Linear Regression

18.05 Spring 2014

Agenda

- Fitting curves to bivariate data
- Measuring the goodness of fit
- The fit vs. complexity tradeoff
- Regression to the mean
- Multiple linear regression

Modeling bivariate data as a function + noise

Ingredients

- Bivariate data (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n).
- Model: $y_i = f(x_i) + E_i$

  where $f(x)$ is some function, E_i random error.
- Total squared error: $\displaystyle\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$

Model allows us to predict the value of y for any given value of x.

- x is called the independent or predictor variable.
- y is the dependent or response variable.

Examples of $f(x)$

- lines: $y = ax + b + E$
- polynomials: $y = ax^2 + bx + c + E$
- other: $y = a/x + b + E$
- other: $y = a\sin(x) + b + E$

Simple linear regression: finding the best fitting line

- Bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$.
- Simple linear regression: fit a line to the data

$$y_i = ax_i + b + E_i, \quad \text{where} \quad E_i \sim \mathrm{N}(0, \sigma^2)$$

and where σ is a fixed value, the same for all data points.

- Total squared error: $\displaystyle\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$
- Goal: Find the values of a and b that give the 'best fitting line'.
- Best fit: (least squares)
  The values of a and b that minimize the total squared error.

Linear Regression: finding the best fitting polynomial

- Bivariate data: $(x_1, y_1), \ldots, (x_n, y_n)$.

- Linear regression: fit a parabola to the data

$$y_i = ax_i^2 + bx_i + c + E_i, \quad \text{where} \quad E_i \sim \mathrm{N}(0, \sigma^2)$$

  and where σ is a fixed value, the same for all data points.

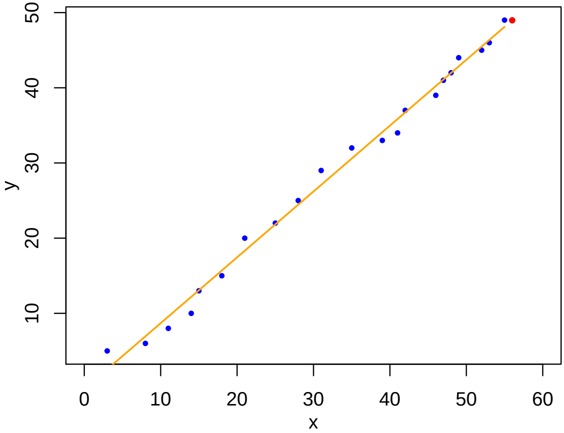
- Total squared error: $\displaystyle\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - ax_i^2 - bx_i - c)^2.$

- Goal:
  Find the values of a, b, c that give the 'best fitting parabola'.

- Best fit: (least squares)
  The values of a, b, c that minimize the total squared error.

  Can also fit higher order polynomials.

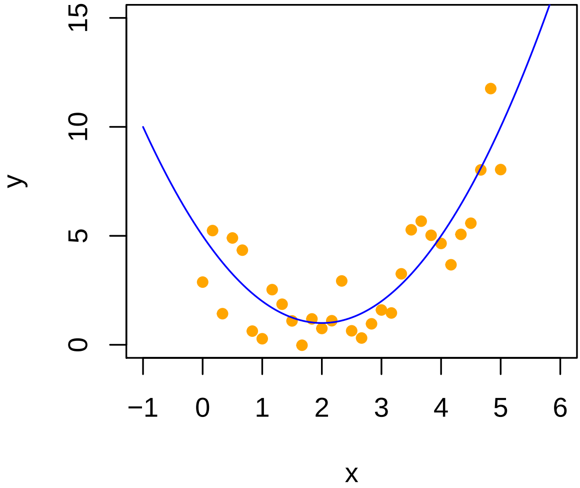
Stamps



Stamp cost (cents) vs. time (years since 1960)
(Red dot = 49 cents is predicted cost in 2016.)

(Actual cost of a stamp dropped from 49 to 47 cents on 4/8/16.)

