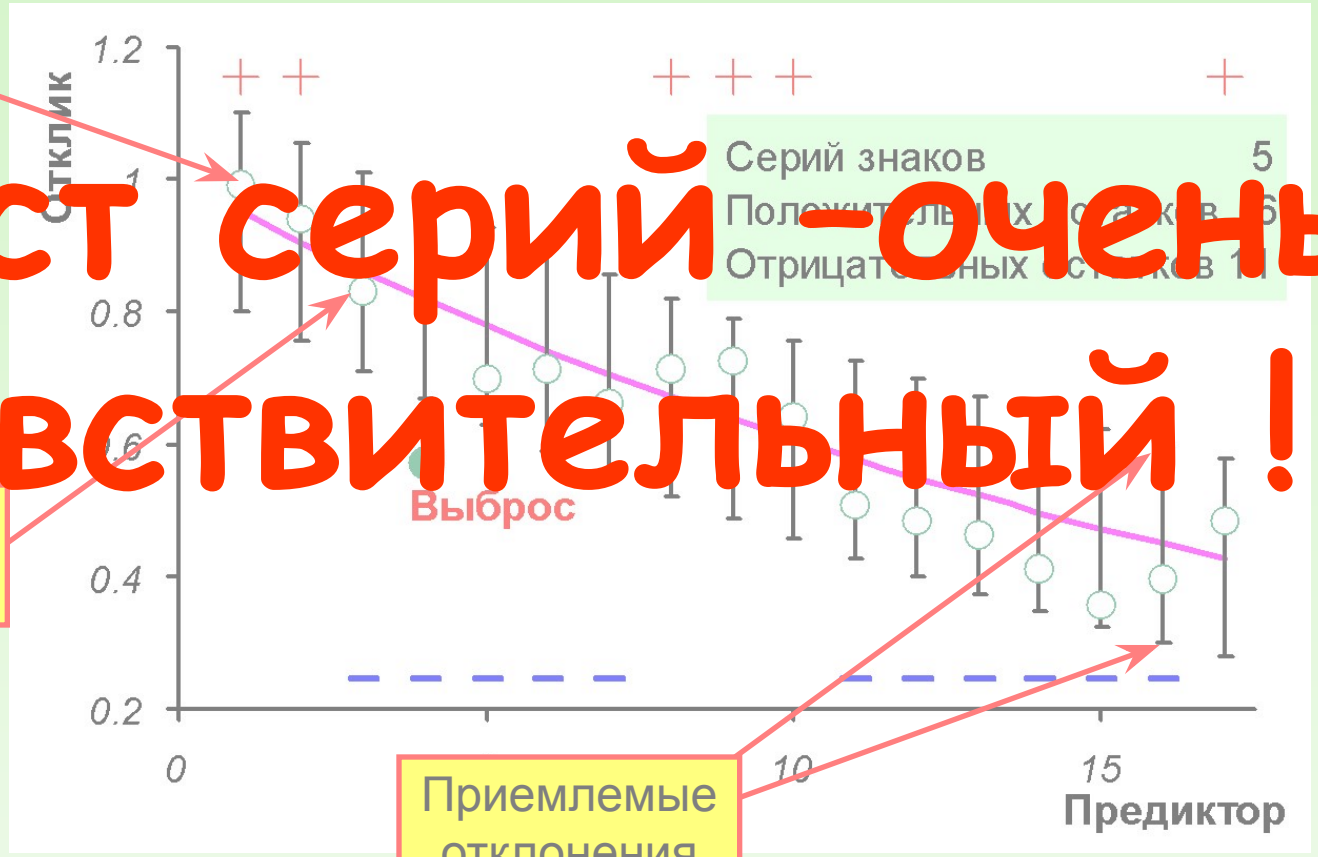


Выбросы и тест серий

Эти тесты используют только остатки, поэтому они могут быть проверены и без реплик

