Mathematical statistics

November 27^{th}, 2018

Lecture 22: Linear regression

Att.

- Final exam:

  Friday, 12/14/2018, 10:30am –12:30pm
  Gore Hall 115

- Course evaluation
- Homework due this Thursday

Overview

Week 1	Probability reviews
Week 2	Chapter 6: Statistics and Sampling Distributions
Week 4	Chapter 7: Point Estimation
Week 7	Chapter 8: Confidence Intervals
Week 10	Chapters 9–10: Tests of Hypothesis
Week 14	Chapter 12: Linear regression

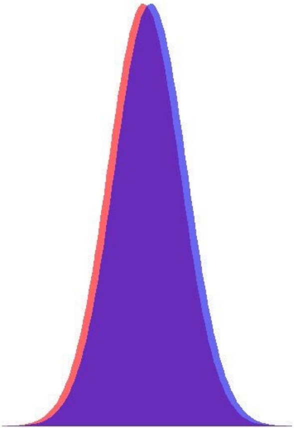
Key steps in statistical inference

- Understand the statistical model [Chapter 6]
- Come up with reasonable estimates of the parameters of interest [Chapter 7]
- Quantify the confidence with the estimates [Chapter 8]
- Testing with the parameters of interest [Chapter 9]

Contexts

- The central mega-example: population mean μ
- Difference between two population means
- Linear regression

Difference between two population means



- Testing:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

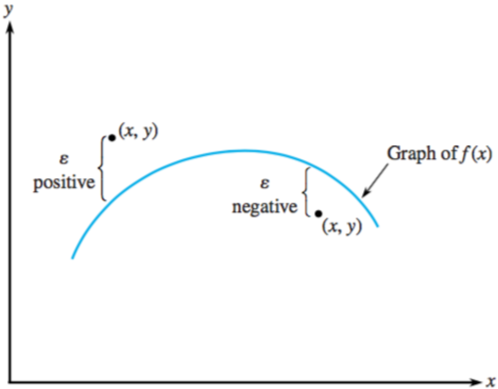
- Works well, even if $|\mu_1 - \mu_2| << \sigma_1, \sigma_2$

Linear regression

Regression analysis

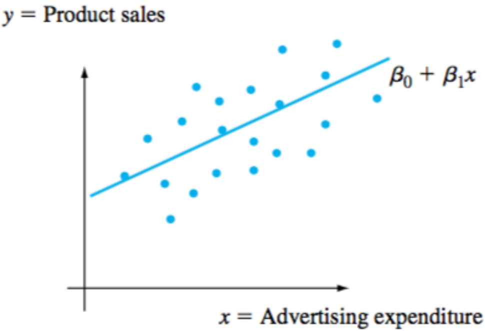
- Objective: determine the relationship between two (or more) variables so that we can gain information about one of them through knowing values of the other(s)
- Many variables in the real world are related, but not in a deterministic fashion

Additive models



$$Y = \text{a deterministic function of } x \ + \ \text{a random deviation}$$
$$= f(x) + \epsilon$$

Linear regression



Mathematical model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

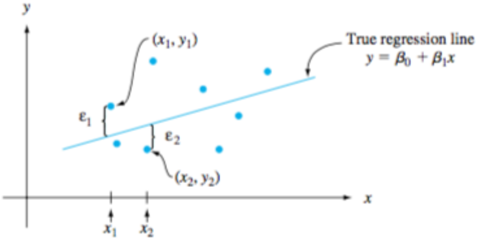
Linear regression

Assumptions:

- There are parameters β_0, β_1 and σ such that for any fixed value of the independent variable x, the dependent variable Y is related to x through the model equation

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

  The random deviation (random variable) ϵ is assumed to be normally distributed with mean value 0 and variance σ^2
- The observed pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are regarded as having been generated independently of one another from the model equation



