Gnumeric: электронная таблица для всех

И.А.Хахаев, © 2007-2010

# 6 Регрессионный анализ в Gnumeric

## *6.1 Небольшое теоретическое введение*

Всегда полезно знать, что и почему вычисляется в той или иной задаче. Поэтому сначала рассмотрим некоторые теоретические основы регрессионного анализа.

Линейный парный регрессионный анализ заключается в определении параметров эмпирической линейной зависимости (1), описывающей связь между некоторым N числом пар значений $x_i$ и $y_i$, обеспечивая при этом наименьшую среднеквадратическую погрешность (метод наименьших квадратов).

$$y(x)=a\cdot x+b \tag{1}$$

Графически это выглядит как проведение прямой в «облаке» точек с координатами $x_i$, $y_i$ так, чтобы величина всех отклонений между значениями y на этой прямой при имеющихся значениях $x_i$ и координатами $y_i$ имеющихся точек отвечала условию (2).

$$U=\sum_{i=1}^{N}(y_i-y(x_i))^2\rightarrow min \tag{2}$$

где $y(x_i)$ – теоретическая зависимость (1). Для этого нужно приравнять к нулю частные производные (3 и 4).

$$\frac{\partial U}{\partial b}=\sum_{i=1}^{N}(y_i-(b+a\cdot x_i)) \tag{3}$$

$$\frac{\partial U}{\partial a}=\sum_{i=1}^{N}(y_i-(b+a\cdot x_i)x_i) \tag{4}$$

Тогда для определения коэффициентов линейной регрессии a и b получаем систему уравнений (5).

$$\begin{cases} b\cdot N+a\cdot\sum_{i=1}^{N}x_i=\sum_{i=1}^{N}y_i \\ b\cdot\sum_{i=1}^{N}x_i+a\cdot\sum_{i=1}^{N}x_i^2=\sum_{i=1}^{N}x_i\cdot y_i \end{cases} \tag{5}$$

Решение этой системы даётся соотношениями 6 и 7.

$$a=\frac{\sum_{i=1}^{N} x_i \cdot - \sum_{i=1}^{N} y_i - N \cdot \sum_{i=1}^{N} x_i \cdot y_i}{\left(\sum_{i=1}^{N} x_i\right)^2 - N \cdot \sum_{i=1}^{N} x_i^2} \quad (6)$$

$$b=\frac{1}{N} \cdot \left(\sum_{i=1}^{N} y_i - a \cdot \sum_{i=1}^{N} x_i\right) \quad (7)$$

Для определения отклонения связи между $x_i$ и $y_i$ от линейной используется коэффициент парной корреляции (8).

$$R=\frac{\sum_{i=1}^{N} x_i \cdot y_i - \left(\sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i\right)/N}{\sqrt{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}} \cdot \sqrt{\sum_{i=1}^{N} y_i^2 - \frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N}}} \quad (8)$$

Если экспериментальная зависимость явно нелинейная, для её интерполяции (аппроксимации) применяются различные нелинейные зависимости (экспоненциальная, степенная с положительными или отрицательными показателями степени, полиномиальные различных порядков и пр.). При этом интерполяционная функция «линеаризуется», т. е. сводится к виду (1) путём замены переменных. Соответственно пересчитываются значения экспериментальных точек и коэффициент парной корреляции показывает успешность этого преобразования. Поскольку знак коэффициента парной корреляции при оценке качества линеаризации не является существенным, часто используется значение $R^2$.

## *6.2 Реализация вычислений на модели*

Вычисление параметров линейной регрессии уже рассматривалось в главе «Инструменты Gnumeric для статистиков», поэтому здесь рассмотрим подробнее процесс добавления и настройки параметров линий регрессии на график с экспериментальными данными. В качестве исходных данных используем таблицу, которая уже применялась в главе про статистику при описании инструментов предсказания и регрессии (рис. 6.1).

| X | Y |
|---|---|
| 1 | 80,54 |
| 2 | 54,21 |
| 3 | 51,01 |
| 4 | 25,26 |
| 5 | 18,43 |
| 6 | 13,11 |
| 7 | 12,75 |
| 8 | 9,07 |
| 9 | 6,4 |
| 10 | 4,43 |
| 11 | 3,39 |
| 12 | 2,16 |
| 13 | 1,7 |
| 14 | 1,14 |
| 15 | 0,65 |

Рисунок 6.1.

