# More Regression

Lecture Notes
CE 311K - McKinney
Introduction to Computer Methods
Department of Civil Engineering
The University of Texas at Austin

## Polynomial Regression

Previously, we fit a straight line to noisy data

$$(x_1,y_1), (x_2,y_2), \cdots (x_n,y_n)$$

using the least-squares criterion. As we have seen, some data are poorly represented by a straight line and for these cases a curve is better suited to fit the data. The most commonly used function for this purpose is the polynomial such as a parabola

$$y = a_0 + a_1 x + a_2 x^2$$

os a cubic

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

or in general an *m*-th degree polynomial:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_m x^m$$

where $a_0,a_1,a_2,\cdots,a_m$ are the constant coefficients of the polynomial.

If the relationship between *x* and *y* were indeed truly *m*-th degree polynomial and there was no noise in the data, then the coefficients could be estimated such that the polynomial passed through all of the data points. However, this is hardly ever the case. As in the linear case, the discrepancy (residual) between the true value of *y* and the polynomial approximation is

$$e_i = y_i - (a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + \cdots + a_m x_i^m)$$

and if there are $n$ such pairs of points ($x_i$, $y_i$), then the sum of squared residuals over all the data points is

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[ y - (a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_m x^m) \right]^2$$

In order to determine the values of the coefficients $a_0, a_1, a_2, \cdots, a_m$, we can minimize $S_r$. The minimization is accomplished by setting the partial derivatives of $S_r$ with respect to each coefficient equal to zero:

$$\begin{aligned} \frac{\partial S_r}{\partial a_0} &= \frac{\partial}{\partial a_0} \left[ \sum_{i=1}^{n} \left( y - a_0 - a_1 x - a_2 x^2 - a_3 x^3 - \cdots - a_m x^m \right)^2 \right] \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial a_0} [\cdots]^2 = \sum_{i=1}^{n} 2[\cdots] \left\{ \frac{\partial}{\partial a_0} [\cdots] \right\} \\ &= \sum_{i=1}^{n} 2 \left[ y - a_0 - a_1 x - a_2 x^2 - a_3 x^3 - \cdots - a_m x^m \right] (-1) \\ &= 0 \end{aligned}$$

Now, dividing by -2 and summing term by term, we have

$$n a_0 + \left[ \sum_{i=1}^{n} x_i \right] a_1 + \left[ \sum_{i=1}^{n} x_i^2 \right] a_2 + \cdots \left[ \sum_{i=1}^{n} x_i^m \right] a_m = \sum_{i=1}^{n} y_i$$

Similarly, the second equation is

$$\begin{aligned} \frac{\partial S_r}{\partial a_1} &= \frac{\partial}{\partial a_1} \left[ \sum_{i=1}^{n} \left( y - a_0 - a_1 x - a_2 x^2 - a_3 x^3 - \cdots - a_m x^m \right)^2 \right] \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial a_1} [\cdots]^2 = \sum_{i=1}^{n} 2[\cdots] \left\{ \frac{\partial}{\partial a_1} [\cdots] \right\} \\ &= \sum_{i=1}^{n} 2 \left[ y - a_0 - a_1 x - a_2 x^2 - a_3 x^3 - \cdots - a_m x^m \right] (-x_i) \\ &= 0 \end{aligned}$$

Dividing by -2 and summing term by term, we have

$$\left[\sum_{i=1}^{n} x_i\right] a_0 + \left[\sum_{i=1}^{n} x_i^2\right] a_1 + \left[\sum_{i=1}^{n} x_i^3\right] a_2 + \cdots + \left[\sum_{i=1}^{n} x_i^{m+1}\right] a_m = \sum_{i=1}^{n} x_i y_i$$

It can now be inferred that the complete set of simultaneous linear equations in the coefficients (the normal equations) of the polynomial is

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^m \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \cdots & \sum_{i=1}^{n} x_i^{m+1} \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \cdots & \sum_{i=1}^{n} x_i^{m+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_i^m & \sum_{i=1}^{n} x_i^{m+1} & \sum_{i=1}^{n} x_i^{m+2} & \cdots & \sum_{i=1}^{n} x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^m y_i \end{bmatrix}$$

**Example:** Given the following data, choose the most suitable low order polynomial and fit it to this data using the least-squares criterion.

