Statistics 2

Simple Linear Regression II: Inference

Casper Albers & Jorge Tendeiro

Lecture 2, 2019 – 2020

Overview

Literature for this lecture

Read:

- Section 9.5.
- Additional text in reader:
  Casper Albers - 'Inference for Correlations'.

Simple linear regression

$$\underbrace{y = \alpha + \beta x}_{\text{Population}} \quad \longrightarrow \quad \underbrace{\widehat{y} = a + bx}_{\text{Sample}}$$

- a: Sample estimate of α.
- b: Sample estimate of β.

Values of a and b vary from sample to sample.
More interesting question:

$$\boxed{\text{What about the population parameters } \alpha \text{ and } \beta\text{?}}$$

Answer: Inference.

Regression model

Population

Population regression equation — $E(Y) = \alpha + \beta x$

link individual score and mean

Statistical model — $y_i = \alpha + \beta x_i + \varepsilon_i$

$\Rightarrow$ population parameters — α, β — $N(0, \sigma)$

Sample

Estimation in sample (OLS) — a, b, and s

Estimated equation — $\widehat{y_i} = a + bx_i$

Regression model

Population

Population unknown

Probability statements about the unknown population

Tests and CI's

Assumption: $\varepsilon \sim \mathcal{N}(0, \sigma)$

Sample

Estimation in sample (OLS)

a, b, and s

Estimated equation

$\widehat{y_i} = a + bx_i$

Beyond the sample

$$\underbrace{y_i = \alpha + \beta x + \varepsilon_i}_{\text{Population}} \quad \longrightarrow \quad \underbrace{y_i = a + bx + e_i}_{\text{Sample}}$$

Inference in regression models depends on crucial assumptions:

- The residuals are normally distributed with equal SD σ: $\varepsilon_i \sim \mathcal{N}(0, \sigma)$.
- The residuals are independent from x.

If these asumptions are met, it can be shown that the sampling distributions of a and b are also normal distributions:

$$a \sim \mathcal{N}(\alpha, \sigma_a) \qquad b \sim \mathcal{N}(\beta, \sigma_b)$$

Beyond the sample

Problem: σ_a and σ_b are unknown, because they depend on σ (the SD of the residuals in the population).
Solution: Use s (from the sample) instead of σ.

Result: The SE for the slope is given as follows

$$\boxed{\sigma_b \simeq SE_b = \frac{s}{\sqrt{\sum (x-\bar{x})^2}}}$$

SE_b is smaller when:

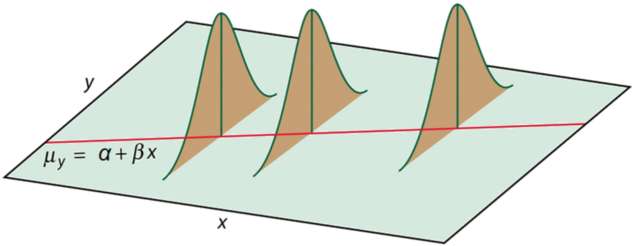
- s decreases, that is, the residuals around the regression line decrease.
- $\sum (x-\bar{x})^2$ increases (e.g., by increasing the sample size).

Notes:

- We don't look at SE_a, not particularly instructive.
- Because we replaced σ by $s = \sqrt{\frac{\sum_i e_i^2}{n-2}}$, the normal distributions is replaced by $t(n-2)$.

Beyond the sample

- Many (sub)populations defined by the values of x.
- Variable y is normally distributed in each (sub)population.
- The expected value (i.e., conditional mean) of y is $E(Y)$ and defined through $E(Y) = \alpha + \beta x$.
- The standard deviation of y is σ, constant.



Beyond the sample

	α	β
CI	$a \pm t^*_{n-2}\text{SE}_a$	$b \pm t^*_{n-2}\text{SE}_b$
Test	$\mathcal{H}_0 : \alpha = 0$ vs $\mathcal{H}_a : \alpha \neq 0$	$\mathcal{H}_0 : \beta = 0$ vs $\mathcal{H}_a : \beta \neq 0$
	$t = \frac{a}{\text{SE}_a} \sim t(n-2)$	$t = \frac{b}{\text{SE}_b} \sim t(n-2)$

Note: Typically we are mostly interested in making inference for β (the slope).

Example – Crime data

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	2.5%	97.5%
1	(Intercept)	209.920	135.613		1.548	0.128	−62.748	482.588
	PovertyRate	25.452	9.260	0.369	2.749	0.008	6.833	44.072

- $a = 209.920$
  $b = 25.452$
- $SE_a = 135.613$
  $SE_b = 9.260$
- CI for β: $b \pm t^*_{50-2} SE_b = 25.452 \pm 2.011 \times 9.260$
- Test: $t = b/SE_b = 2.749 \rightarrow p = .008$. Reject $\mathcal{H}_0$.