На рис. 6.2 показан график с исходными данными (круглые точки).

Поскольку линейная регрессия для таких данных, очевидно, даёт плохие результаты, будем пытаться использовать нелинейные модели. Тогда этот процесс можно будет называть «non-linear fitting» – «нелинейная подгонка».

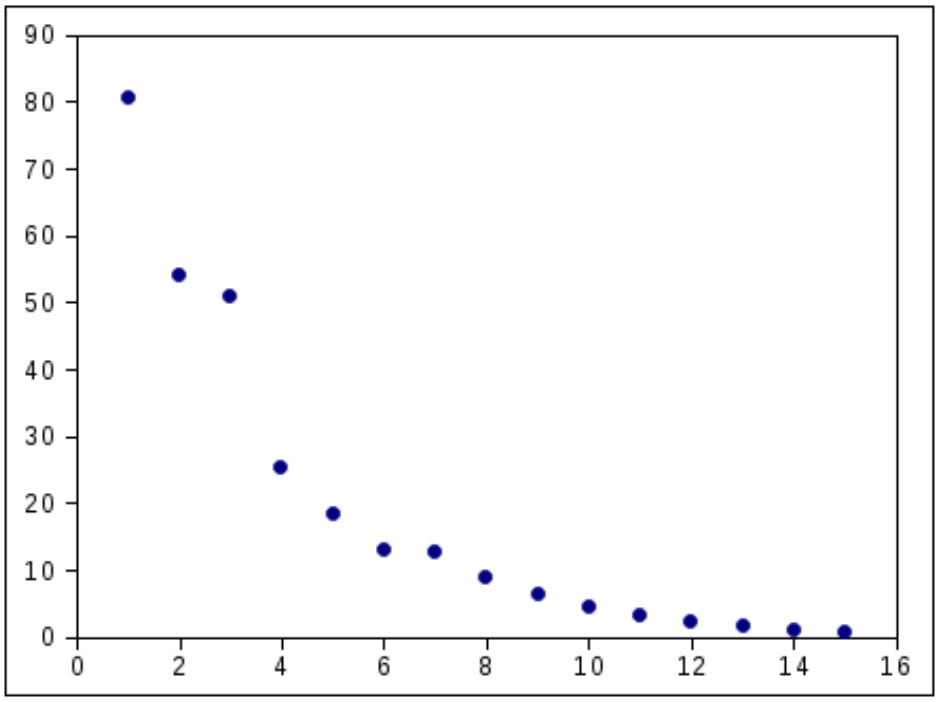


Рисунок 6.2. График исходных данных

Для добавления кривых регрессии вызовем диалог настройки графика, выберем серию исходных данных (Y) и используем кнопку «+Добавить» для выбора добавляемого объекта (рис. 6.3).

Во вложенном меню «Trend Line» («линия тренда») имеется набор классов

кривых (уравнений интерполяции, рис. 6.4).