Положительные остатки

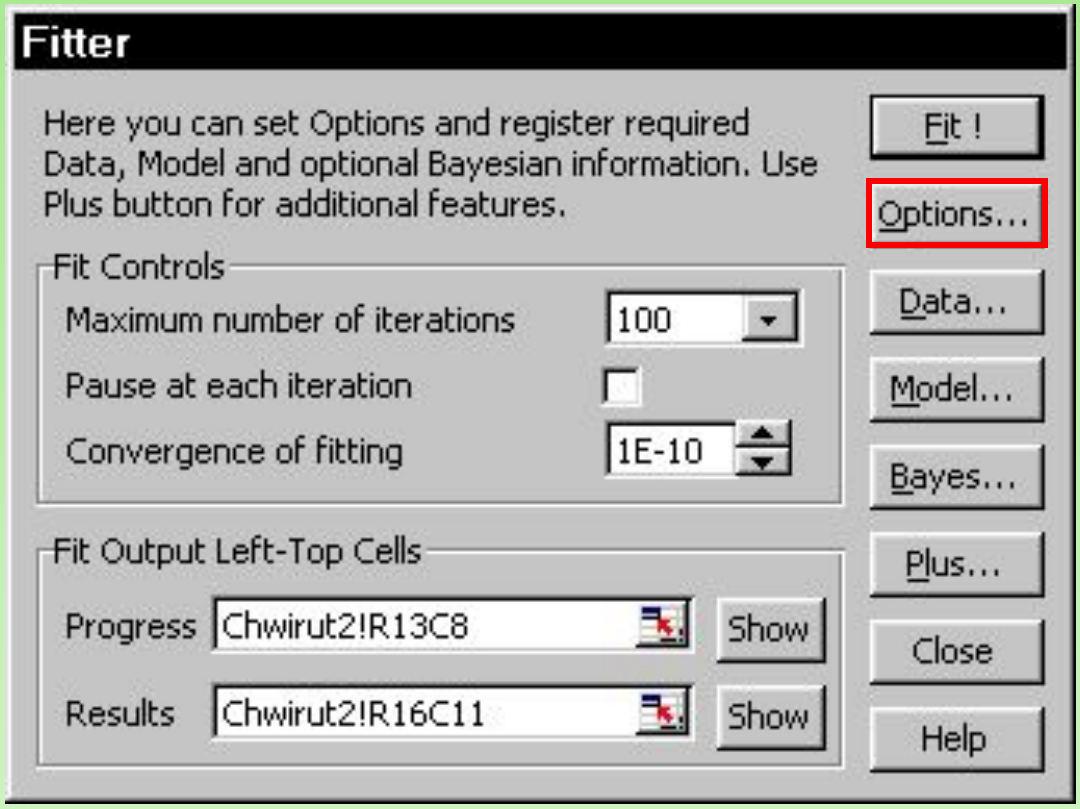
Отрицательные остатки

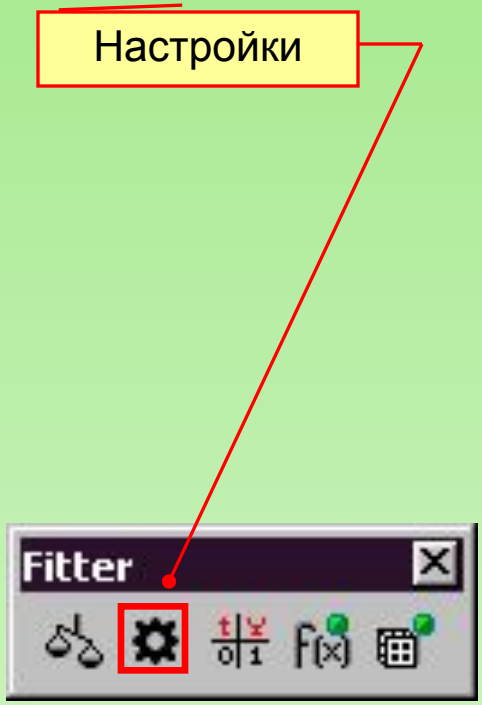


Тест серий – очень чувствительный !

