MATH 205: Statistical methods

November 29th, 2021

Lecture 22: Linear regression (cont.)

Announcements

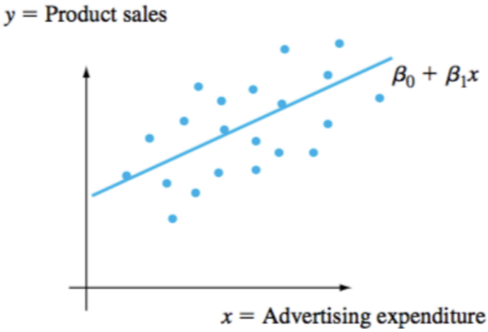
- Homework 5 due on Wednesday, 12/01 (11:59 pm)
- Final exam:

  12/13/2021, Monday
  3:30PM - 5:30PM
  Kirkbride Hall Room 205

- Plan for the next two weeks
  - Other topics of the syllabus: Linear regression and Analysis of Variance
  - Review of materials + Practice exam
- **There will be no lab for the rest of the semester**

Linear regression

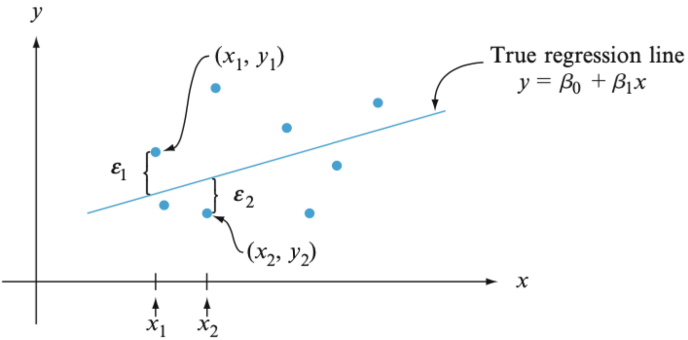
Linear regression



Mathematical model:

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

Linear regression



Why do we do regression?

- Regression to Make Predictions
  $\rightarrow$ You already knew how to do this!

- Regression to spot trends $\rightarrow$ Are you sure that $\beta_1 \neq 0$?

Using correlation to predict

Procedure 2.1 (Predicting a Value Using Correlation) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. Assume we have an x value x_0 for which we want to give the best prediction of a y value, based on this data. The following procedure will produce a prediction:

- Transform the data set into standard coordinates, to get

$$\hat{x}_i = \frac{1}{\mathsf{std}(x)}(x_i - \mathsf{mean}(\{x\}))$$
$$\hat{y}_i = \frac{1}{\mathsf{std}(y)}(y_i - \mathsf{mean}(\{y\}))$$
$$\hat{x}_0 = \frac{1}{\mathsf{std}(x)}(x_0 - \mathsf{mean}(\{x\})).$$

- Compute the correlation

$$r = \mathsf{corr}(\{(x, y)\}) = \mathsf{mean}(\{\hat{x}\hat{y}\}).$$

- Predict $\hat{y}_0 = r\hat{x}_0$.
- Transform this prediction into the original coordinate system, to get

$$y_0 = \mathsf{std}(y) r \hat{x}_0 + \mathsf{mean}(\{y\})$$

Linear regression: settings

Assumption

1. *$x_1, x_2, \dots, x_n$ are fixed design points (non-random)*
2. *Linear model:*
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
where $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are random sample from $\mathcal{N}(0, \sigma^2)$
3. *Let assume (for now), that σ is known*

We want to make inferences about the trend, so β_1 is important

Estimate β_1

The true value of β_1 will be estimated by

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - x)^2}$$

We first note that

$$\sum (x_i - \bar{x})\bar{Y} = \bar{Y} \cdot \sum x_i - \bar{x} = 0$$

We can write $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - x)^2}$$

thus $\hat{\beta}_1$ is a linear combination of independent normal random variables

Problem

We have

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - x)^2} = \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - x)^2}$$

where

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

thus $\hat{\beta}_1$ is a linear combination of independent normal random variables Y_i.
Tasks:

- What are $E[Y_i]$ and $Var(Y_i)$ in terms of x_i, β_0 and β_1?
- What are $E[\bar{Y}]$ in terms of $\bar{x}$, β_0 and β_1?
- What are $E[\hat{\beta}_1]$ and $Var[\hat{\beta}_1]$ in terms of β_0, β_1 and x_i's.

Linear regression: σ is known

Problem

We have

$$\frac{\hat{\beta} - \beta_1}{\sigma/\sqrt{S_{xx}}}$$

follows standard normal distribution, where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Use this to construct a 95% confidence interval of β_1.

Linear regression: σ is unknown

Theorem

If we define

$$S^2 = \frac{\sum [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2}{n-2}$$

then the random variable

$$\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

follows the t−distribution with degrees of freedom $(n-2)$.

Confidence intervals

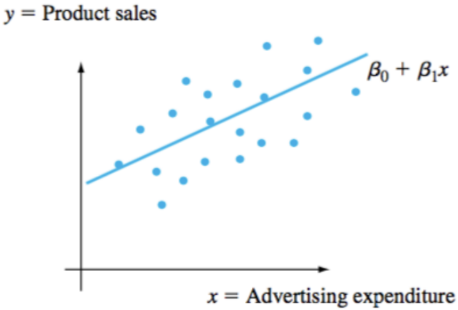
A $100(1-\alpha)\%$ **CI for the slope $\boldsymbol{\beta_1}$** of the true regression line is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$$

Note: When $n > 40$, $t_{\alpha/2,n-2} \approx z_{\alpha/2}$

Testing with β_1

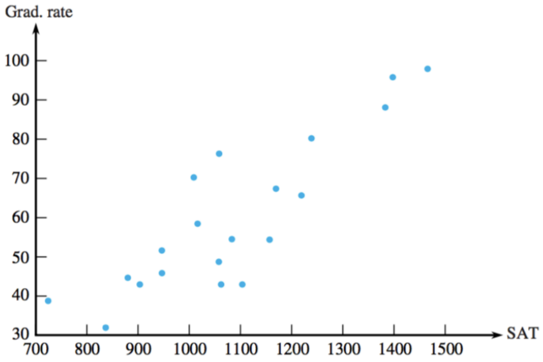
β_1 characterizes relation between x and Y



Question: Does increase advertising expense help increase sales?
$\rightarrow$ Testing $H_0 : \beta_1 = 0$ against $H_a : \beta_1 > 0$

Example

Is it possible to predict graduation rates from SAT scores?



Assume that

$$\hat{\beta}_1 = .08855; s = 10.29; S_{xx} = 704125; n = 20.$$