Chapter 11: Simple linear regression

(Ott & Longnecker Sections: 11.1-11.5)

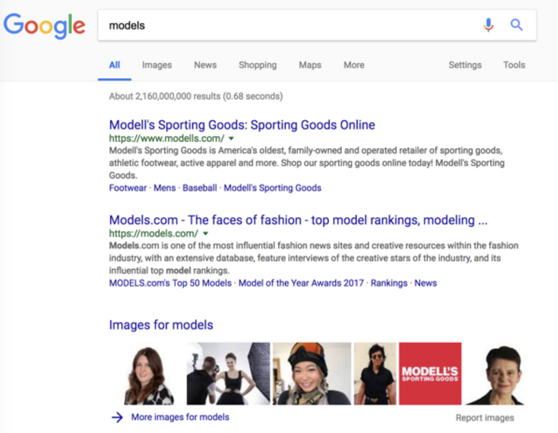
Duzhe Wang

Part 1

https://dzwang91.github.io/stat371/

UNIVERSITY OF WISCONSIN–MADISON

Models


Google
models
All
Images
News
Shopping
Maps
More
Settings
Tools
About 2,160,000,000 results (0.68 seconds)
Modell's Sporting Goods: Sporting Goods Online
https://www.modells.com/
Modell's Sporting Goods is America's oldest, family-owned and operated retailer of sporting goods, athletic footwear, active apparel and more. Shop our sporting goods online today! Modell's Sporting Goods.
Footwear · Mens · Baseball · Modell's Sporting Goods
Models.com - The faces of fashion - top model rankings, modeling ...
https://models.com/
Models.com is one of the most influential fashion news sites and creative resources within the fashion industry, with an extensive database, feature interviews of the creative stars of the industry, and its influential top model rankings.
MODELS.com's Top 50 Models · Model of the Year Awards 2017 · Rankings · News
Images for models
MODELL'S
SPORTING GOODS
More images for models
Report images

Models

Non-math/statistics Models

What is a math/statistics model

- A statistics model describes relationship between different variables.
- Types:
  - Deterministic models (no randomness)
  - Probabilistic models (with randomness)

What is a math/statistics model

- A statistics model describes relationship between different variables.
- Types:
  - Deterministic models (no randomness)
  - Probabilistic models (with randomness)
- An example of deterministic models: Body mass index (BMI) is a measure of body fat based

$$BMI = \frac{WeightinKilograms}{(HeightinMeters)^2}$$

What is a math/statistics model

- A statistics model describes relationship between different variables.
- Types:
  - Deterministic models (no randomness)
  - Probabilistic models (with randomness)
- An example of deterministic models: Body mass index (BMI) is a measure of body fat based

$$BMI = \frac{WeightinKilograms}{(HeightinMeters)^2}$$

- We'll introduce one of the most important and popular probabilistic models: regression model.

Example

Sir Francis Galton (1822-1911) was interested in how children resemble their parents. One simple measure of this is height. So Galton (actually his disciple, Karl Pearson) measured the heights of father son pairs (in inches) at maturity. In the actual study, 1078 pairs were measured. For convenience, we will use a small subsample of $n = 14$ pairs:

Example continued

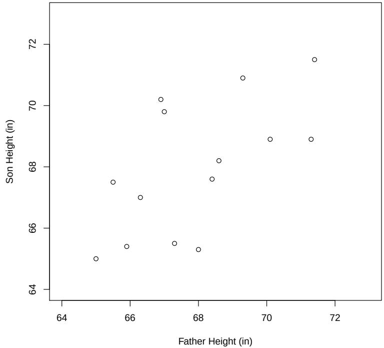
Family	Father's Height	Son's Height
1	71.3	68.9
2	65.5	67.5
3	65.9	65.4
4	68.6	68.2
5	71.4	71.5
6	68.4	67.6
7	65.0	65.0
8	66.3	67.0
9	68.0	65.3
10	67.3	65.5
11	67.0	69.8
12	69.3	70.9
13	70.1	68.9
14	66.9	70.2

Predict sons' heights from father's heights. Deterministic or Probabilistic?

Example continued

- Scatterplot: for each father-son pair, put a point in the two-dimensional plane whose x-coordinate is the father's height and whose y-coordinate is the son's height.
- The response variable is the variable we'd like to predict. By convention, in regression we put the response variable on the vertical axis.
- The predictor variable is the variable we will use to make the prediction.
- The statistical technique of estimating and/or inferring a relationship between two variables is called regression.

Example continued



- It seems that as father's height increases, so does son's height. On a genetic basis, we expect this.
- The nature of the relationship seems approximately linear. However, we also see that sometimes short fathers have tall sons, and vice versa. The relationship is not perfect.

Example continued

- A linear model: an equation that captures that the expected son's heights are linear functions of father's heights.

$$\text{Son's height} = \beta_0 + \beta_1 * \text{Father's height} + \text{Random error}$$

Example continued

- A linear model: an equation that captures that the expected son's heights are linear functions of father's heights.

$$\text{Son's height} = \beta_0 + \beta_1 * \text{ Father's height} + \text{Random error}$$

- β_0 is the intercept.
- β_1 is the slope.
- The Random error term picks up sources of variation in an individual son's height that are not due to his father's height (mother's genetics, environmental factors, etc.) and which cause the points to be "off line."
- Our hope is that the random error term is truly random, so there are no other systemic/structural sources of variation explaining a son's height (if there were, we should try and find them and put them in the model!).

Example continued

- Denote the height of son i by y_i, the height of father i by x_i, and the random error by ϵ_i, so that the model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- Our goal is to estimate the values of β_0 and β_1 from data.