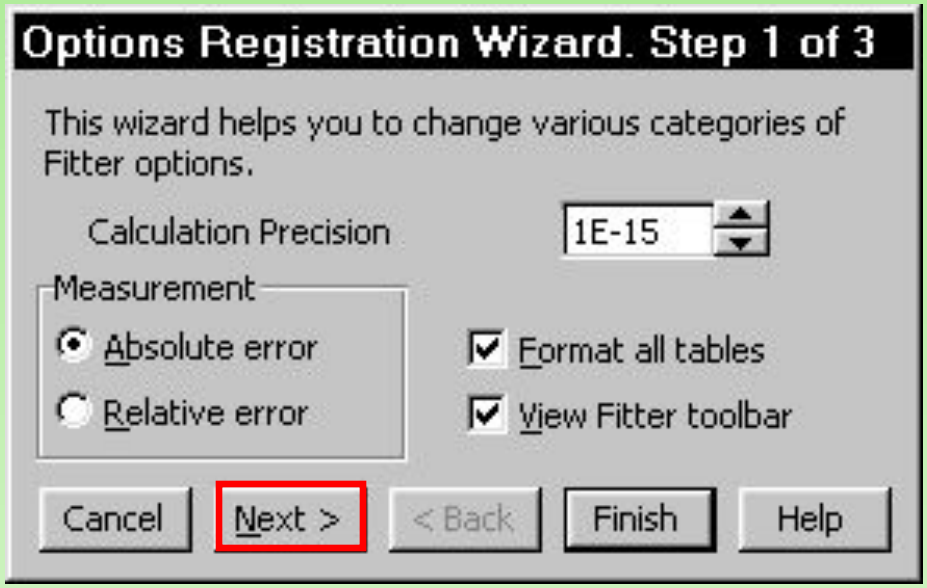
Приемлемые отклонения

Регистрация настроек проверки гипотез



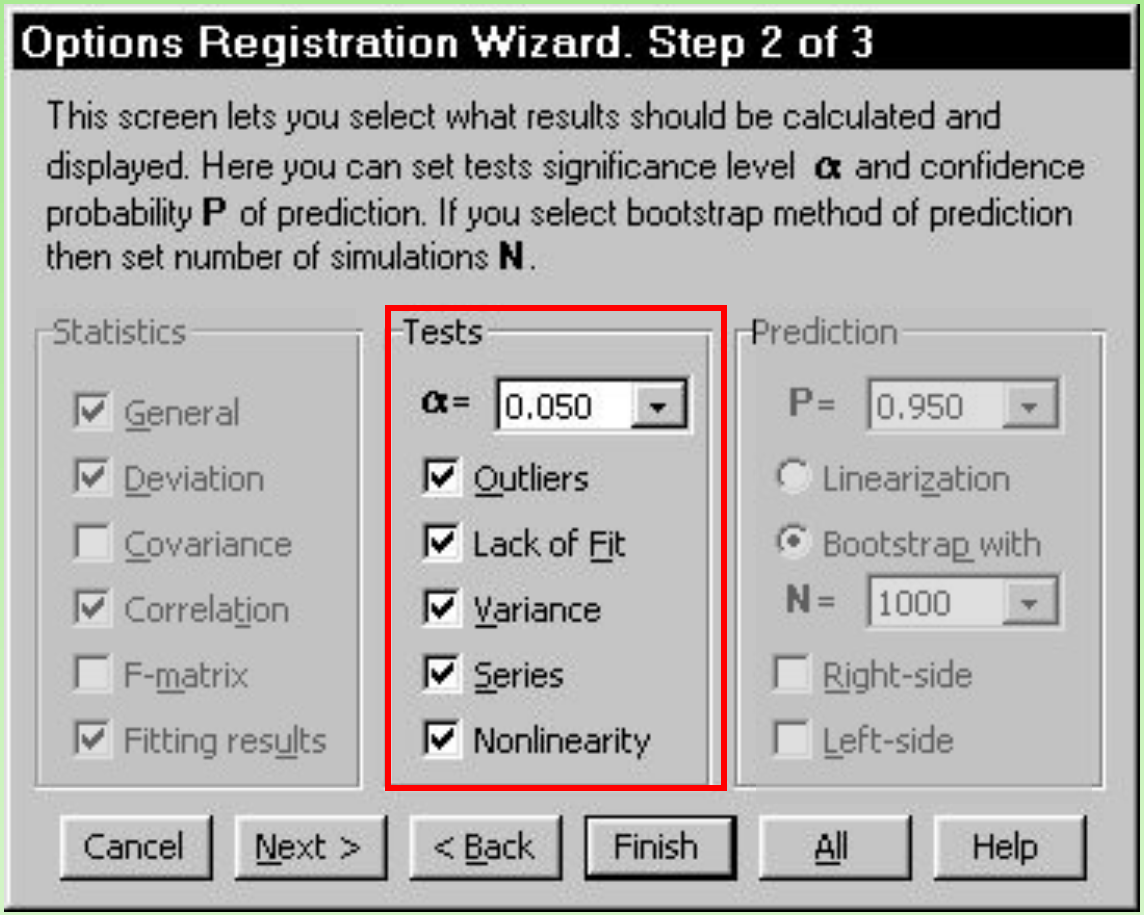


Регистрация настроек . Шаг 1



Продолжить

Регистрация настроек гипотез. Шаг 2



Уровень значимости

Тест выбросов

Тест адекватности

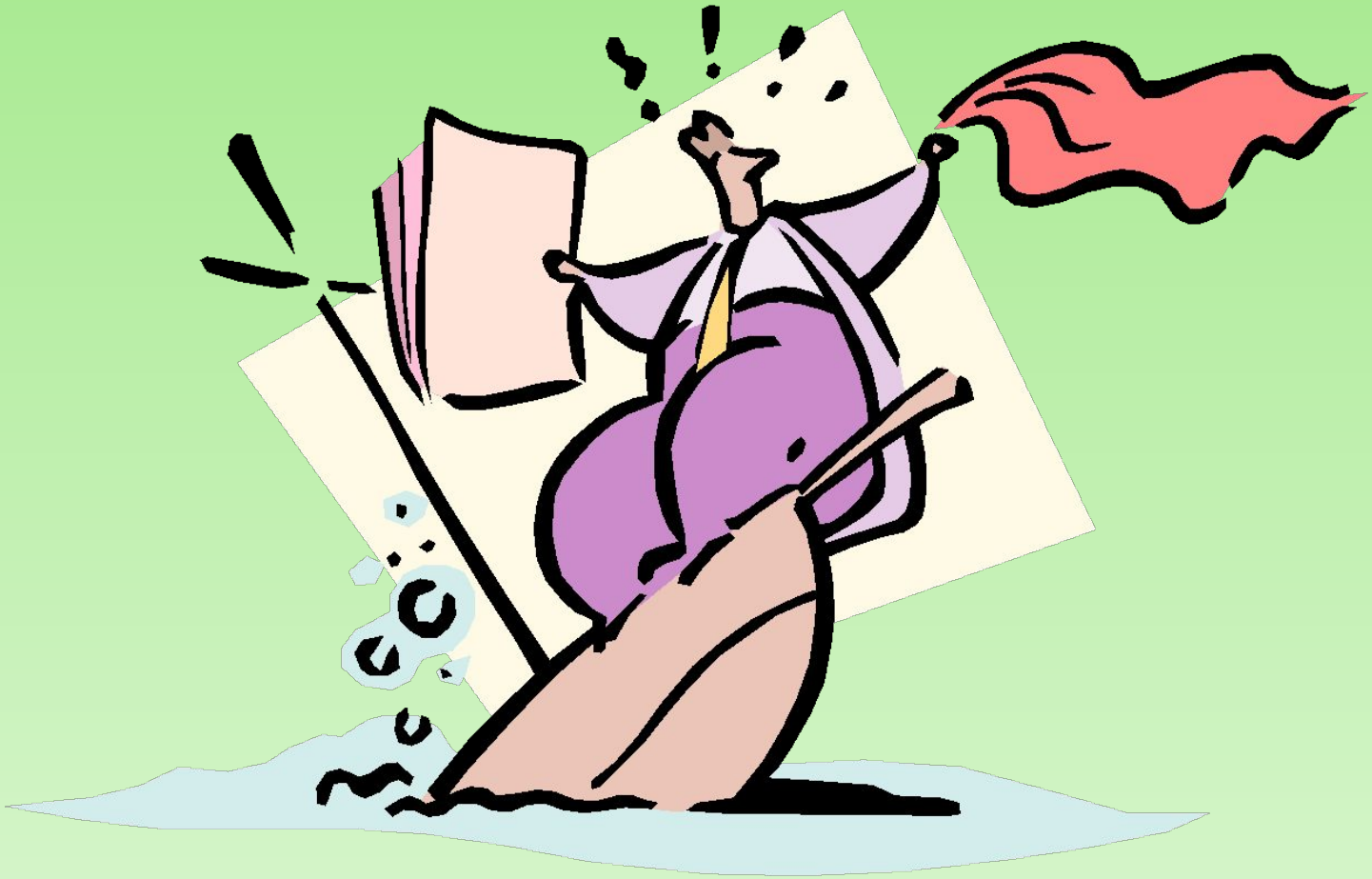
