MATH 205: Statistical methods

November 17th, 2021

Lecture 21: Linear regression

Announcements

- Homework 5 due on Wednesday, 12/01 (11:59 pm)
- Quiz on the lecture Monday after Thanksgiving (Hypothesis testing)

Comparing the mean of two populations

Assumption

1. *$X_1, X_2, \ldots, X_m$ is a random sample from a population with mean μ_1 and variance σ_1^2.*
2. *$Y_1, Y_2, \ldots, Y_n$ is a random sample from a population with mean μ_2 and variance σ_2^2.*
3. *The X and Y samples are independent of each other.*

Properties of $\bar{X} - \bar{Y}$

Proposition

The expected value of $\overline{X} - \overline{Y}$ is $\mu_1 - \mu_2$, so $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. The standard deviation of $\overline{X} - \overline{Y}$ is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Confidence intervals

When both population distributions are normal, standardizing $\overline{X} - \overline{Y}$ gives a random variable Z with a standard normal distribution. Since the area under the z curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$, it follows that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate $\mu_1 - \mu_2$ yields the equivalent probability statement

$$P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right) = 1 - \alpha$$

Testing the difference between two population means

- Setting: independent normal random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ with known values of σ_1 and σ_2. Constant Δ_0.
- Null hypothesis:

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

- Alternative hypothesis:
  (a) $H_a : \mu_1 - \mu_2 > \Delta_0$
  (b) $H_a : \mu_1 - \mu_2 < \Delta_0$
  (c) $H_a : \mu_1 - \mu_2 \neq \Delta_0$
- When $\Delta_0 = 0$, the test (c) becomes

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

Testing the difference between two population means

Assume that we want to test the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$ against each of the following alternative hypothesis

(a) $H_a : \mu_1 - \mu_2 > \Delta_0$

(b) $H_a : \mu_1 - \mu_2 < \Delta_0$

(c) $H_a : \mu_1 - \mu_2 \neq \Delta_0$

We use the test statistic:

$$z = \frac{(\bar{x} - \bar{y}) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}.$$

and derive the p-value in the same way as the one-sample tests.

Large-sample tests/confidence intervals

- Central Limit Theorem: $\bar{X}$ and $\bar{Y}$ are approximately normal when $n > 30 \to$ so is $\bar{X} - \bar{Y}$. Thus

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

  is approximately standard normal

- When n is sufficiently large $S_1 \approx \sigma_1$ and $S_2 \approx \sigma_2$
- Conclusion:

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

  is approximately standard normal when n is sufficiently large

If $m, n > 40$, we can ignore the normal assumption and replace σ by S

Large-sample CIs

Proposition

Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ *with a confidence level of approximately* $100(1-\alpha)\%$ *is*

$$\bar{x} - \bar{y} \ \pm \ z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where $-$*gives the lower limit and* $+$ *the upper limit of the interval. An upper or lower confidence bound can also be calculated by retaining the appropriate sign and replacing* $z_{\alpha/2}$ *by* z_α*.*

Large-sample tests

Proposition

Use of the test statistic value

$$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

along with the previously stated upper-, lower-, and two-tailed rejection regions based on z critical values gives large-sample tests whose significance levels are approximately α. These tests are usually appropriate if both $m > 40$ and $n > 40$. A P-value is computed exactly as it was for our earlier z tests.

Example

Let μ_1 and μ_2 denote true average tread lives for two competing brands of size P205/65R15 radial tires.

(a) Test

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

at level 0.05 using the following data: $m = 45$, $\bar{x} = 42,500$, $s_1 = 2200$, $n = 45$, $\bar{y} = 40,400$, and $s_2 = 1900$.

(b) Construct a 95% CI for $\mu_1 - \mu_2$.

Example

The article "Gender Differences in Individuals with Comorbid Alcohol Dependence and Post-Traumatic Stress Disorder" (Amer. J. Addiction, 2003: 412–423) reported the accompanying data on total score on the Obsessive-Compulsive Drinking Scale (OCSD).

Gender	Sample Size	Sample Mean	Sample SD
Male	44	19.93	7.74
Female	40	16.26	7.58

Formulate hypotheses and carry out an appropriate analysis. Does your conclusion depend on whether a significance level of .05 or .01 was employed?

