Intervals estimation in linear regression

Statistics

Anton Afanasev

Higher School of Economics

DSBA 211
March 4, 2023

Seminar Overview

1. Quiz

2. Simple linear regression
   - Model
   - Ordinary least squares
   - Variance
   - Normal regression

3. Intervals estimation
   - Slope and intercept
   - Confidence interval on interpolated value
   - Prediction interval on observed value

Quiz

Three random samples were taken from normal distributions with same standard deviations.

Sample 1	10	20	
Sample 2	5	15	25
Sample 3	8	32	

1. Complete ANOVA table.
2. Test $H_0 : \mu_1 = \mu_2 = \mu_3$ against $H_1 :$ not H_0 at 5% significance level.
3. Construct $CI(\mu_1)_{95\%}$.
4. Construct $CI(\mu_1 - \mu_2)_{95\%}$.
5. Test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ at 5% significance level.
6. Construct $SCI(\mu_1 - \mu_2, \mu_1 - \mu_3)_{90\%}$.

Simple linear regression model

- Simple linear regression model of n observations $\{x_i, y_i\}_{i=1}^n$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where x – regressor,

y – regressand,

β_0 – intercept of true regression line,

β_1 – slope of true regression line,

ε – disturbance term with $\mathsf{E}(\varepsilon) = 0$ and $\mathsf{V}(\varepsilon) = \sigma^2 > 0$.

- True regression line:

$$\mathsf{E}(y) = \beta_0 + \beta_1 x.$$

- Assumptions (to satisfy Gauss-Markov theorem):
  1. values of regressor x_i are constants,
  2. noise instants are uncorrelated: $\mathsf{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j.$

Sample regression line

- Under assumptions from above y_i are uncorrelated random variables with

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_i \quad \text{and} \quad \mathsf{V}(y_i) = \sigma^2.$$

- Sample regression line:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

- $\widehat{y}$ is an estimate of true population value of y at the point x.
- Values of regressand on the sample line, corresponding to x_i:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

- Estimate of ε_i is a difference between observation y_i and estimate $\widehat{y}_i$, denoted as e_i:

$$e_i = \widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i.$$

OLS estimates

- Quadratic minimization of $\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2$:

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- Results:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{SS_{xy}}{SS_{xx}}.$$

Moments of $\widehat{\beta}_0$ and $\widehat{\beta}_1$

- Expected values:

$$\mathsf{E}\left(\widehat{\beta}_0\right) = \beta_0, \qquad \mathsf{E}\left(\widehat{\beta}_1\right) = \beta_1.$$

- Variances:

$$\mathsf{V}\left(\widehat{\beta}_0\right) = \frac{\overline{x^2}\sigma^2}{SS_{xx}}, \qquad \mathsf{V}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{SS_{xx}}.$$

- Standard errors:

$$\text{S.E.}\left(\widehat{\beta}_0\right) = \sigma\sqrt{\frac{\overline{x^2}}{SS_{xx}}}, \qquad \text{S.E.}\left(\widehat{\beta}_1\right) = \frac{\sigma}{\sqrt{SS_{xx}}}.$$

Estimation of variance

- Model of simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \mathsf{E}(\varepsilon_i) = 0, \mathsf{V}(\varepsilon_i) = \sigma^2.$$

- Variation, created by ε_i (residual, error):

$$RSS = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right)^2$$

  with $n - 2$ degrees of freedom.

- Point estimate of σ^2:

$$\widehat{\sigma^2} = \frac{RSS}{n-2} = MSE.$$

Problem 1

For a sample of twenty monthly observations, a financial analyst wants to regress the percentage of return (Y) of the common stock of a corporation on the percentage rate of return (X) of the S&P 500 index. The following information is available:

$$\sum_{i=1}^{20} y_i = 22.6 \quad \sum_{i=1}^{20} x_i = 25.4 \quad \sum_{i=1}^{20} x_i^2 = 145.7 \quad \sum_{i=1}^{20} x_i y_i = 150.5 \quad \sum_{i=1}^{20} y_i^2 = 196.2$$

1. Estimate the linear regression of Y on X.
2. Interpret the slope of the sample regression line.
3. Interpret the intercept of the sample regression line.

Normal regression

- Now assume $\varepsilon_1, \ldots, \varepsilon_n \sim$ i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$.
- Subsequently $y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$ and they are independent.
- Linear combination of normal variables is normal:

$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\overline{x^2}\sigma^2}{SS_{xx}}\right) \qquad \text{and} \qquad \widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SS_{xx}}\right).$$

- Pivot functions to estimate β_0 and β_1:

$$\frac{\widehat{\beta}_0 - \beta_0}{\text{S.E.}\left(\widehat{\beta}_0\right)} \sim \mathcal{N}(0,1) \qquad \text{and} \qquad \frac{\widehat{\beta}_1 - \beta_1}{\text{S.E.}\left(\widehat{\beta}_1\right)} \sim \mathcal{N}(0,1).$$

- Can't be used when σ^2 is unknown (almost always).

Estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$

- Let's replace standard errors with estimated standard errors, where we replace σ^2 with its point estimate – MSE:

$$\text{E.S.E.}\left(\widehat{\beta}_0\right) = \sqrt{\frac{\overline{x^2} \cdot MSE}{SS_{xx}}}, \qquad \text{E.S.E.}\left(\widehat{\beta}_1\right) = \sqrt{\frac{MSE}{SS_{xx}}}.$$

- According to Fisher's lemma:

$$\frac{(n-2) \cdot MSE}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}.$$

- Pivot functions then:

$$\frac{\widehat{\beta}_0 - \beta_0}{\text{E.S.E.}\left(\widehat{\beta}_0\right)} \sim t_{n-2} \qquad \text{and} \qquad \frac{\widehat{\beta}_1 - \beta_1}{\text{E.S.E.}\left(\widehat{\beta}_1\right)} \sim t_{n-2}.$$

- MSE is independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Intervals estimation for β_0 and β_1

- $(1-\alpha)\%$ confidence intervals for β_0 and β_1:

$$(\beta_0)_{1-\alpha} \in \widehat{\beta}_0 \pm t_{n-2;\,\alpha/2} \cdot \text{E.S.E.}\left(\widehat{\beta}_0\right),$$
$$(\beta_1)_{1-\alpha} \in \widehat{\beta}_1 \pm t_{n-2;\,\alpha/2} \cdot \text{E.S.E.}\left(\widehat{\beta}_1\right).$$

- Test statistic for $\text{H}_0 : \beta_0 = b_0$:

$$T_{n-2}\Big|_{\text{H}_0} = \frac{\widehat{\beta}_0 - b_0}{\text{E.S.E.}\left(\widehat{\beta}_0\right)}.$$

- Test statistic for $\text{H}_0 : \beta_1 = b_1$:

$$T_{n-2}\Big|_{\text{H}_0} = \frac{\widehat{\beta}_1 - b_1}{\text{E.S.E.}\left(\widehat{\beta}_1\right)}.$$

Problem 2

Doctors are interested in the relationship between the dosage of a medicine and the time required for a patient's recovery. The following table shows, for a sample of five patients, dosage levels (in grams) and recovery times (in hours). These patients have similar characteristics except for medicine dosages.

Dosage levels	1.2	1.0	1.5	1.2	1.4
Recovery time	25	40	10	27	16

1. Estimate the linear regression of recovery time on dosage.
2. Find and interpret 90% confidence interval for the slope of the population regression line.
3. Would the sample regression derived in part 1 be useful in predicting recovery time for a patient given 2.5 grams of this drug? Explain your answer.

Confidence interval for $\mathsf{E}(y)$

Goal

To estimate real population value of y, corresponding to arbitrary x

- Simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

- $\mathsf{E}(y)$ is a value on true regression line for given x:

$$\mathsf{E}(y) = \beta_0 + \beta_1 x.$$

- $\mathsf{E}(y)$ is a function of x, so it's denoted as $\mu(x)$.
- Point estimate of $\mu(x)$ is a value on sample regression line:

$$\widehat{\mu}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

The estimator is unbiased.

Normal regression in $\mathsf{E}(y)$ estimation

- In normal regression:

$$\widehat{\mu}(x) \sim \mathcal{N}\left(\mu(x), \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \cdot \frac{\sigma^2}{SS_{xx}}\right).$$

- Possible pivot function:

$$\frac{\widehat{\mu}(x) - \mu(x)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \cdot \frac{\sigma^2}{SS_{xx}}}} \sim \mathcal{N}(0,1).$$

- And again, it can't be used when σ^2 is unknown.

Confidence interval for $\mathsf{E}(y)$

- Replacing σ^2 with MSE:

$$\frac{\widehat{\mu}(x) - \mu(x)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \cdot \frac{MSE}{SS_{xx}}}} \sim t_{n-2}.$$

- $(1-\alpha)\%$ confidence interval for $\mu(x)$:

$$(\mu(x))_{1-\alpha} \in \widehat{\mu}(x) \pm t_{n-2;\,\alpha/2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \cdot \frac{MSE}{SS_{xx}}}.$$

Prediction interval for y

Goal

To predict observation y, corresponding to arbitrary x

- $(1-\alpha)\%$ prediction interval of a random variable y is an interval, which contains y with probability $1-\alpha$.
- Confidence intervals – for unknown parameters (constants). Prediction intervals – for random variables.
- Assumption: $y(x)$ is independent of $y_1, \ldots, y_n$.
- Simple linear regression model:

$$y(x) = \beta_0 + \beta_1 x + \varepsilon = \mu(x) + \varepsilon.$$

- In normal regression:

$$y(x) \sim \mathcal{N}\left(\mu(x), \sigma^2\right).$$

Prediction interval for y

- Both $y(x)$ and $\widehat{\mu}(x)$ are normally distributed with mean $\mu(x)$:

$$y(x) - \widehat{\mu}(x) \sim \mathcal{N}\left(0, \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \cdot \frac{\sigma^2}{SS_{xx}}\right)$$

- Replacing σ^2 with MSE right away:

$$\frac{y(x) - \widehat{\mu}(x)}{\sqrt{MSE\left(1 + \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \cdot \frac{1}{SS_{xx}}\right)}} \sim t_{n-2}.$$

- $(1-\alpha)\%$ prediction interval for $y(x)$:

$$(y(x))_{1-\alpha} \in \widehat{\mu}(x) \pm t_{n-2;\,\alpha/2} \cdot \sqrt{MSE\left(1 + \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \cdot \frac{1}{SS_{xx}}\right)}.$$

Problem 3

The midterm and final exam scores of 10 students in a statistics course are tabulated as shown.

Midterm	Final	Midterm	Final
70	87	67	73
74	79	70	83
80	88	64	79
84	98	74	91
80	96	82	94

1. Calculate the least squares regression line of final exam scores on midterm exam scores for these data.
2. Plot the points and the least squares regression line on the same graph.

Problem 3

3. Find the value of MSE.
4. Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 > 0$ at 2.5% significance level.
5. Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at 5% significance level.
6. Find a 95% confidence interval for $\mathsf{E}(final)$, when $midterm = 68, 75$ and 82.
7. Find a 95% prediction interval for $final$, when $midterm = 68, 75$ and 82.

Look at the time!