Тест дисперсий

Тест серий

Коэффициент нелинейности

Закончить

8. Мультиколлинеарность



Что такое мультиколлинеарность

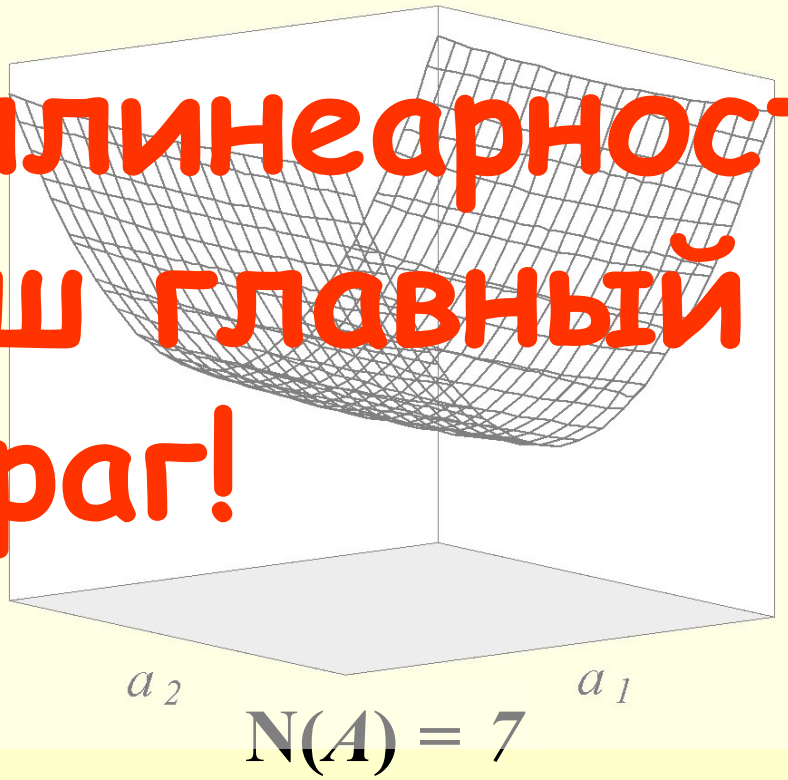
Мультиколлинеарность – это вырождение матрицы A

Разброс собственных значений:

$$N(A) \equiv \log_{10}\left(\frac{\lambda_{max}}{\lambda_{min}}\right)$$

это мера вырождения матрицы ***A***

Целевая функция $Q(a)$



Мультиколлинеарность – это наш главный враг!

Причины мультиколлинеарности

“Строгая” мультиколлинеарность	“Нестрогая” мультиколлинеарность
$y = a_1 a_2 x$	$y = a_1\left(1 - e^{-a_2 x}\right) \approx a_1 a_2 x$

Планирование эксперимента!