Summary: Inference in regression

Confidence Interval for β:

$$\boxed{b \pm t^* \mathrm{SE}_b}$$

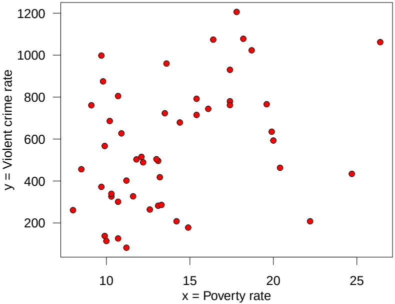
with t^* critical value t_{n-2}-distribution.

Test for β: $\mathcal{H}_0$: $\beta = 0$ vs. $\mathcal{H}_a$: $\beta \neq 0$:

$$\boxed{t = b/\mathrm{SE}_b}$$

Under $\mathcal{H}_0$, t has the t_{n-2}-distribution.

Correlation: Crime data



Pearson Correlations

		PovertyRate	ViolentCrime
PovertyRate	Pearson's r	—	
	p-value	—	
ViolentCrime	Pearson's r	0.369	—
	p-value	0.008	—

Just as a and b, the estimate r of ρ will vary per sample.

Inference in correlation

Test $\mathcal{H}_0$: $\rho = 0$ vs. $\mathcal{H}_a$: $\rho \neq 0$:

- Test statistic

$$\boxed{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

  Under $\mathcal{H}_0$, t has the t_{n-2}-distribution.

- This test only works for $\mathcal{H}_0$: $\rho = \rho_0$ when $\rho_0 = 0$.

Example – Crime data

Pearson Correlations

		PovertyRate	ViolentCrime
PovertyRate	Pearson's r	—	
	p-value	—	
ViolentCrime	Pearson's r	0.369	—
	p-value	0.008	

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.369\sqrt{48}}{\sqrt{1-0.369}} = 4.05$$

$$t^*_{48} = 2.011.\ \text{Reject } \mathcal{H}_0\ (\alpha = 5\%).$$

Confidence intervals

$\boxed{\text{estimate} \pm \text{critical value} \times \text{standard error}}$

Simple Linear Regression

- Sampling distribution of b: $\mathcal{N}(\beta, \sigma_b)$.
- $\Rightarrow$ CI: $b \pm t^* \mathrm{SE}_b$.
- t^* critical value from t_{n-2} distribution.

Correlation

- Sampling distribution of r is not normal. Not even symmetrical.
- An interval in the form '$r \pm$ something $\times$ SE_r' is not appropriate.

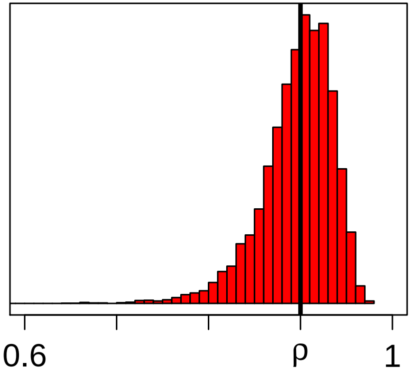
Sampling distribution of r

- When $\rho = 0$, the sampling distribution of r is approximately normal.
- That is why for $\mathcal{H}_0$: $\rho = 0$ a t-test is still possible.
- When $\rho \neq 0$, the sampling distribution is not symmetric:
  - Suppose that $\rho = 0.9$. Sample values 0.2 lower (thus 0.7) are possible. Sample values 0.2 higher (thus 1.1) are impossible.
  - $-1 \leq r \leq 1$. Skewed sampling distribution.
- Work-around to get CI's and tests: Fisher Z-transformation.

See Ex. 9.64 and Additional Text 1

Sampling distribution of r

From population with $\rho = 0.9$.
10,000 samples of size $n = 30$ have been drawn.
For each sample, r has been computed.



Fisher Z-transformation

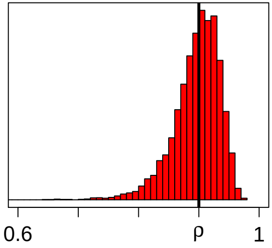
- General idea
  Not a normal distribution?
  $\Rightarrow$ transform to (approximate) normality.
- Transform r such that the transformed correlation r_z is (approximately) normal.
- Fisher Z-transformation: $\boxed{r_z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)}$
- r_z is approximately normal with
  - Mean $= \rho_z$ (with $\rho_z = \frac{1}{2}\log[(1+\rho)/(1-\rho)]$).
  - SD $= 1/\sqrt{n-3}$.

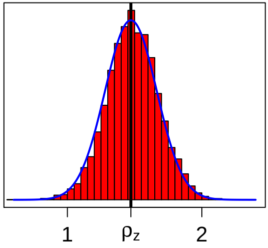
Note: Whenever in this Course we use 'log' we mean the natural logarithm, ('ln').

Sampling distribution r_z

Population with correlation $\rho = 0.90$.
10,000 samples of size $n = 30$.
Histogram of sample correlation r.

