Tufts UNIVERSITY

Bruce M. Boghosian

Motivation for regression

The method of least squares

Higher-degree polynomial fits

Nonlinear models

- Exponential regression
- Logarithmic regression
- Logistic regression
- Other nonlinear models

Summary

Regression

The Method of Least Squares

Bruce M. Boghosian

Tufts UNIVERSITY | School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

1 Motivation for regression

2 The method of least squares

3 Higher-degree polynomial fits

4 Nonlinear models
- Exponential regression
- Logarithmic regression
- Logistic regression
- Other nonlinear models

5 Summary

Regression

- Find *relationship* between two quantities.
- One or both quantities is a random variable.
- The relationship might be that the second is a degree-m polynomial function of the first.
- It might be that no such exact relation exists.
- There might be a "closest" degree-m polynomial for the given data.
- Finding that polynomial is an example of *regression*.

Approach to the problem

Bruce M. Boghosian

Motivation for regression

The method of least squares

Higher-degree polynomial fits

Nonlinear models

Exponential regression

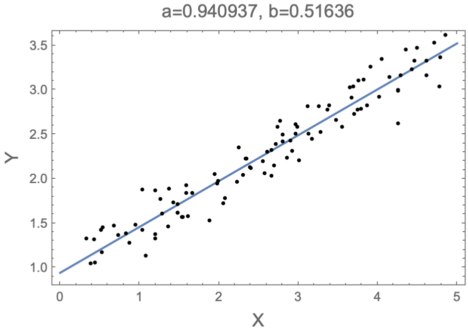
Logarithmic regression

Logistic regression

Other nonlinear models

Summary

- Suppose we have n data pairs (x_i, y_i), for $i = 1, \ldots, n$.
- Find the "best" linear fit, $y = a + bx$.
- This would be called *linear regression*.
- Result would be the optimum parameters, a and b.



The method of least squares

- Suppose we have n data pairs (x_i, y_i), for $i = 1, \dots, n$.
- We wish to fit this to a polynomial $p(x) = a + \sum_{j=1}^{m} b_j x^j$
- Find the coefficients that minimize

$$L = \sum_{i=1}^{n} [y_i - p(x_i)]^2$$

- Result will be optimal a, and b_j where $j = 1, \dots, m$.
- In the above, note that
  - n is the number of data points.
  - m is the order of the polynomial.

Linear least-squares regression

Bruce M. Boghosian

Motivation for regression
The method of least squares
Higher-degree polynomial fits
Nonlinear models
Exponential regression
Logarithmic regression
Logistic regression
Other nonlinear models
Summary

- Let the polynomial be linear, $p(x) = a + bx$
- Find the coefficients that minimize

$$L(a, b) = \sum_{i=1}^{n} [y_i - p(x_i)]^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

- Find minimum L by setting partial derivatives to zero

$$0 = \frac{\partial L}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + bx_i)]$$

$$0 = \frac{\partial L}{\partial b} = -2 \sum_{i=1}^{n} [y_i - (a + bx_i)]\, x_i$$

Linear least-squares regression (continued)

- Find minimum L by setting partial derivatives to zero

$$0 = \frac{\partial L}{\partial a} = -2 \sum_{i=1}^{n} \left[y_i - (a + b x_i) \right]$$

$$0 = \frac{\partial L}{\partial b} = -2 \sum_{i=1}^{n} x_i \left[y_i - (a + b x_i) \right]$$

- Rearrange to obtain

$$\left(\sum_{i=1}^{n} 1 \right) a + \left(\sum_{i=1}^{n} x_i \right) b = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i \right) a + \left(\sum_{i=1}^{n} x_i^2 \right) b = \sum_{i=1}^{n} x_i y_i$$

- Two simultaneous equations in two unknowns for a and b

Linear least-squares regression (continued)

- Two simultaneous equations

$$\left(\sum_{i=1}^{n} 1\right) a + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} y_i$$
$$\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b = \sum_{i=1}^{n} x_i y_i$$

- Define the moments $M_j := \sum_{i=1}^{n} x_i^j$ and $N_j := \sum_{i=1}^{n} x_i^j y_i$
- Simultaneous equations are

$$M_0 a + M_1 b = N_0$$
$$M_1 a + M_2 b = N_1$$

- Note that $M_0 = n$.