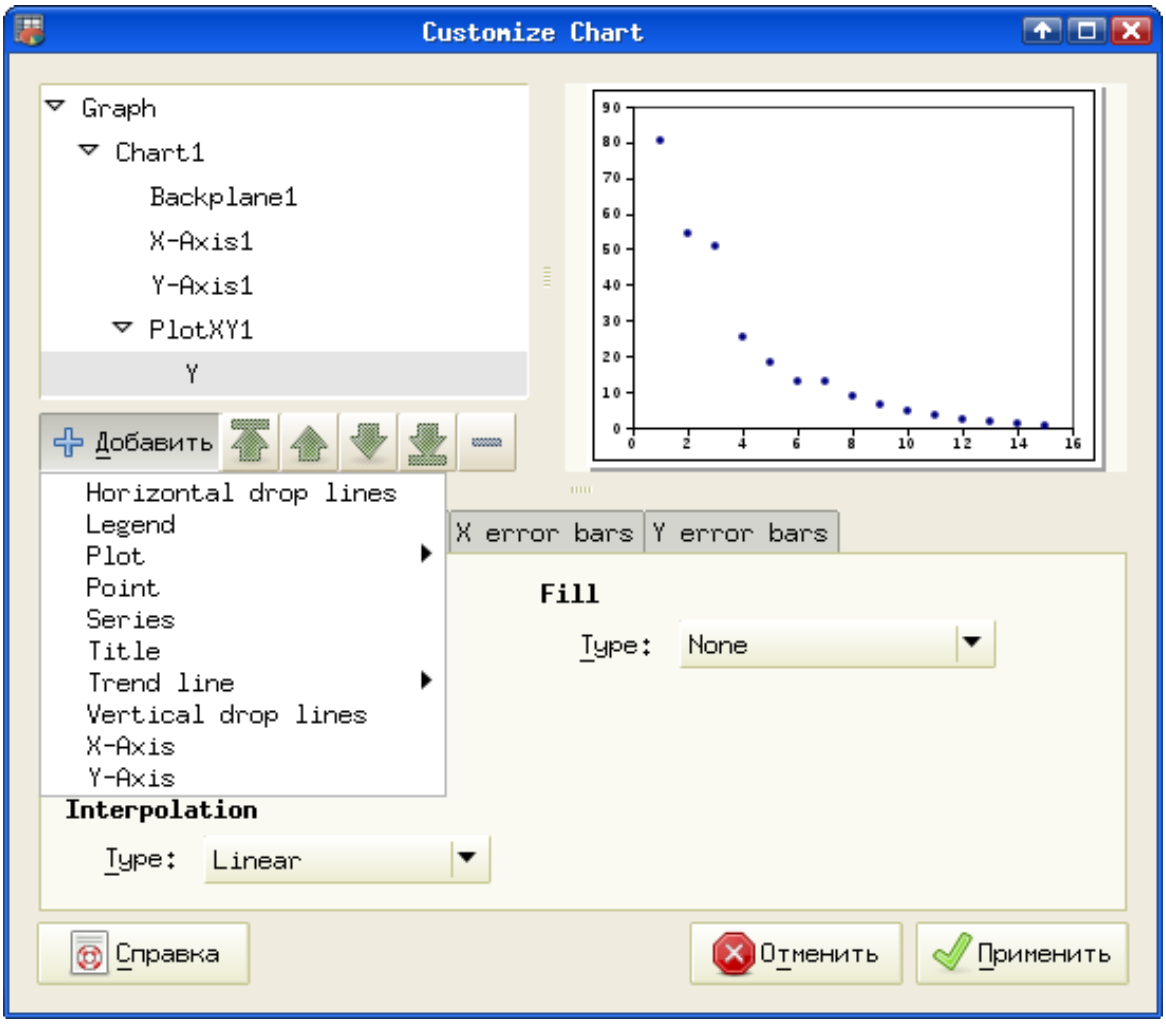


Рисунок 6.3. Выбор добавляемого объекта

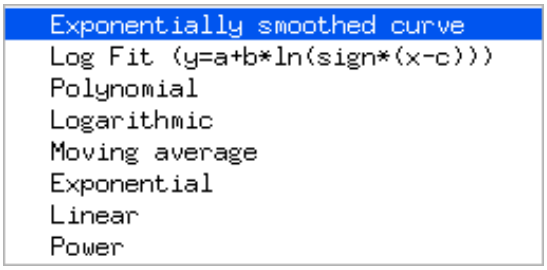


Рисунок 6.4. Определение класса интерполяционных функций

Заметим, что в списке вариантов присутствуют «Экспоненциальное сглаживание» и «Скользящее среднее», которые рассматривались в главе про статистику.