**Table.** Data for polynomial fitting example.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *x* | 0 | 1.0 | 1.5 | 2.3 | 2.5 | 4.0 | 5.1 | 6.0 | 6.5 | 7.0 | 8.1 | 9.0 |
| *y* | 0.2 | 0.8 | 2.5 | 2.5 | 3.5 | 4.3 | 3.0 | 5.0 | 3.5 | 2.4 | 1.3 | 2.0 |
| *x* | 9.3 | 11.0 | 11.3 | 12.1 | 13.1 | 14.0 | 15.5 | 16.0 | 17.5 | 17.8 | 19.0 | 20.0 |
| *y* | -0.3 | -1.3 | -3.0 | -4.0 | -4.9 | -4.0 | -5.2 | -3.0 | -3.5 | -1.6 | -1.4 | -0.1 |

The data are plotted in the following Figure. The data appear to have a maximum near $x = 5$ and a minimum near $x = 15$. The lowest order polynomial which can reproduce such behavior is a cubic. The least-squares equations (normal equations) for this set of data ($n = 24$) and for $m = 3$ are

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 & \sum_{i=1}^{n} x_i^6 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \\ \sum_{i=1}^{n} x_i^3 y_i \end{bmatrix}$$

$$\begin{bmatrix} 24 & 229.6 & 3060.2 & 46342.8 \\ 229.6 & 3060.2 & 46342.8 & 752835.2 \\ 3060.2 & 46342.8 & 752835.2 & 12780147.7 \\ 46342.8 & 752835.2 & 12780147.7 & 223518116.8 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} -1.30 \\ -316.9 \\ -6037.2 \\ -9943.36 \end{bmatrix}$$

Gauss elimination yields

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} -0.3593 \\ 2.3051 \\ -0.3532 \\ 0.0121 \end{bmatrix}$$