Parabolic fit



Board question: make it fit

Bivariate data:

$$(1,3),\ (2,1),\ (4,4)$$

1. Do (simple) linear regression to find the best fitting line.
Hint: minimize the total squared error by taking partial derivatives with respect to a and b.

2. Do linear regression to find the best fitting parabola.

3. Set up the linear regression to find the best fitting cubic. but don't take derivatives.

4. Find the best fitting exponential $y = \mathrm{e}^{ax+b}$.
Hint: take $\ln(y)$ and do simple linear regression.

Solutions

1. Model $\hat{y}_i = ax_i + b$.

$$\begin{aligned} \text{total squared error} \;=\; T \;&=\; \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - ax_i - b)^2 \\ &= (3-a-b)^2 + (1-2a-b)^2 + (4-4a-b)^2 \end{aligned}$$

Take the partial derivatives and set to 0:

$$\begin{aligned} \frac{\partial T}{\partial a} &= -2(3-a-b) - 4(1-2a-b) - 8(4-4a-b) &= 0 \\ \frac{\partial T}{\partial b} &= -2(3-a-b) - 2(1-2a-b) - 2(4-4a-b) &= 0 \end{aligned}$$

A little arithmetic gives the system of simultaneous linear equations and solution:

$$\begin{array}{lll} 42a & +14b & = 42 \\ 14a & +6b & = 16 \end{array} \quad \Rightarrow \quad a = 1/2,\; b = 3/2.$$

The least squares best fitting line is $y = \frac{1}{2}x + \frac{3}{2}$.

Solutions continued

2. Model $\hat{y}_i = ax_i^2 + bx_i + c$.

Total squared error:

$$\begin{aligned} T &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - ax_i^2 - bx_i - c)^2 \\ &= (3 - a - b - c)^2 + (1 - 4a - 2b - c)^2 + (4 - 16a - 4b - c)^2 \end{aligned}$$

We didn't really expect people to carry this all the way out by hand. If you did you would have found that taking the partial derivatives and setting to 0 gives the following system of simultaneous linear equations.

$$\begin{array}{llll} 273a & +73b & +21c & = 71 \\ 73a & +21b & +7c & = 21 \\ 21a & +7b & +3c & = 8 \end{array} \quad \Rightarrow \quad a = 1.1667, \; b = -5.5, \; c = 7.3333.$$

The least squares best fitting parabola is $y = 1.1667x^2 + -5.5x + 7.3333$.

Solutions continued

3. Model $\hat{y}_i = ax_i^3 + bx_i^2 + cx_i + d$.

Total squared error:

$$T = \sum (y_i - \hat{y}_i)^2$$
$$= \sum (y_i - ax_i^3 - bx_i^2 - cx_i - d)^2$$
$$= (3 - a - b - c - d)^2 + (1 - 8a - 4b - 2c - d)^2 + (4 - 64a - 16b - 4c$$

In this case with only 3 points, there are actually many cubics that go through all the points exactly. We are probably overfitting our data.

4. Model $\hat{y}_i = \mathrm{e}^{ax_i + b} \Leftrightarrow \ln(y_i) = ax_i + b$.

Total squared error:

$$T = \sum (\ln(y_i) - \ln(\hat{y}_i))^2$$
$$= \sum (\ln(y_i) - ax_i - b)^2$$
$$= (\ln(3) - a - b)^2 + (\ln(1) - 2a - b)^2 + (\ln(4) - 4a - b)^2$$

Now we can find a and b as before. (Using R: $a = 0.18$, $b = 0.41$)

What is linear about linear regression?

Linear in the parameters a, b, ...

$$y = ax + b.$$
$$y = ax^2 + bx + c.$$

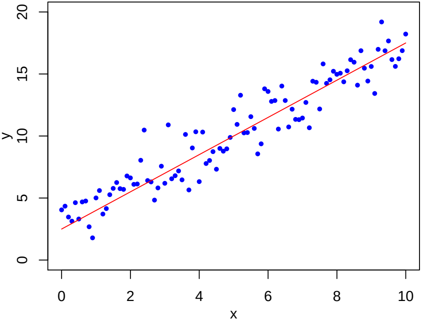
It is **not** because the curve being fit has to be a straight line –although this is the simplest and most common case.