В качестве первой попытки описания экспериментальных данных выберем

вариант интерполяции полиномом (Polynomial) 3-го порядка (рис. 6.5).

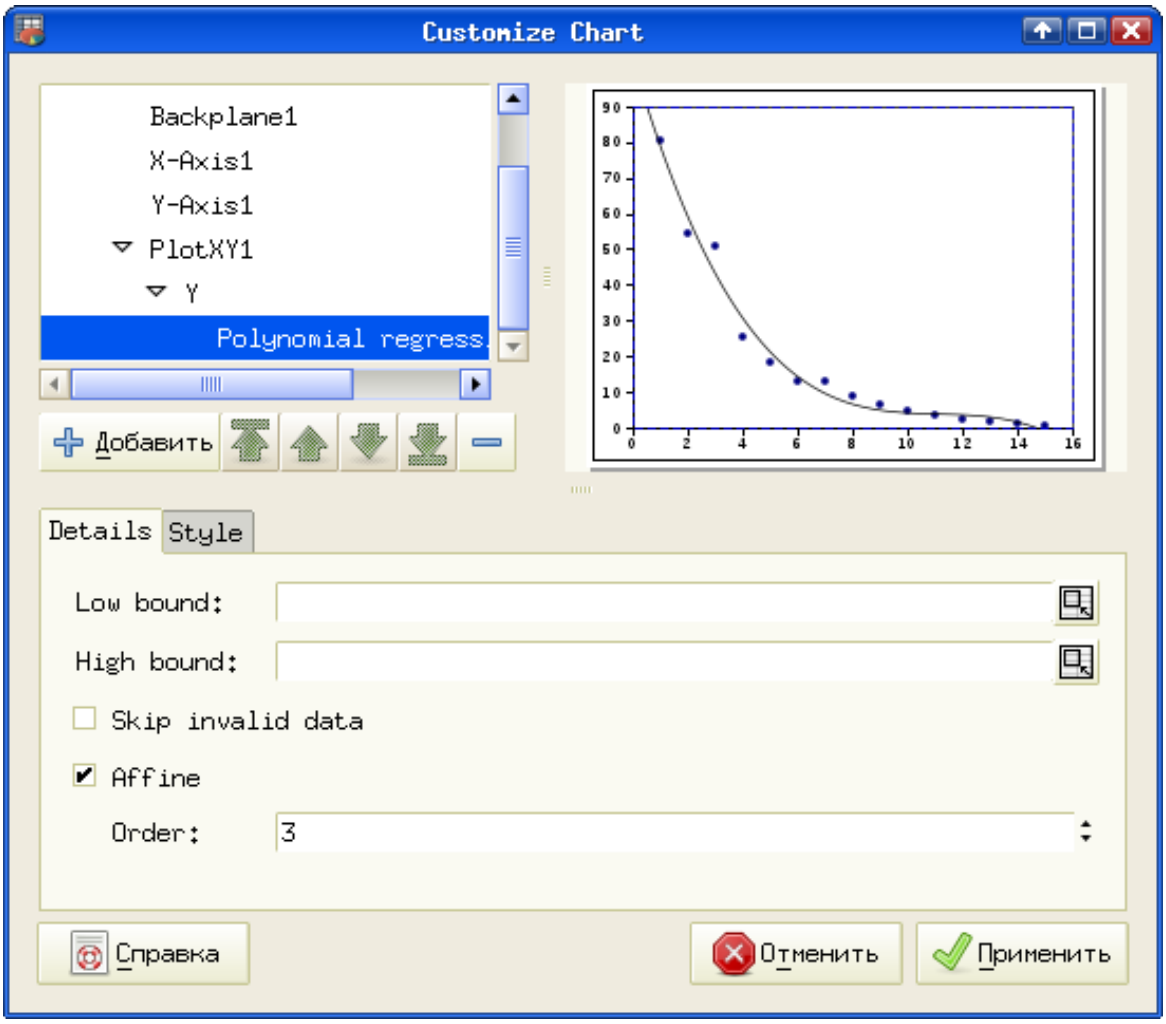


Рисунок 6.5. Настройка полиномиальной регрессии и предварительный вид графика

Список «Order» («Порядок») позволяет выбрать максимальную степень аргумента (порядок) в полиноме, а на вкладке Style (Стиль) можно настроить внешний вид линии.