Thus the equations for the interpolating polynomial is

$$y = 0.0121x^3 - 0.3532x^2 + 2.3051x - 0.3593$$

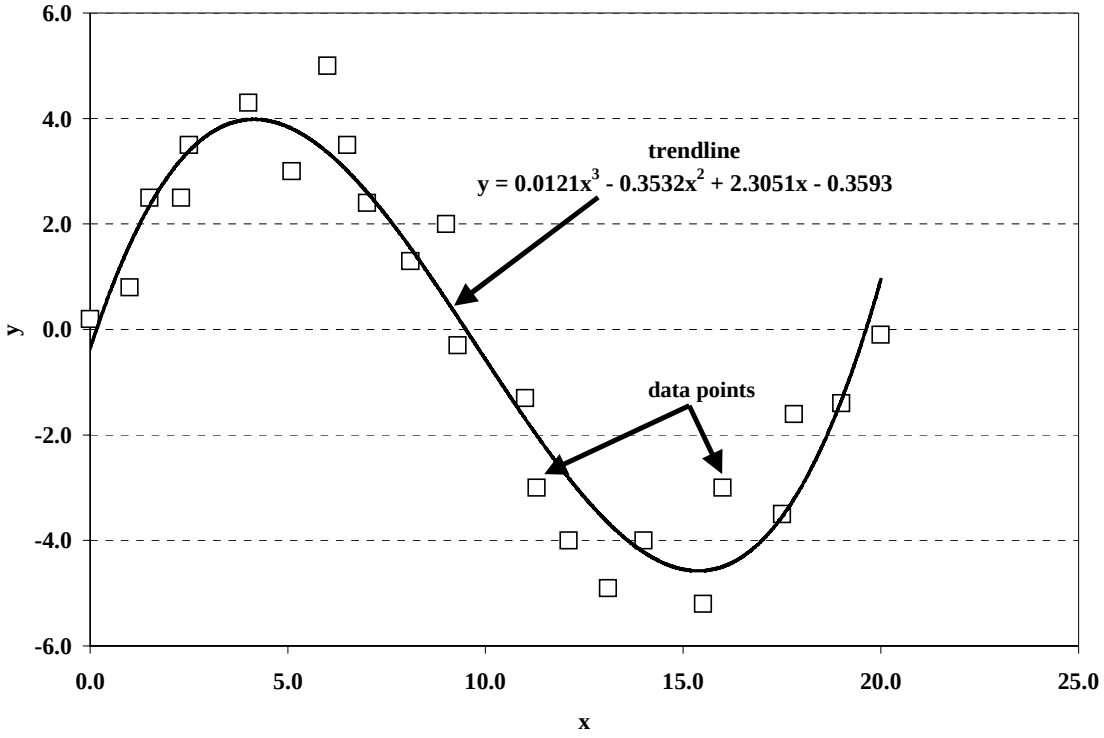



**Figure.** Plot of data for polynomial fitting example

## Linearization of Nonlinear Relationships

In order to apply the techniques of linear least-squares regression, the function whose coefficients are being approximated must be linear in the coefficients. Many relationships among independent and dependent variables in engineering are not linear. However, in many cases a transformation can be applied to the relationships to render them linear in the coefficients. Consider an exponential relationship,

$$y = ae^{bx}$$

where the base is the number $e$, and $a$ and $b$ are constants. If we take the natural logarithm of both sides of the equation, we have

$$\ln(y) = \ln(a) + bx$$

which is a linear relationship between $\ln(y)$ and $x$. The coefficients to be determined in this expression are $\ln(a)$ and $b$. A power law relationship be written as

$$y = ax^b$$

If we take the natural logarithm of both sides of this equation, we have

$$\ln(y) = \ln(a) + b\ln(x)$$

or

$$\log_{10}(y) = \log_{10}(a) + b\log_{10}(x)$$

which is a linear relationship between $\ln(y)$ and $\ln(x)$. Again, the coefficients to be determined in this expression are $\ln(a)$ and $b$.

**Example.** Given the data in the following table, use the least-squares criterion to fit a function of the form $Ax^B$ to these data.:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $x$ | 1.2 | 2.8 | 4.3 | 5.4 | 6.8 | 7.9 |
| $y$ | 2.1 | 11.5 | 28.1 | 41.9 | 72.3 | 91.4 |

The power law relationship is

$$y = Ax^B$$

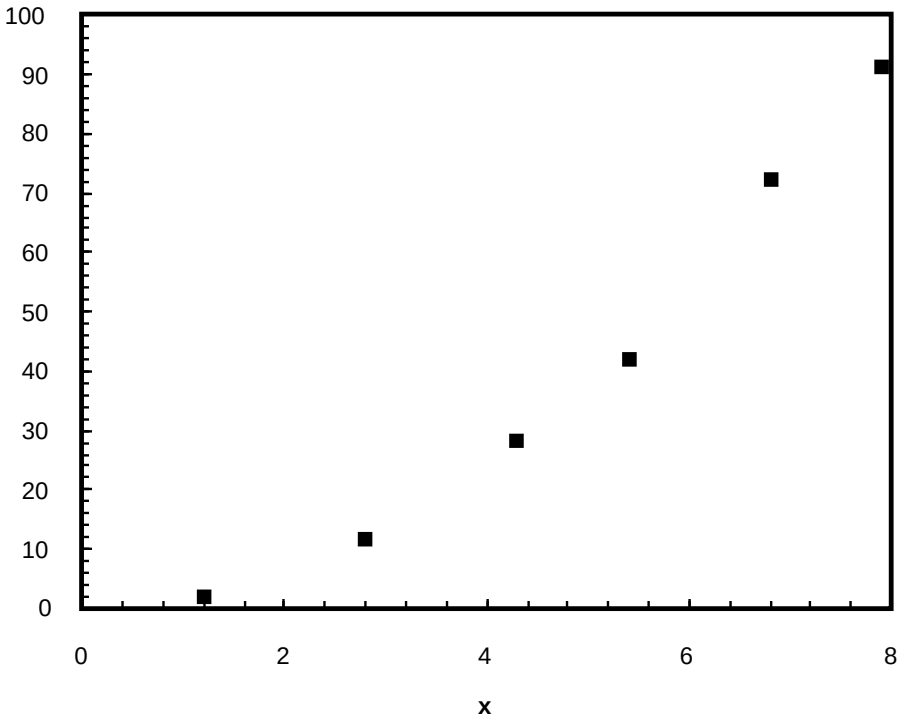
Take the natural logarithm of both sides