Linear least-squares regression (continued)

- Simultaneous equations for a and b

$$\begin{aligned} M_0 a + M_1 b &= N_0 \\ M_1 a + M_2 b &= N_1 \end{aligned}$$

- Solutions can be written (using, e.g., Cramer's rule)

$$a = \frac{M_2 N_0 - M_1 N_1}{M_0 M_2 - M_1^2} \quad \text{and} \quad b = \frac{M_0 N_1 - M_1 N_0}{M_0 M_2 - M_1^2}$$

- Alternatively, solve above for b and then use

$$a = \frac{N_0 - M_1 b}{M_0}.$$

Linear least-squares regression (continued)

- Writing the answers in terms of the data

$$a = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

$$b = \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2}$$

- Alternatively, solve above for b and then use

$$a = \frac{1}{n}\left(\sum_{i=1}^n y_i\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right) b.$$

which can also be written $a = \overline{y} - \overline{x}\ b$.

Linear least-squares regression (continued)

Bruce M. Boghosian

Motivation for regression
The method of least squares
Higher-degree polynomial fits
Nonlinear models
Exponential regression
Logarithmic regression
Logistic regression
Other nonlinear models
Summary

- The preceding is the content of Theorem 11.2.1 in Larsen & Marx.
- Another way to understand these results is to note

$$M_0 = n, \quad M_1 = n\overline{x}, \quad M_2 = n\overline{x^2}, \quad N_0 = n\overline{y} \quad N_1 = n\overline{xy}$$

- Then the solutions

$$a = \frac{M_2 N_0 - M_1 N_1}{M_0 M_2 - M_1^2} \quad \text{and} \quad b = \frac{M_0 N_1 - M_1 N_0}{M_0 M_2 - M_1^2}$$

become

$$a = \frac{\overline{x^2}\ \overline{y} - \overline{x}\ \overline{xy}}{\overline{x^2} - (\overline{x})^2} \quad \text{and} \quad b = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - (\overline{x})^2}$$

Polynomial least-squares regression

- For linear regression, we had the simultaneous equations

$$M_0 a + M_1 b = N_0$$
$$M_1 a + M_2 b = N_1$$

- For higher-degree polynomials, the pattern continues

$$\begin{array}{llllllllll}
M_0 & a & +M_1 & b_1 & +M_2 & b_2 & +\cdots & +M_m & b_m = & N_0 \\
M_1 & a & +M_2 & b_1 & +M_3 & b_2 & +\cdots & +M_{m+1} & b_m = & N_1 \\
M_2 & a & +M_3 & b_1 & +M_4 & b_2 & +\cdots & +M_{m+2} & b_m = & N_2 \\
\vdots & & \vdots & & \vdots & & & \vdots & \vdots & \vdots \\
M_m & a & +M_{m+1} & b_1 & +M_{m+2} & b_2 & +\cdots & +M_{m+m} & b_m = & N_m
\end{array}$$

- Must solve $m+1$ equations for the $m+1$ unknowns

$$a,\ b_1,\ b_2, \ldots,\ b_m.$$

- Result is polynomial $p(x) = a + b_1 x + b_2 x^2 + \cdots + b_m x^m$.

Exponential growth

- Malthusian population growth model (Malthus 1798)
- Fractional growth rate is constant

$$\frac{1}{P}\frac{dP}{dt} = b$$

- Results in differential equation

$$\frac{dP}{dt} = bP$$

- General solution to this equation is

$$P(t) = ae^{bt}$$

- We collect data (t_i, P_i) and wish to find "best" a and b.

Exponential regression

- Suppose we have n data pairs (x_i, y_i), for $i = 1, \ldots, n$.
- We have reason to believe that it fits to $y = ae^{bx}$.
- You could try fitting by minimizing

$$L(a, b) = \sum_{i=1}^{n} \left(y_i - ae^{bx_i}\right)^2$$

- Setting partial derivatives to zero leaves you with difficult-to-solve simultaneous nonlinear equations.
- Fortunately, there is an easier way.