Для того чтобы узнать коэффициенты полинома ещё раз нажмём кнопку «+Добавить» и увидим, что в списке объектов появился объект Equation (Уравнение), как показано на рис. 6.6

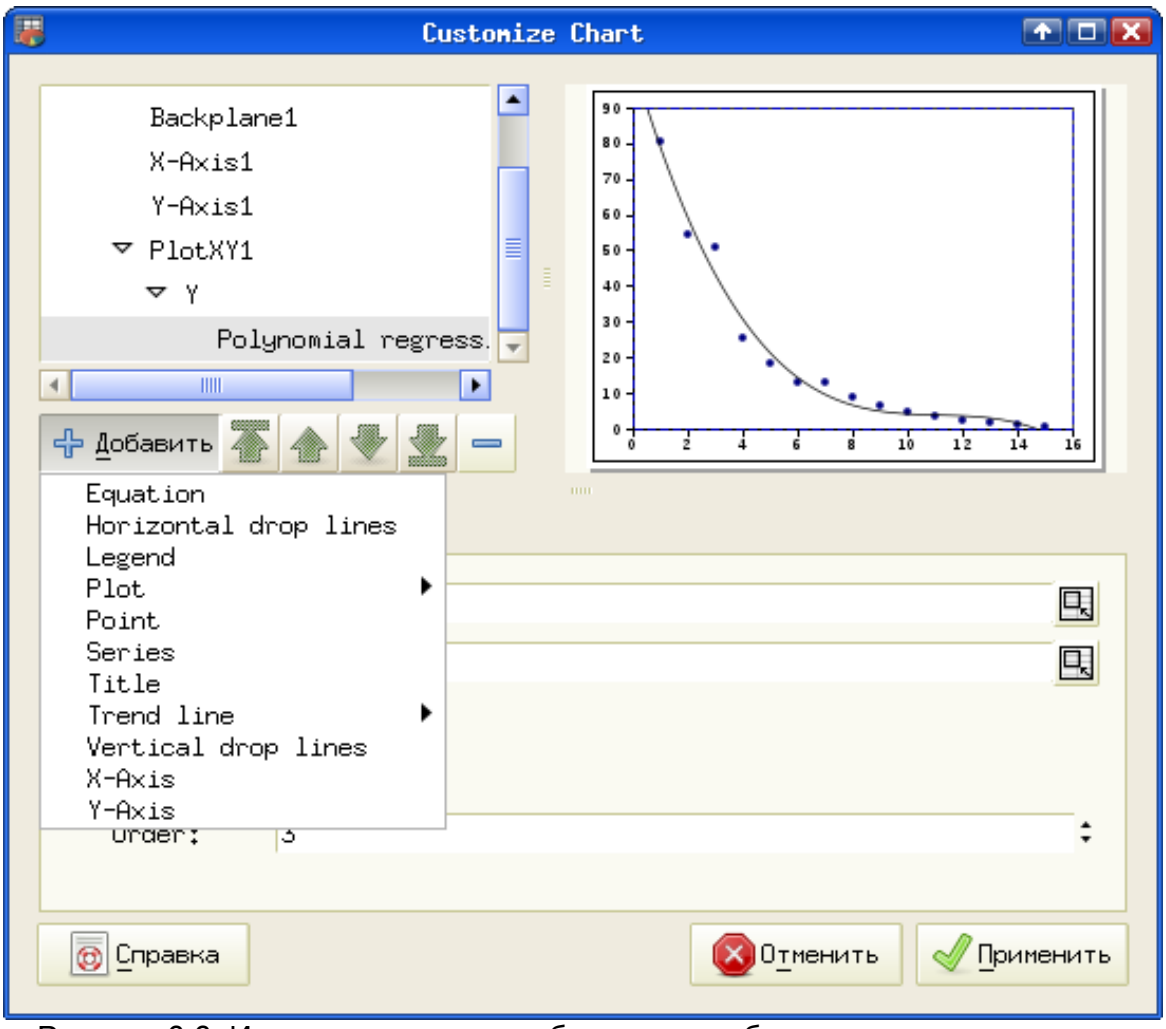


Рисунок 6.6. Изменение списка добавляемых объектов в зависимости от выбранного объекта графика

Добавляемое на график уравнение кривой имеет собственный диалог настроек (рис. 6.7).

