Linear Regression
18.05 Spring 2022

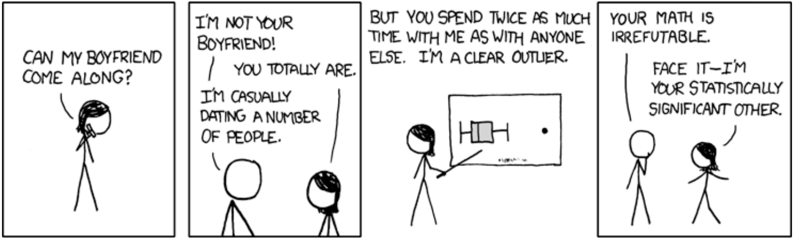

CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

Image courtesy of xkcd. License: CC BY-NC.

https://xkcd.com/539/

Thanks for a great semester!!

Announcements/Agenda

Announcements

- Posted a lot of review questions for final exam.
- Office hours for this week will be announced on Canvas.

Agenda

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

RQuiz

- Overall excellent. Most people did the extra-credit
- Class average was 47/50. Median was 50/50
- **2b. standardization:** $z = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}}$. A number of people forgot the $\sqrt{n}$
- **1c. Bayesian update**:

  `likelihood = dbinom(num_cured, num_patients, theta_values)`

  bayes_numerator = likelihood $*$ prior

  posterior $= \dfrac{\texttt{bayes\_numerator}}{\texttt{sum(bayes\_numerator)}}$.

Modeling bivariate data as a function + noise

Ingredients

- Bivariate data (x_1, y_1), (x_2, y_2), ..., (x_n, y_n).
- Model: $y = f(x) + E$ where $f(x)$ is a function we pick (the model) and E is random error.
- Total squared error: $\displaystyle\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$

The model predicts or explains 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.
- Typically: the error is from an imperfect model or imperfect measurement

Examples of $f(x)$

- lines: $y = ax + b$
- polynomials: $y = ax^2 + bx + c$
- other: $y = a/x + b$
- other: $y = c\mathrm{e}^{ax}$

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

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

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

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

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

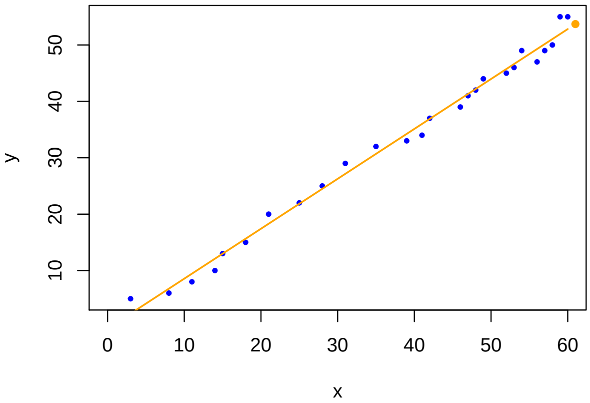
- Total squared error: $\displaystyle\sum_{i=1}^{n} E_i^2 \;=\; \sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c)^2$.

- Goal: Find a, b, c giving 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.

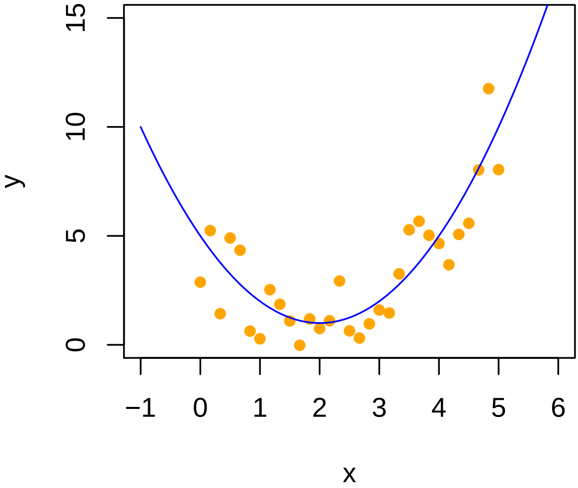
Stamps



Stamp cost (cents) vs. time (years since 1960)
(Orange dot = 53.6 cents is the predicted cost in 2021.)