Example

Research has shown that good hip range of motion and strength in throwing athletes results in improved performance and decreased body stress. The article "Functional Hip Characteristics of Baseball Pitchers and Position Players" (Am. J. Sport. Med., 2010: 383–388) reported on a study involving samples of 40 professional pitchers and 40 professional position players.

For the pitchers, the sample mean trail leg total arc of motion (degrees) was 75.6 with a sample standard deviation of 5.9, whereas the sample mean and sample standard deviation for position players were 79.6 and 7.6, respectively.
Assuming normality, test appropriate hypotheses to decide whether true average range of motion for the pitchers is less than that for the position players (as hypothesized by the investigators).

Example

A letter in the Journal of the American Medical Association (May 19, 1978) reports that of 215 male physicians who were Harvard graduates and died between November 1974 and October 1977, the 125 in full-time practice lived an average of 48.9 years beyond graduation, whereas the 90 with academic affiliations lived an average of 43.2 years beyond graduation.

Does the data suggest that the mean lifetime after graduation for doctors in full-time practice exceeds the mean lifetime for those who have an academic affiliation?

Review: Using correlation to predict

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

$$\hat{x}_i = \frac{1}{\mathsf{std}(x)}(x_i - \mathsf{mean}(\{x\}))$$

$$\hat{y}_i = \frac{1}{\mathsf{std}(y)}(y_i - \mathsf{mean}(\{y\}))$$

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

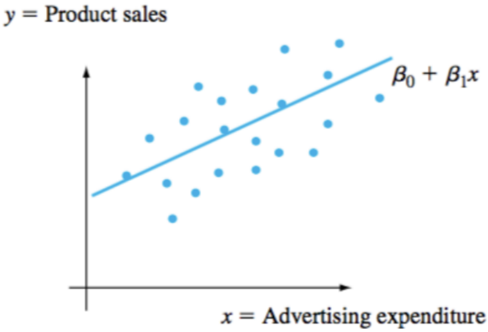
- Compute the correlation

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

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

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

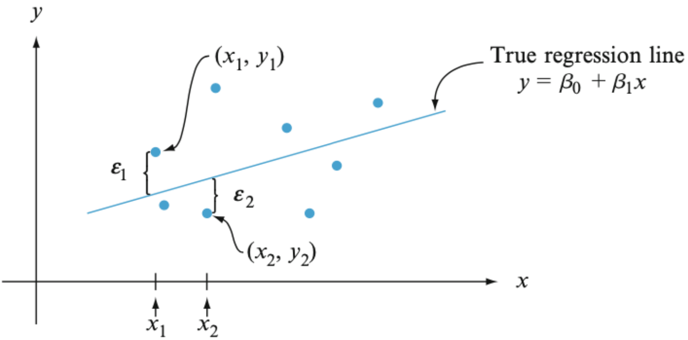
Linear regression



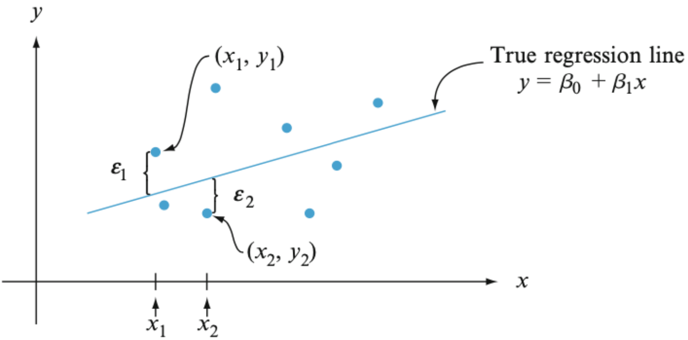
Mathematical model:

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

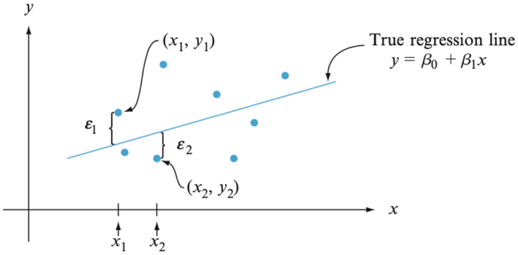
Linear regression



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 $\beta_1 > 0$?

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

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.