Подготовка данных и модели

$$((a+b)+c)+d \neq a+(b+(c+d)) \quad \text{т.к.} \quad 1+10^{-20}=1$$

64-х битное представление числа в компьютере

Цель сделать матрицу A регулярной, т.е. уменьшить $\mathrm{N}(A)$

Средства

Шкалирование	$X \rightarrow mX$
Центрирование	$X \rightarrow X - X_0$
Перепараметризация	$a \rightarrow \phi(a) \quad x \rightarrow \psi(x) \quad y \rightarrow \chi(y)$

Компьютеру трудно – нужно ему помочь!

Пример: закон Аррениуса

Обычная форма записи	Улучшенная форма записи
$k \exp\left(-\frac{E}{RT}\right)$	$\exp(a_1 - a_2 X)$
Шкалирование и центрирование	$X = \frac{\frac{1000}{T} - X_0}{\frac{1}{n}\sum_{i=1}^{n}\frac{1000}{T_i}}$
Перепараметризация	$a_1 = \ln(k) - a_2 X_0 \qquad a_2 = \frac{E}{1000R}$

Простые преобразования, дающие большой выигрыш!

1

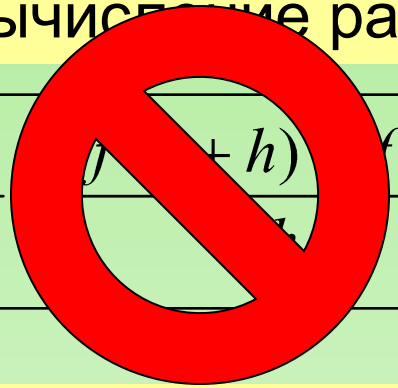
$N(A) = 20$	$N(A) = 2$

Вычисление производных и точность

N(A)	A^{-1}	$y=f(a,x)$	$0=f(y,a,x)$	$dy/dx=f(y,a,x)$
10	10+2=12	12+0=12	12+2=14	14+2=16
6	6+2=8	8+0=8	8+2=10	10+2=12

1) Численное вычисление разностных производных

$$\frac{\partial f}{\partial a} \quad f(a+h) \quad f(a)$$



2) Автовычисление аналитических производных

```
f=exp(-a*t)
```

```
df/da=-t*exp(-a*t)
```

9. Прогноз срока службы ПВХ изоляции

Постановка задачи

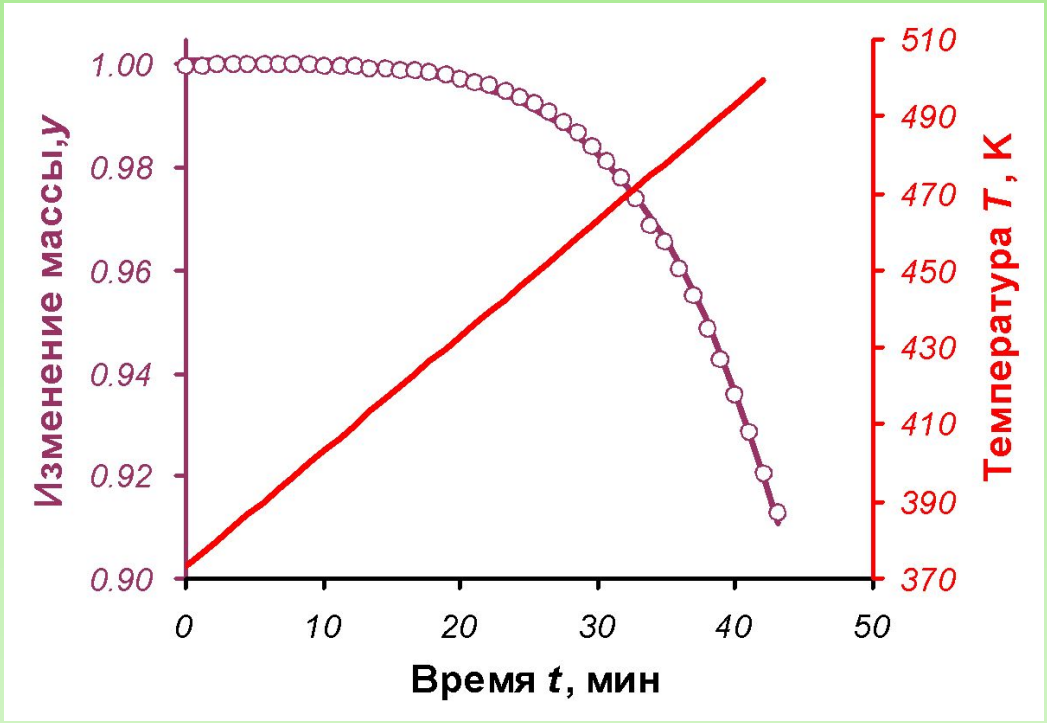
Объект	Кабельная изоляция из ПВХ
Цель	Прогнозирование срока службы
Эксперимент	Термогравиметрический метод
Обработка	Нелинейная регрессия + Fitter