Histogram of transformed sample correlation r_z. Approximately $\mathcal{N}(\rho_z = 1.47, sd = 0.192)$





Confidence interval for ρ

- $r_z \sim \mathcal{N}(\rho_z, 1/\sqrt{n-3})$
- CI for ρ_z: $\boxed{r_z \pm z^* \frac{1}{\sqrt{n-3}}}$, z^* from $\mathcal{N}(0,1)$.
- CI for ρ:
  - Transform the CI for ρ_z back to one for ρ.
  - Inverse Fisher Z-transformation:

$$r = \frac{\mathrm{e}^{2r_z} - 1}{\mathrm{e}^{2r_z} + 1}$$

- CI for ρ:

$$\boxed{(\mathrm{LB}, \mathrm{UB}) = \left(\frac{\mathrm{e}^{2\mathrm{LB}_z} - 1}{\mathrm{e}^{2\mathrm{LB}_z} + 1}, \frac{\mathrm{e}^{2\mathrm{UB}_z} - 1}{\mathrm{e}^{2\mathrm{UB}_z} + 1} \right)}$$

Example – Crime data

- $n = 50, \; r = 0.358.$
- $r_z = \frac{1}{2} \log\left(\frac{1+0.358}{1-0.358}\right) = 0.375.$
- CI for ρ_Z:

$$0.375 \pm 1.96 \times \frac{1}{\sqrt{50-3}} \Rightarrow (0.089, 0.661).$$

- CI for ρ:

$$\left(\frac{e^{2\times 0.089} - 1}{e^{2\times 0.089} + 1}; \frac{e^{2\times 0.661} - 1}{e^{2\times 0.661} + 1}\right) = (0..088, 0.579).$$

- Note that the estimate $r = .358$ does not lie in the center of this CI.

Tests for ρ

- Remember the general formula for a test statistic:

$$\boxed{\text{test statistic} = \frac{\text{estimate} - \text{value } H_0}{\text{SE}}.}$$

- Not applicable for ρ directly, but applicable for ρ_Z (for which a sampling distribution is available).
- $\mathcal{H}_0$: $\rho = \rho_0$ vs. $\mathcal{H}_a$: $\rho \neq \rho_0$.
  Test statistic:

$$\boxed{Z = \frac{r_z - \rho_z}{1/\sqrt{n-3}} \sim \mathcal{N}(0,1).}$$

- p-value for this test on ρ_Z is used for ρ.

Example – Crime data

- $r = 0.358 \Rightarrow r_z = 0.375$.
- $\mathcal{H}_0$: $\rho = 0.30$ vs. $\mathcal{H}_a$: $\rho > 0.30$.
- Test
  $\mathcal{H}_0$: $\rho_z = 0.310$ vs $\mathcal{H}_a$: $\rho_z > 0.310$.
- $$Z = \frac{r_z - \rho_z}{1/\sqrt{n-3}} = \frac{0.375 - 0.310}{1/\sqrt{47}} = 0.446.$$
- **Conclusion:** Do not reject $\mathcal{H}_0$.

Confidence intervals

1. Confidence Interval for model parameter β

$$b \pm t^*_{n-2}\mathrm{SE}_b.$$

2. Confidence Interval for mean response $E(Y)$

$$E(Y) \pm t^*_{n-2}\mathrm{SE}_{\widehat{\mu}}.$$

3. Prediction Interval for a value of y

$$\widehat{y} \pm t^*_{n-2}\mathrm{SE}_{\widehat{y}}.$$

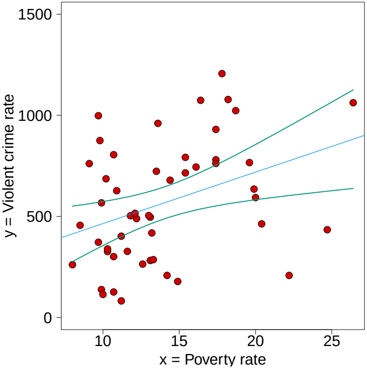
All based on the t-distribution with $n-2$ df's.

Intervals for $E(Y)$ and $\widehat{y}$

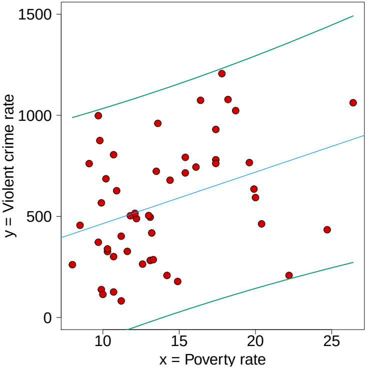
- Filling a value for x in the regression line $a + bx$ implies:
  1. Estimating the mean response: $E(Y) = a + bx$.
  2. Predicting a value of y: $\widehat{y} = a + bx$.
- For both there is a SE, but the prediction-SE is larger:
  - The width of the interval for $E(Y)$ describes the uncertainty in estimating $E(Y)$.
  - Prediction $\widehat{y} = E(Y) + \widehat{\varepsilon}_y$.
  - It has additional variance because individual values are spread around the mean $E(Y)$.

Example – Crime data

Confidence interval



Prediction interval



Next lecture

Contents:

- Model assumptions and violations
  Causality & Association

Read:

Agresti, Section 9.6, Ch. 10