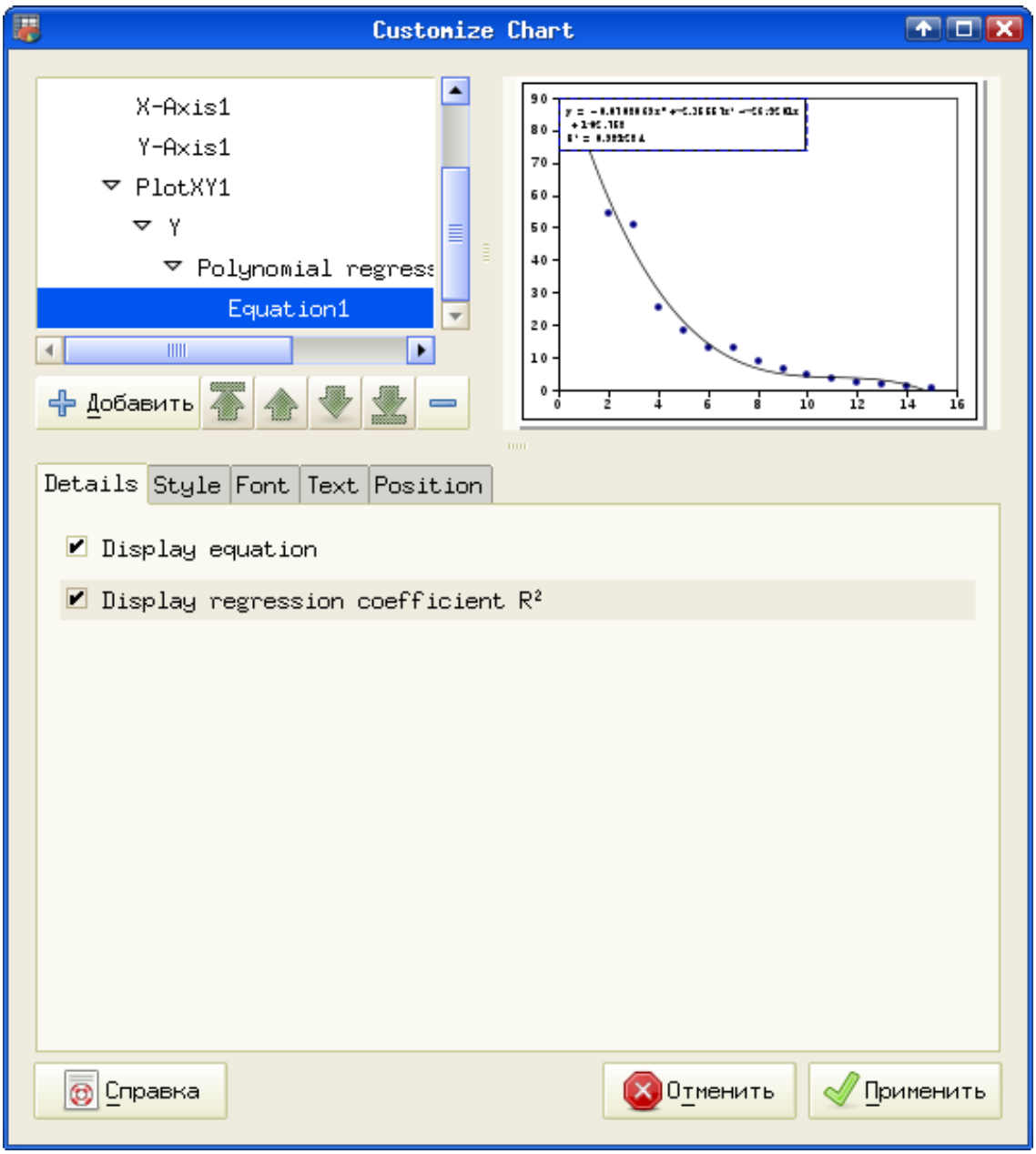


Рисунок 6.7. Добавление уравнения кривой на график

Режим «Показывать коэффициент регрессии $R^2$» позволяет вывести под уравнением значение коэффициента парной корреляции, характеризующего «качество» интерполяции. Чем ближе это значение к 1, тем лучше подобрано уравнение регрессии.

На вкладке Position (Расположение) можно задать желаемое место уравнения на графике в относительных единицах (рис. 6.8).