Предсказать значения, которые нельзя измерить – это экстраполяция!

ТГА эксперимент и данные

ТГА эксперимент

ТГА данные



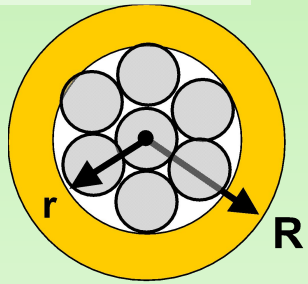
Это схема эксперимента, а не грешник в аду!

Переменные в примере ТГА

Измеряемые	Оцениваемые
Отклик	**Промежуточные**
$y=m/m_0$ Изменение массы	c Концентрация катализатора
Факторы	**Параметры**
t Время	y_0 Начальное значение y
C_0 Исходная концентрация	k_0 Константа скорости испарения
v Скорость нагрева	E Энергия активации
T_0 Начальная температура	
F Удельная поверхность образца	

Нелинейная задача малой размерности!

Удельная поверхность образца $F=\frac{S}{V}=\frac{2R}{R^2-r^2}$



Модель испарения пластификатора

Закон испарения	$\frac{dy}{dt} = -kC,\ y(0) = y_0$
Изменение объема	$C = 1 - \frac{1 - C_0}{y}$
Закон Аррениуса	$k = A\exp\left(-\frac{E}{RT}\right)$
Рост температуры	$T = T_0 + vt$

Диффузия здесь не важна!

Пример ТГА на рабочем листе Fitter

Data

v	T0	C	F	t	y	f	Left
3	373	0.3	2.4	0.00	1.000	1.001	1.001
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]01
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]000
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]000
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]000
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]
3	373	0.3	2.4	[illegible]	[illegible]	[illegible]	[illegible]000
0	293	0.4	2	[illegible]	[illegible]	[illegible]	[illegible]915
0	293	0.4	2	[illegible]	[illegible]	[illegible]	0.844
0	293	0.4	2	8E+06		0.821	0.788
0	293	0.4	2	1E+07		0.781	0.746
0	293	0.4	2	1E+07		0.748	0.714

$$\frac{dy}{dt} = -kC, \quad y(0) = y_0$$

$$C = 1 - \frac{1-C_0}{\dots}$$

$$k = F\exp\left(k_0 - \frac{E}{RT}\right)$$

$$T = T_0 + vt$$

Просто!