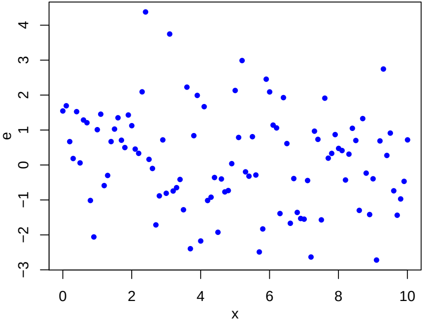
Notice: in the board question you had to solve a system of simultaneous linear equations.

Fitting a line is called simple linear regression.

Homoscedastic

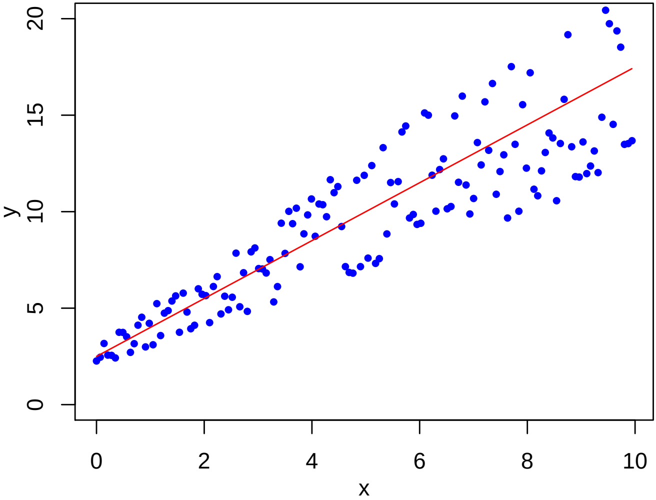
BIG ASSUMPTIONS: the E_i are independent with the same variance σ^2.





Regression line (left) and residuals (right).
Homoscedasticity = uniform spread of errors around regression line.

Heteroscedastic



Heteroscedastic Data

Formulas for simple linear regression

Model:

$$y_i = ax_i + b + E_i \quad \text{where} \quad E_i \sim \mathrm{N}(0, \sigma^2).$$

Using calculus or algebra:

$$\hat{a} = \frac{s_{xy}}{s_{xx}} \quad \text{and} \quad \hat{b} = \bar{y} - \hat{a}\,\bar{x},$$

where

$$\bar{x} = \frac{1}{n}\sum x_i \qquad s_{xx} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$\bar{y} = \frac{1}{n}\sum y_i \qquad s_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}).$$

WARNING: This is just for simple linear regression. For polynomials and other functions you need other formulas.

Board Question: using the formulas plus some theory

Bivariate data: $(1,3)$, $(2,1)$, $(4,4)$

1.(a) Calculate the sample means for x and y.

1.(b) Use the formulas to find a best-fit line in the xy-plane.

$$\hat{a} = \frac{s_{xy}}{s_{xx}} \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}$$

$$s_{xy} = \frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y}) \qquad s_{xx} = \frac{1}{n-1}\sum(x_i - \overline{x})^2.$$

2. Show the point $(\overline{x}, \overline{y})$ is always on the fitted line.

3. Under the assumption $E_i \sim \mathrm{N}(0, \sigma^2)$ show that the least squares method is equivalent to finding the MLE for the parameters (a, b).

Hint: $f(y_i \,|\, x_i, a, b) \sim \mathrm{N}(ax_i + b, \sigma^2)$.