(Actual cost of a first class stamp in 2021 was 55 cents.)

Parabolic fit



Board question: make it fit

We are given bivariate data: $(1,3),\ (2,1),\ (4,4)$.

(a) Do (simple) linear regression to find the best fitting line.
(i) Give the model for simple linear regression.
(ii) Write down the formula for the total squared error.
(iii) Use calculus to find the parameters that minimize the total squared error.

(b) Do linear regression to find the best fitting parabola. (Really just set this up and get as far as needing to solve equations to find the coefficients.)

(c) Find the best fitting exponential $y = \mathrm{e}^{ax+b}$. (As before, set up the equations but don't solve them.)
Hint: take $\ln(y)$ and do simple linear regression.

(d) For data $(x_1, y_1), \ldots, (x_n, y_n)$. Set up the linear regression to find the best fitting cubic. Don't try to take derivatives or actually find the formulas for the coefficients.

What is linear about linear regression?

Linear in the parameters a, b, ...

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

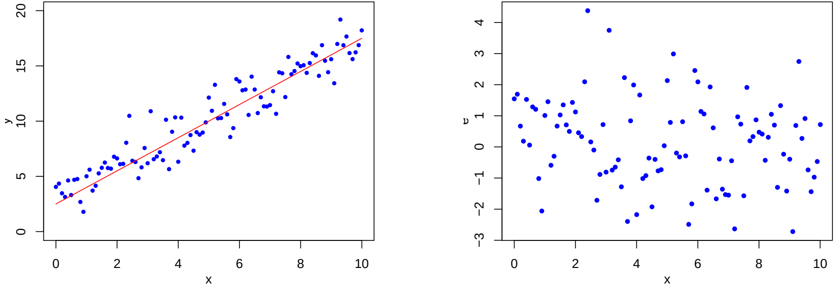
It is **not** because the curve being fit has to be a straight line –although this is the simplest and most common case.

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

Fitting a line is called simple linear regression.

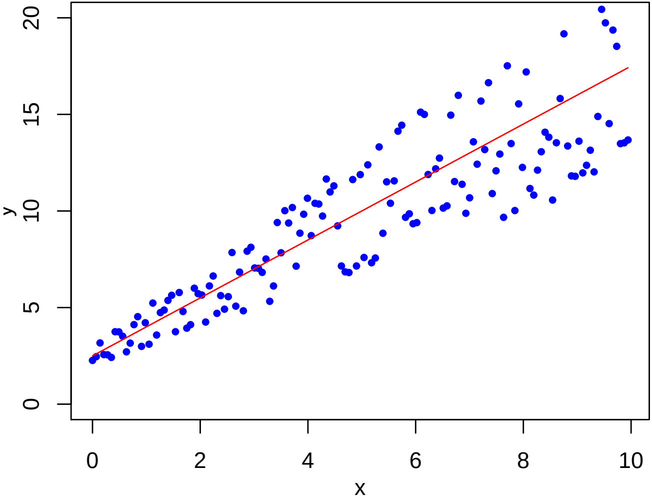
Homoscedastic

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:

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

Using calculus or algebra:

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

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: These formulas are just for simple linear regression. For polynomials and other functions you need other formulas.

Board Question: using the formulas plus some theory

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

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

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

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

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

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

(d) (For fun later!) 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: $f(y_i \,|\, x_i, a, b) \sim \mathrm{N}(ax_i + b, \sigma^2)$.

Measuring the fit

(Not responsible for this on the final exam.)

$y = (y_1, \cdots, y_n) =$ 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 $=$ squared error unexplained by model

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

- $R^2 = (TSS - RSS)/TSS$ is measure of goodness-of-fit (called coefficient of determination)

- 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
https://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: $(x_{i,1}, x_{i,2}, \dots, x_{i,m}, y_i)$ (n data points: $i = 1, \dots, n$)

Model $y_i = a_1 x_{i,1} + a_2 x_{i,2} + \dots + a_m x_{i,m} + E$, where E is random error as before.

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

y_i is the response variable.

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

The total squared error is

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