$$\ln(y) = \ln(A) + B\ln(x)$$

which is a linear relationship between $\ln(y)$ and $\ln(x)$. The coefficients to be determined are $\ln(A)$ and $B$. Another way to look at this is

$$Y = a + BX$$

which is a linear relationship between $Y = \ln(y)$ and $X = \ln(x)$. The coefficients to be determined are $a = \ln(A)$ and $B$.



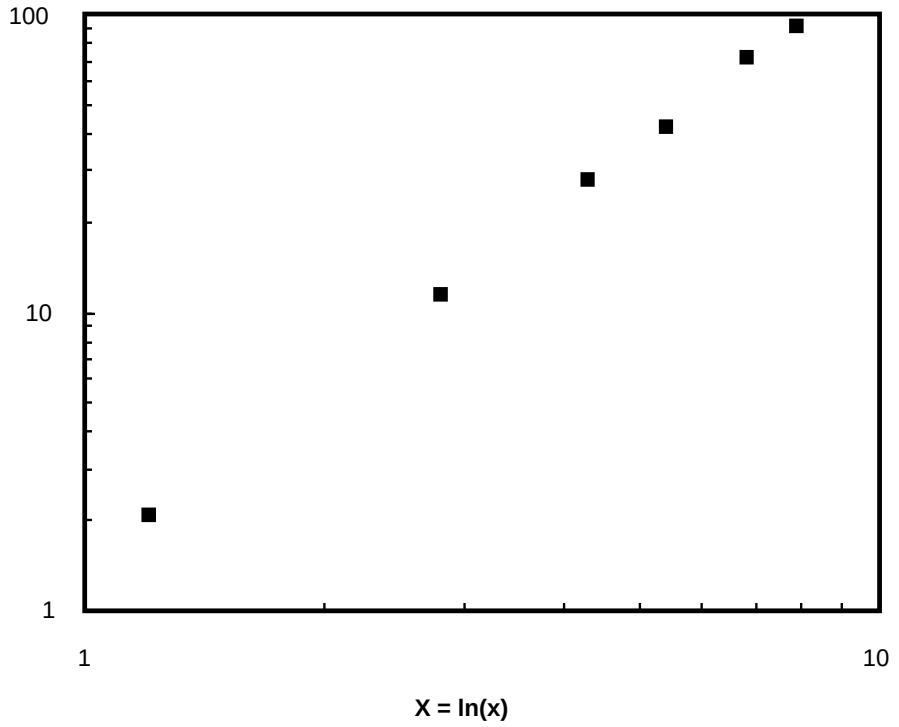




**Figure.** Plot of data on arithmetic and log-log axes.

The normal equations are

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix} \begin{bmatrix} a \\ B \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