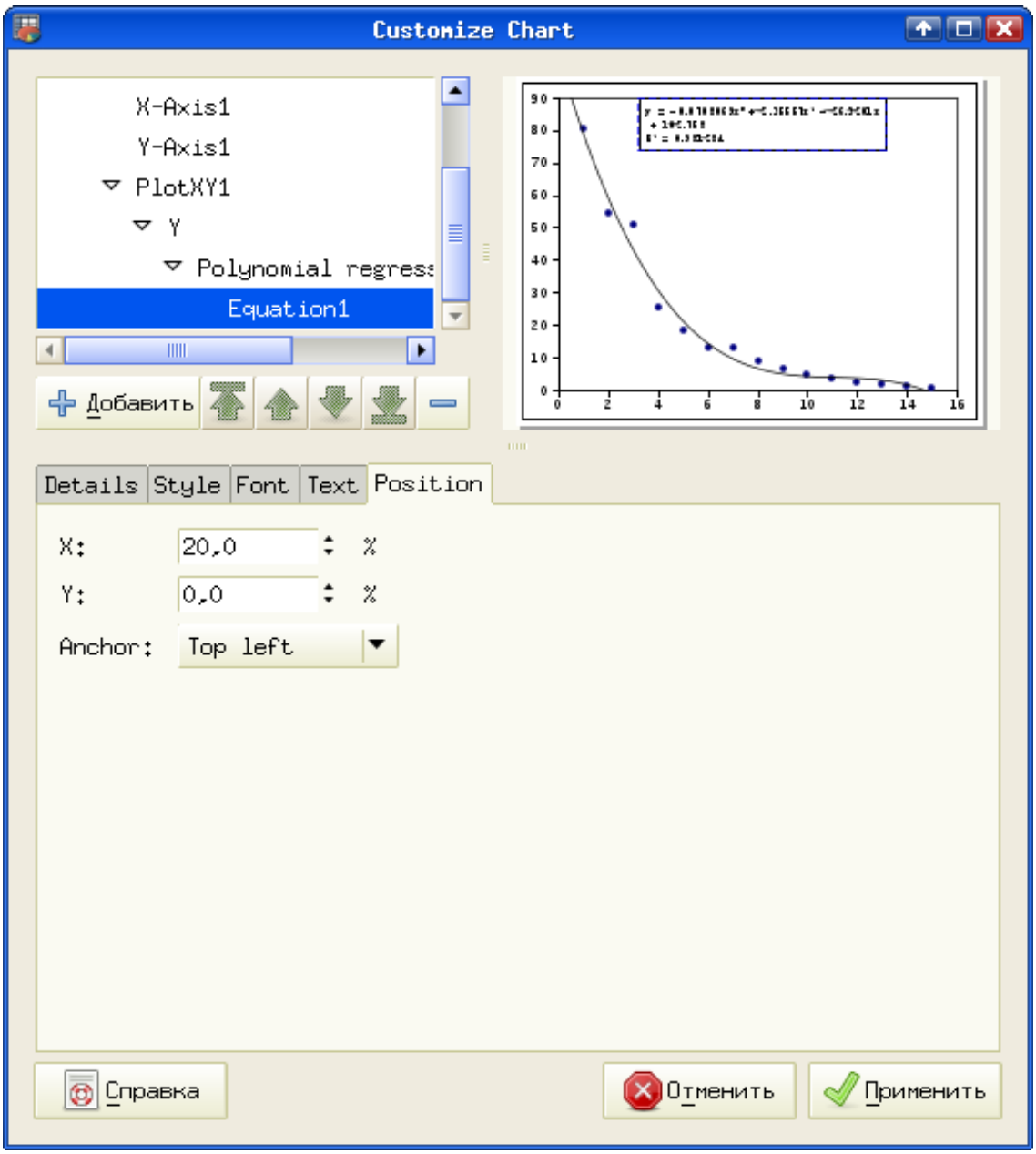


Рисунок 6.8. Настройка расположения уравнения

На вкладках Style (Стиль) и Font (Шрифт) задаётся стиль обрамления области с уравнением и шрифт для отображения уравнения.

Для сравнения добавим интерполяцию степенной функцией (Power), установив стиль линии «точки» и толщину в 2 точки экрана (рис. 6.9). В этом случае модель имеет вид $y(x)=A*x^b$, а на графике отображается линеаризованный вариант уравнения (через натуральный логарифм).

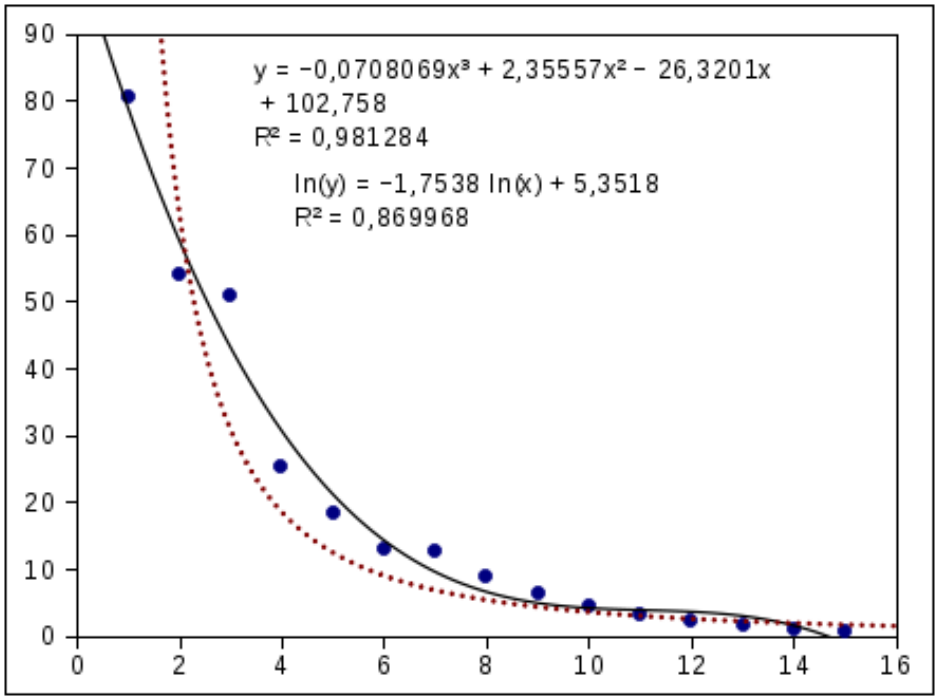


Рисунок 6.9. Исходные данные и два варианта интерполяции

Теперь можно пробовать другие варианты функций и следить за значением критерия $R^2$. Наилучшим описанием будет такое, при котором это значение, как уже упоминалось, будет максимально близко к 1.

Вариант подгонки экспоненциальной зависимостью вида $y(x)= A*e^{bx}$ уже был показан в главе про статистику.

Таким образом, использование Gnumeric для подгонки экспериментальных данных даёт неплохие результаты для не очень сложных зависимостей и позволяет избежать использования громоздких и дорогостоящих математических пакетов программ.