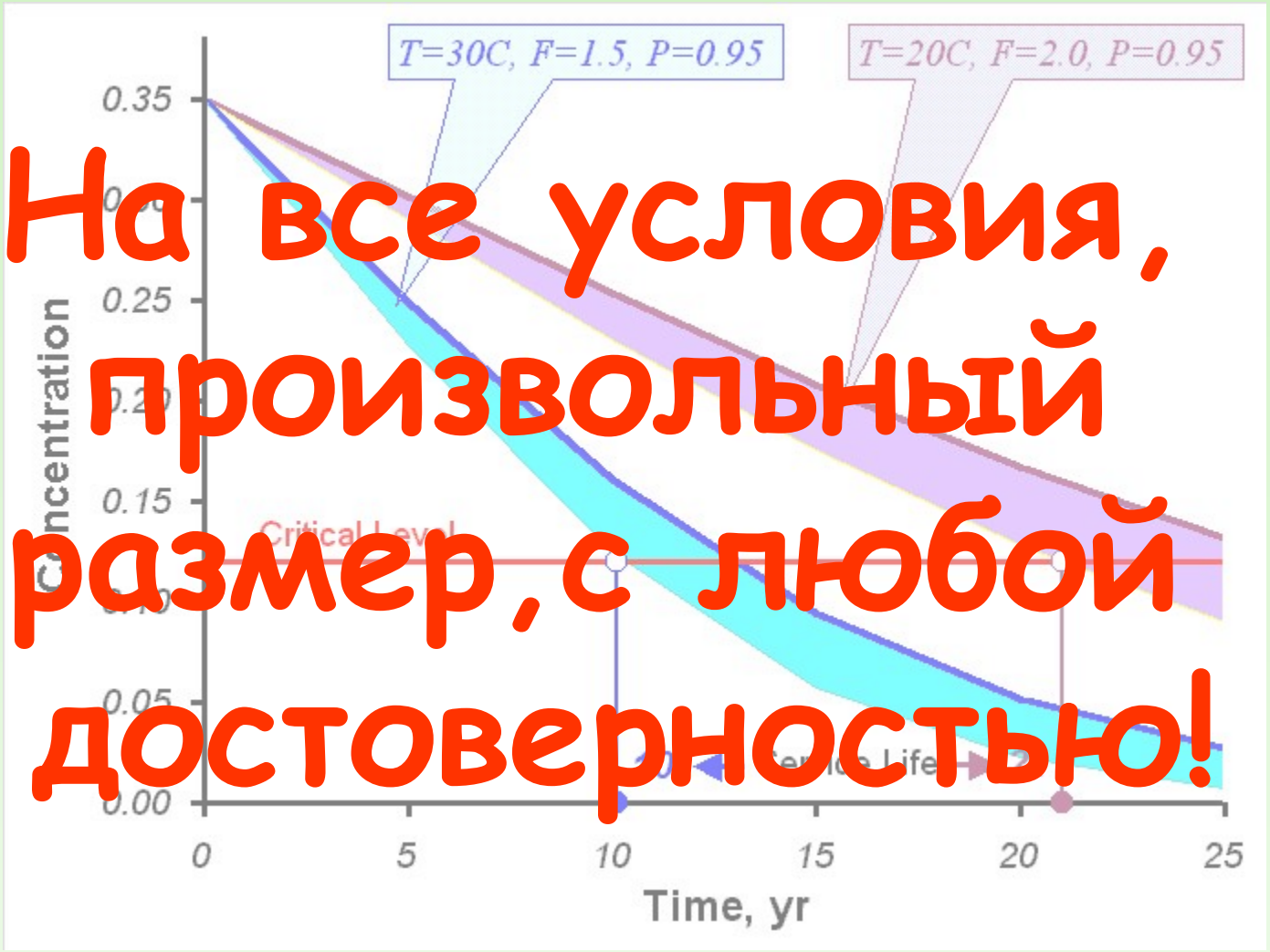
Fitter

```
'TGA Desorbtion Model
D[y]/D[t]=-k*[1-(1-C0)/y];y(0)=y0
k=F*exp(k0-E/R/T)
T=T0+v*t
R=1.98717
y0=?
k0=?
E=?
```

Parameters estimation

Name	Initial	Final	Deviation
y0	1	1.00088	0.0002
k0	10	13.9964	0.24016
E	10000	18052.3	225.424

Прогноз срока службы



На все условия, произвольный размер, с любой достоверностью!

10. Выводы

Формальный подход

Загадочная природа

Содержательный подход

Спасибо за внимание!