| $x_i$ | $X_i = \ln(x_i)$ | $X_i^2$ | $y_i$ | $Y_i = \ln(y_i)$ | $X_i Y_i$ |
|---|---|---|---|---|---|
| 1.2 | 0.18 | 0.03 | 2.1 | 0.74 | 0.14 |
| 2.8 | 1.03 | 1.06 | 11.5 | 2.44 | 2.51 |
| 4.3 | 1.46 | 2.13 | 28.1 | 3.34 | 4.87 |
| 5.4 | 1.69 | 2.84 | 41.9 | 3.74 | 6.3 |
| 6.8 | 1.92 | 3.67 | 72.3 | 4.28 | 8.21 |
| 7.9 | 2.07 | 4.27 | 91.4 | 4.52 | 9.33 |

$$\sum_{i=1}^{5} X_i = \sum_{i=1}^{5} \ln(x_i) = 8.34$$
$$\sum_{i=1}^{5} X_i^2 = \sum_{i=1}^{5} \ln(x_i)^2 = 14.0$$
$$\sum_{i=1}^{5} Y_i = \sum_{i=1}^{5} \ln(y_i) = 19.1$$
$$\sum_{i=1}^{5} X_i Y_i = \sum_{i=1}^{5} \ln(x_i)\ln(y_i) = 31.4$$

Plugging in the numerical values from the data table, the normal equations are

$$\begin{bmatrix} 6 & 8.34 \\ 8.34 & 14.0 \end{bmatrix} \begin{bmatrix} a \\ B \end{bmatrix} = \begin{bmatrix} 19.1 \\ 31.4 \end{bmatrix}$$

Solution yields

$a = ?$ *so* $A = \exp(a) = ?$

$B = ?$

# Example - Carbon Adsorption

Adsorption involves the accumulation of dissolved substances at interfaces of and between material phases. Adsorption may occur as the result of the attraction of a surface or interface for a chemical species, such as the adsorption of substances from water by activated carbon (Weber and DiGiano, 1996) as commonly used in home water filters. Carbon is well known for its adsorptive properties. Activated carbon is regularly used to remove taste and odors from drinking water since carbon has a unique ability to remove synthetic organic chemicals from water supplies.

Adsorption is the process where molecules of a liquid or gas are attached to and then held at the surface of a solid. Physical adsorption is the process whereby surface tension causes molecules to be held at the surface of a solid. Chemical adsorption occurs when a chemical reaction occurs to cause molecules to be held at the surface by chemical bonding. Physical adsorption occurs on activated carbon. The large surface area of the carbon makes it an excellent adsorbent material. Macropores in the surface of the activated carbon granules provide an entrance into the interior of the granual. Adsorption requires three processes: (1) diffusion through a liquid phase to reach the carbon granule, (2) diffusion of molecules through macropores in the carbon granule to an adsorption site, and (3) adsorption of the molecule to the surface. These processes occur at different rates for different molecules of different substances.

Sorption studies are conducted by equilibrating known quantities of sorbent (say, carbon) with solutions of solute (the pollutant). Plots of the resulting data relating the variation of solid-phase concentration, or amount of the solute (pollutant) sorbed per unit mass of solid (carbon), to the variation of the solution-phase concentration are termed sorption isotherms (Weber and DiGianno, 1996). They are referred to as isotherms because the data are collected at constant temperature.

To evaluate the effectiveness of using activated carbon to remove pollutants from water, the first step is to perform a liquid-phase adsorption isotherm test. Data are generated by adding known weights of carbon to water containing a known concentration of pollutant. The carbon-water mixture is mixed at constant temperature, then the carbon is removed by filtration. The residual pollutant concentration in the water is measured and the amount of pollutant adsorbed on to the carbon is calculated. This value if divided by the weight of carbon to determine the carbon loading (q).

Several models have been developed to represent sorption isotherms mathematically. These include:

(1) Linear isotherm model

$$q = Kc$$

where $q$ is the mass of pollutant sorbed per unit mass of carbon at equilibrium with a solution of pollutant concentration $c$, and $K$ is called the distribution coefficient. The distribution coefficient can be determined by fitting a straight line through the origin to the data.