Mathematical model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Problem 1

Problem

Recall that

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

and

$$\bar{Y} = \frac{Y_1 + Y_2 + \dots Y_n}{n}$$

Compute $E[Y_i]$, $Var[Y_i]$, $E[\bar{Y}]$ *and* $Var[\bar{Y}]$
(in term of $x_1, x_2, \dots, x_n$, β_0, β_1 *and* σ*.)*

Estimating the parameters by maximum likelihood

- Question 1: Given x_i, β_0, β_1 and σ , what is the distribution of Y_i?

$$f_Y(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{[y_i-(\beta_0+\beta_1 x_i)]^2}{2\sigma^2}}$$

- The joint density function is

$$f_{joint}(y_1, y_2, \dots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_i [y_i-(\beta_0+\beta_1 x_i)]^2}$$

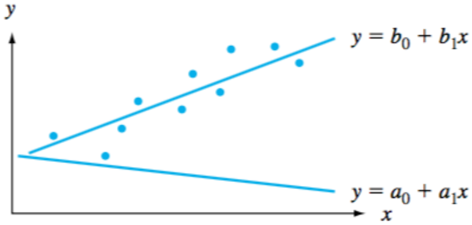
- The maximum likelihood estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

The principle of least squares

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$



The principle of least squares

PRINCIPLE OF LEAST SQUARES

The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1 x$ is

$$\text{height of point} - \text{height of line} = y_i - (b_0 + b_1 x_i)$$

The sum of squared vertical deviations from the points $(x_1, y_1), \ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates**, are those values that minimize $f(b_0, b_1)$. That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are such that $f(\hat{\beta}_0, \hat{\beta}_1) \le f(b_0, b_1)$ for any b_0 and b_1. The **estimated regression line** or **least squares line** is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The principle of least squares

- Taking partial derivatives

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$
$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

- Normal equations

$$\begin{aligned} nb_0 + (\sum x_i)b_1 &= \sum y_i \\ (\sum x_i)b_0 + (\sum x_i^2)b_1 &= \sum x_i y_i \end{aligned}$$

Least squares estimates

- Estimates

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Computing formulas

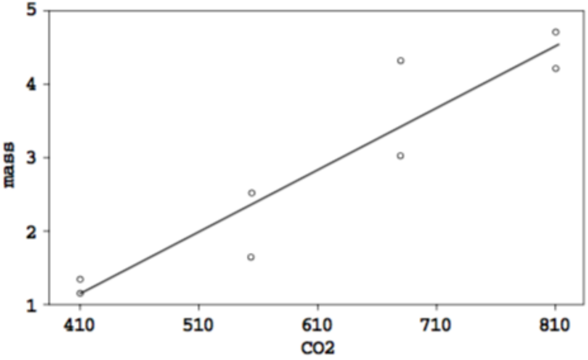
$$S_{xy} = \left(\sum x_i y_i\right) - \frac{(\sum x_i)(\sum y_i)}{n}$$

and

$$S_{xx} = \left(\sum x_i^2\right) - \frac{(\sum x_i)^2}{n}$$

Example

There were two trees at each of four levels of $CO2$ concentration, and the mass of each tree was measured after 11 months. The observations are obtained with $x =$ atmospheric concentration of $CO2$ (mL/L, or ppm) and $y =$ tree mass (kg).



Example

Obs	x	y	x^2	xy	y^2
1	408	1.1	166,464	448.8	1.21
2	408	1.3	166,464	530.4	1.69
3	554	1.6	306,916	886.4	2.56
4	554	2.5	306,916	1385.0	6.25
5	680	3.0	462,400	2040.0	9.00
6	680	4.3	462,400	2924.0	18.49
7	812	4.2	659,344	3410.4	17.64
8	812	4.7	659,344	3816.4	22.09
Sum	4908	22.7	3,190,248	15,441.4	78.93

What are $\hat{\beta}_0, \hat{\beta}_1$?

Problem 2: Estimating σ by maximum likelihood

Problem

The joint density function of $(Y_1, Y_2, \ldots, Y_n)$ *is*

$$f_{joint}(y_1, y_2, \ldots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2}$$

What is the maximum likelihood estimator of σ*?*