Solution

<u>**answer:**</u> **1.** (a) $\bar{x} = 7/3$, $\bar{y} = 8/3$.
(b)

$$s_{xx} = (1 + 4 + 16)/3 - 49/9 = 14/9, \quad s_{xy} = (3 + 2 + 16)/3 - 56/9 = 7/9.$$

So

$$\hat{a} = \frac{s_{xy}}{s_{xx}} = 7/14 = 1/2, \quad \hat{b} = \bar{y} - \hat{a}\bar{x} = 9/6 = 3/2.$$

(The same answer as the previous board question.)

2. The formula $\hat{b} = \bar{y} - \hat{a}\bar{x}$ is exactly the same as $\bar{y} = \hat{a}\bar{x} + \hat{b}$. That is, the point $(\bar{x}, \bar{y})$ is on the line $y = \hat{a}x + \hat{b}$

Solution to 3 is on the next slide.

3. Our model is $y_i = ax_i + b + E_i$, where the E_i are independent. Since $E_i \sim \mathrm{N}(0, \sigma^2)$ this becomes

$$y_i \sim \mathrm{N}(ax_i + b, \sigma^2)$$

Therefore the likelihood of y_i given x_i, a and b is

$$f(y_i \mid x_i, a, b) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(y_i - ax_i - b)^2}{2\sigma^2}}$$

Since the data y_i are independent the likelihood function is just the product of the expression above, i.e. we have to sum exponents

$$\text{likelihood } = f(y_1, \ldots, y_n \,|x_1, \ldots, x_n, a, b) = \mathrm{e}^{-\frac{\sum_{i=1}^n (y_i - ax_i - b)^2}{2\sigma^2}}$$

Since the exponent is negative, the maximum likelihood will happen when the exponent is as close to 0 as possible. That is, when the sum

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2$$

is as small as possible. This is exactly what we were asked to show.

Measuring the fit

$y = (y_1, \cdots, y_n)$ = data values of the response variable.

$\hat{y} = (\hat{y}_1, \cdots, \hat{y}_n)$ = 'fitted values' of the response variable.

- TSS $= \sum (y_i - \overline{y})^2$ = total sum of squares = total variation.

- RSS $= \sum (y_i - \hat{y}_i)^2$ = residual sum of squares.
  RSS = unexplained by model squared error (due to random fluctuation)

- RSS/TSS = unexplained fraction of the total error.

- $R^2 = 1 - RSS/TSS$ is measure of goodness-of-fit

- R^2 is the fraction of the variance of y explained by the model.

Overfitting a polynomial

- Increasing the degree of the polynomial increases R^2
- Increasing the degree of the polynomial increases the complexity of the model.
- The optimal degree is a tradeoff between goodness of fit and complexity.
- If all data points lie on the fitted curve, then $y = \hat{y}$ and $R^2 = 1$.

R demonstration!

Outliers and other troubles

Question: Can one point change the regression line significantly?

Use mathlet
http://mathlets.org/mathlets/linear-regression/

Regression to the mean

- Suppose a group of children is given an IQ test at age 4. One year later the same children are given another IQ test.
- Children's IQ scores at age 4 and age 5 should be positively correlated.
- Those who did poorly on the first test (e.g., bottom 10%) will tend to show improvement (i.e. regress to the mean) on the second test.
- A completely useless intervention with the poor-performing children might be misinterpreted as causing an increase in their scores.
- Conversely, a reward for the top-performing children might be misinterpreted as causing a decrease in their scores.

This example is from Rice *Mathematical Statistics and Data Analysis*

A brief discussion of multiple linear regression

Multivariate data: $(x_{i,1}, x_{i,2}, \ldots, x_{i,m}, y_i)$ (n data points: $i = 1, \ldots, n$)

Model $\hat{y}_i = a_1 x_{i,1} + a_2 x_{i,2} + \ldots + a_m x_{i,m}$

$x_{i,j}$ are the explanatory (or predictor) variables.

y_i is the response variable.

The total squared error is

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_1 x_{i,1} - a_2 x_{i,2} - \ldots - a_m x_{i,m})^2$$

MIT OpenCourseWare
https://ocw.mit.edu

18.05 Introduction to Probability and Statistics
Spring 2014

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.