(2) Langmuir isotherm model

$$q = \frac{Qbc}{1 + bc}$$

where $Q$ is the maximum adsorption capacity, and $b$ is a rate constant, and

(3) Freundlich isotherm model

$$q = K(c)^n$$

where, $K$ is called the specific capacity, an indicator of sorption capacity at a specific pollutant concentration; $n$ is a measure of the energy of the sorption reaction. Both of the parameters can be determined fitting a straight line to the logarithmic transformation

$$\ln q = \ln K + (n) \ln c$$

or

$$\log_{10} q = \log_{10} K + n \log_{10} c$$

**Table.** Adsorption data for a pollutant (phenol).

| c | q | logc | logq |
|---|---|---|---|
| 2.8 | 77.8 | 0.45 | 1.89 |
| 3.1 | 90.9 | 0.49 | 1.96 |
| 12.1 | 132.7 | 1.08 | 2.12 |
| 18 | 153.6 | 1.26 | 2.19 |
| 30.4 | 171.4 | 1.48 | 2.23 |
| 36.2 | 185.4 | 1.56 | 2.27 |
| 48.5 | 196.2 | 1.69 | 2.29 |
| 46.4 | 187.2 | 1.67 | 2.27 |
| 63 | 193.4 | 1.80 | 2.29 |
| 71.4 | 232.6 | 1.85 | 2.37 |
| 78.1 | 204.4 | 1.89 | 2.31 |
| 87.7 | 206.2 | 1.94 | 2.31 |
| 102 | 210.8 | 2.01 | 2.32 |
| 109 | 218 | 2.04 | 2.34 |
| 102 | 230.5 | 2.01 | 2.36 |
| 180 | 259.2 | 2.26 | 2.41 |
| 273 | 271.4 | 2.44 | 2.43 |
| 353 | 285.2 | 2.55 | 2.46 |
| 434 | 294.3 | 2.64 | 2.47 |
| 526 | 279.9 | 2.72 | 2.45 |
| 600 | 268 | 2.78 | 2.43 |