Exponential regression (continued)

- We want to fit data to $y = ae^{bx}$.
- Note that this is equivalent to $\ln y = \ln a + bx$.
- So the pairs $(x_i, \ln y_i)$ have a linear relation.
- Do linear least-square fit to obtain coefficients A and B.
- Then identify $\ln a = A$ and $b = B$, or $a = e^A$ and $b = B$.
- Note that all the y_i must be positive for this to work.

Power-law growth

Bruce M. Boghosian

Motivation for regression
The method of least squares
Higher-degree polynomial fits
Nonlinear models
Exponential regression
Logarithmic regression
Logistic regression
Other nonlinear models
Summary

- Fractional growth rate decays in time

$$\frac{1}{P}\frac{dP}{dt} = \frac{b}{t}$$

- Results in differential equation

$$\frac{dP}{dt} = b\frac{P}{t}$$

- General solution to this equation is

$$P(t) = at^b$$

- We collect data (t_i, P_i) and wish to find "best" a and b.

Logarithmic regression

- We want to fit data to $y = ax^b$.
- Note that this is equivalent to $\ln y = \ln a + b \ln x$.
- So the pairs $(\ln x_i, \ln y_i)$ have a linear relation.
- Do linear least-square fit to obtain coefficients A and B.
- Then identify $\ln a = A$ and $b = B$, or $a = e^A$ and $b = B$.
- Note that all x_i and y_i must be positive for this to work.

Logistic growth

- Malthusian growth leads to infinite population as $t \to \infty$.
- Before this happens, resources are exhausted and reproduction decreases to zero at *carrying capacity* $P = L$.
- Verhulst (1838) proposed logistic growth

$$\frac{1}{P}\frac{dP}{dt} = b\left(1 - \frac{P}{L}\right)$$

- Results in differential equation

$$\frac{dP}{dt} = bP\left(1 - \frac{P}{L}\right)$$

- General solution to this equation is

$$P(t) = \frac{L}{1 + e^{a-bt}} \to L \quad \text{as } t \to \infty$$

- We collect data (t_i, P_i) and wish to find “best” a and b.

Logistic regression

- We want to fit data to $y = \frac{L}{1+e^{a+bx}}$, where L is known.
- Note that this is equivalent to $\frac{L}{y} - 1 = e^{a+bx}$, or

$$\ln\left(\frac{L}{y} - 1\right) = a + bx$$

- So the pairs $(x_i, \ln(L/y_i - 1))$ have a linear relation.
- Do linear least-square fit to obtain coefficients a and b.
- Note that all $\frac{L}{y_i} - 1$ must be positive for this to work.
- Since $y = \frac{L}{1+e^{a+bx}} < L$, choose L greater than $\max_i y_i$.
- You can learn to "eyeball" the data to make a guess for L.

Transformations for other nonlinear models

- Other models and corresponding differential equations

Differential equation	Solution	Linear fit
$\frac{dy}{dx} = -by^2$	$y = \frac{1}{a+bx}$	$\frac{1}{y}$ versus x
$\frac{dy}{dx} = +b\frac{y^2}{x^2}$	$y = \frac{x}{b+ax}$	$\frac{1}{y}$ versus $\frac{1}{x}$
$\frac{dy}{dx} = -\frac{b(1-y)\ln(1-y)}{x}$	$y = 1 - e^{-\frac{x^b}{a}}$	$\ln\ln\left(\frac{1}{1-y}\right)$ versus $\ln x$

- In all cases above,
  - parameter b appears in the differential equation.
  - parameter a is the arbitrary constant in the solution.

Summary

Bruce M. Boghosian

Motivation for regression
The method of least squares
Higher-degree polynomial fits
Nonlinear models
Exponential regression
Logarithmic regression
Logistic regression
Other nonlinear models
Summary

- We have introduced the topic of regression of data.
- We introduced the methodology behind least-squares fits.
- We began with linear fits.
- We showed how this could be extended to arbitrary polynomial fits.
- We then showed how transformations could be applied to data so fits to other nonlinear functions could be handled.