MATH 205: Statistical methods

December 1st, 2021

Lecture 23: Linear regression (cont.)

Announcements

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

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

- Next week: Practice exam + Review of materials
- **There will be no lab for the rest of the semester**

Linear regression: settings

Assumption

1. *$x_1, x_2, \ldots, x_n$ are fixed design points (non-random)*
2. *Linear model:*

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ 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

Simplest case: σ is known

Estimate β_1

The true value of β_1 will be estimated by

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

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

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

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.

Problem

We have

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

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.

Confidence interval for β_1: σ is known

Recalling that

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

and

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

A $100(1-\alpha)\%$ confidence interval for the slope β_1 of the true regression line is

$$\left(\hat{\beta}_1 - z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}}, \hat{\beta}_1 + z_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}}\right)$$

Confidence interval for β_1: σ is known

A $100(1-\alpha)\%$ confidence upper bound for the slope β_1 of the true regression line is

$$\left(-\infty, \hat{\beta}_1 + z_\alpha \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

Testing about the slope β_1

- Null hypothesis

$$H_0 : \beta_1 = \Delta$$

  where Δ is a constant.

- The alternative hypothesis will be either:
  - $H_a : \beta_1 > \Delta$
  - $H_a : \beta_1 < \Delta$
  - $H_a : \beta_1 \neq \Delta$

Testing about the slope β_1: example

It is well known that the more beer you drink, the more your blood alcohol level rises. Suppose we have the following data on student beer consumption

Student	1	2	3	4	5	6	7	8	9	10
Beers	5	2	9	8	3	7	3	5	3	5
BAL	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06	0.02	0.05

Make a scatterplot and fit the data with a regression line. Test the hypothesis that another beer raises your BAL by 0.02 percent against the alternative that it is less.

$$H_0 : \beta_1 = 0.02$$
$$H_a : \beta_1 < 0.02$$

How do we do testing?

- Let's assume that the null hypothesis is correct
  $\rightarrow$ this means $\beta_1 = \Delta$
- This implies that
  $$\frac{\hat{\beta}_1 - \Delta}{\sigma/\sqrt{S_{xx}}}$$
  follows standard normal distribution.
- Note that this $z - value$ is something we can compute from data
- This means, depending on the alternative hypothesis, we can quantify the p-value associated with this $z - value$
- Comparing this p-value with significance level $\rightarrow$ complete testing procedure

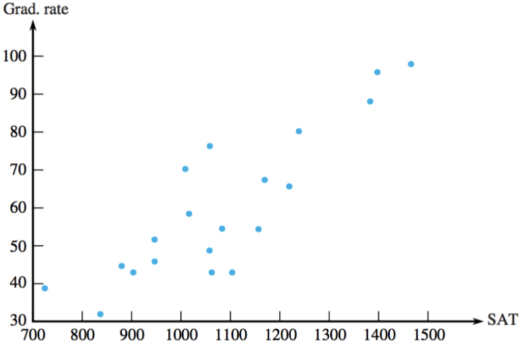
Example

Based on the average SAT score of entering freshmen at a university, can we predict the percentage of those freshmen who will get a degree there within 6 years? A random sample of 20 universities is obtained:

University	Grad rate	SAT
Princeton	98	1465.00
Brown	96	1395.00
Johns Hopkins	88	1380.00
Pittsburgh	65	1215.00
SUNY-Binghamton	80	1235.00
Kansas	58	1011.10
Dayton	76	1055.54
Illinois Inst Tech	67	1166.65
Arkansas	48	1055.54
Florida Inst Tech	54	1155.00
New Mexico Inst Mining	42	1099.99
Temple	54	1080.00
Montana	45	944.43
New Mexico	42	899.99
South Dakota	51	944.43
Virginia Commonwealth	42	1060.00
Widener	70	1005.00
Alabama A&M	38	722.21
Toledo	44	877.77
Wayne State	31	833.32

Example

Is it possible to predict graduation rates from SAT scores?



$\rightarrow$ It seems that a linear model is appropriate.

Example

Problem

Assume that σ is known to be 15, and the computed summary from the dataset is

$$\hat{\beta}_1 = 0.08855; \quad S_{xx} = 704125; \quad n = 20$$

- *Construct a 95% confidence interval of the slope of the true regression line β_1*
- *Conduct a test of hypothesis*

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

General case: 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 $(n-2)$.

Testing about the slope β_1: example

It is well known that the more beer you drink, the more your blood alcohol level rises. Suppose we have the following data on student beer consumption

Student	1	2	3	4	5	6	7	8	9	10
Beers	5	2	9	8	3	7	3	5	3	5
BAL	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06	0.02	0.05

Make a scatterplot and fit the data with a regression line. Test the hypothesis that another beer raises your BAL by 0.02 percent against the alternative that it is less.

$$H_0 : \beta_1 = 0.02$$
$$H_a : \beta_1 < 0.02$$