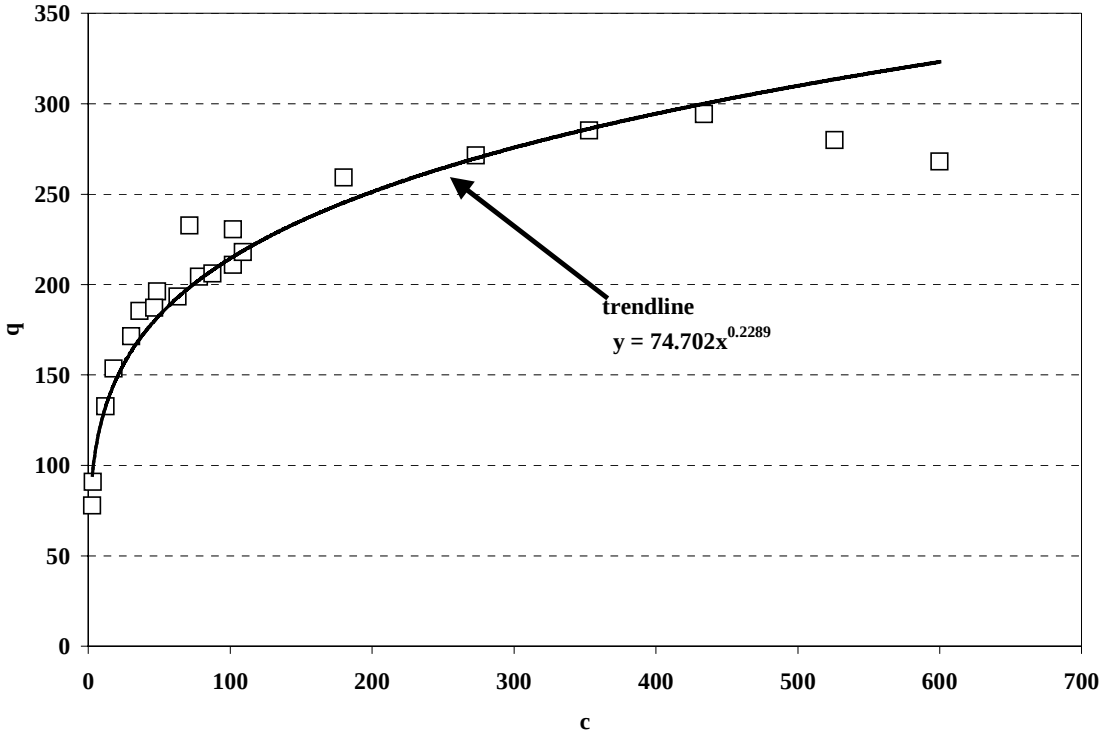



**Figure.** Phenol isotherm (arithmetic scales on axes).

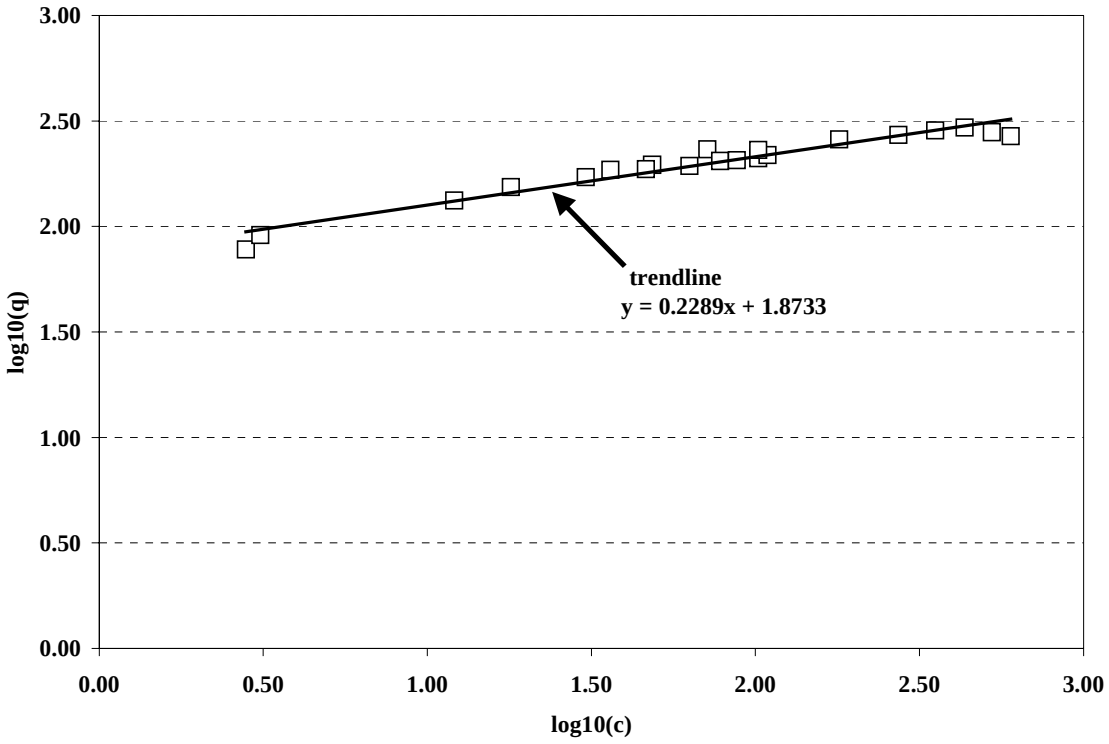



**Figure.** Phenol isotherm (logarithmic scales on axes).

Arithmetic axes:

$$q = K(c)^n$$

so

$$K = 74.702, \text{ and } n = 0.2289$$

Logarithmic axes:

$$\log_{10} q = \log_{10} K + n \log_{10} c$$

so

$$logK = 1.8733$$

or

$$K = 10^{1.6733} = 74.696$$

and

$$n = 0.2289$$