M8S1 - Regression Inference

Professor Jarad Niemi

STAT 226 - Iowa State University

November 29, 2018

Regression Inference

- Review of population mean inference
  - Assumptions
  - Confidence interval
  - p-value
  - Hypothesis test
- Regression inference
  - Assumptions
  - Confidence interval
  - p-value
  - Hypothesis test

Population mean assumptions

What is an inference? Making a statement about the population based on a sample.

What are our assumptions when making an inference about a population mean?

- Data are independent
- Data are normally distributed
- Data are identically distributed with a common mean and a common variance

This is encapsulated with the statistical notation

$$Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$$

Statistics for a population mean

If we have the assumption $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$,

- What is our estimator for μ?
  sample mean, so $\hat{\mu} = overliney$
- What is our estimator for σ^2?
  sample variance, s^2
- What is the standard error of $\hat{\mu}$?
  $SE[\hat{\mu}] = s/\sqrt{n}$

Confidence intervals for a population mean

If we have the assumption $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, what is the formula to construct a $100(1-\alpha)\%$ confidence interval for the population mean μ?

$$\overline{y} \pm t_{n-1,1-\alpha/2} s/\sqrt{n}$$

where $P(T_{n-1} > t_{n-1,1-\alpha/2}) = \alpha/2$.

More generally, we have

$$\hat{\mu} \pm t^* \times SE[\hat{\mu}]$$

where

- $\hat{\mu}$ is the estimator of the population mean
- t^* is the appropriate t-critical value
- $SE[\hat{\mu}]$ is the standard error of the estimator

t-statistic for a population mean

Suppose you have the null hypothesis

$$H_0 : \mu = m_0$$

What is the formula for the t-statistic?

$$t = \frac{\overline{y} - m_0}{s/\sqrt{n}} = \frac{\hat{\mu} - m_0}{SE[\hat{\mu}]}$$

Thus we have the estimator minus the hypothesized value in the numerator and the standard error of the estimator in the denominator.

If the null hypothesis is true, what is the distribution for t?

$$t \sim T_{n-1}$$

Hypothesis test for population mean

Suppose you have the hypotheses

$$H_0 : \mu = m_0 \qquad \text{versus} \qquad H_a : \mu > 0$$

How can you calculate the p-value for this test?

$$p\text{-value} = P(T_{n-1} > t) = P\left(T_{n-1} > \frac{\hat{\mu} - m_0}{SE[\hat{\mu}]}\right)$$

At level α, you

- reject H_0 if p-value $\le \alpha$ and conclude that there is statistically significant evidence that $\mu > 0$
- fail to reject H_0 if p-value $> \alpha$ and conclude that there is insufficient evidence that $\mu > 0$.

Assumptions

In statistical notation, the regression assumptions can be written as

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

for some unknown population intercept (β_0), population slope (β_1), and error for individual i (ϵ_i).

What are the assumptions for the regression model?

- Errors are independent
- Errors are normal
- Errors are identically distributed with a mean of 0 and a variance of σ^2
- Linear relationship between the explanatory variable and the mean of the response:

$$E[Y_i] = \beta_0 + \beta_1 x_i$$

You might also see regression written like

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Statistics for regression

(You do not need to know the formulas.)

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

- For the slope (β_1), the estimator is the sample slope
$$\hat{\beta}_1 = b_1 = r \times s_y / s_x$$
- For the intercept (β_0), the estimator is the sample intercept
$$\hat{\beta}_0 = b_0 = \overline{y} - b_1 \overline{x}$$
- For the variance (σ^2), the estimator is
$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y - b_0 - b_1 x_i)^2$$

Standard errors for regression

(You do not need to know the formulas.)

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

The important standard errors are

$$SE[\hat\beta_1] = SE[b_1] = \hat\sigma \sqrt{\frac{1}{(n-1)s_x^2}}$$

and

$$SE[\hat\beta_0] = SE[b_0] = \hat\sigma \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}}$$

We can use these to construct confidence intervals and pvalues.

Confidence intervals for regression

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

$100(1-\alpha)\%$ confidence interval for the slope:

$$b_1 \pm t_{n-2,1-\alpha/2} \times SE[b_1]$$

$100(1-\alpha)\%$ confidence interval for the intercept:

$$b_0 \pm t_{n-2,1-\alpha/2} \times SE[b_0]$$

To remember the degrees of freedom, it is always the sample size minus the number of parameters in the mean. In this case, there are two parameters in the mean: β_0 and β_1.

Hypothesis tests

Although alternative hypothesis tests can be constructed for different hypothesized values, the vast majority of the time we are testing versus a hypothesized value of 0 and typically only caring about the slope.

Suppose you have these hypotheses about the slope

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_a : \beta_1 \neq 0$$

Then our t-statistic is

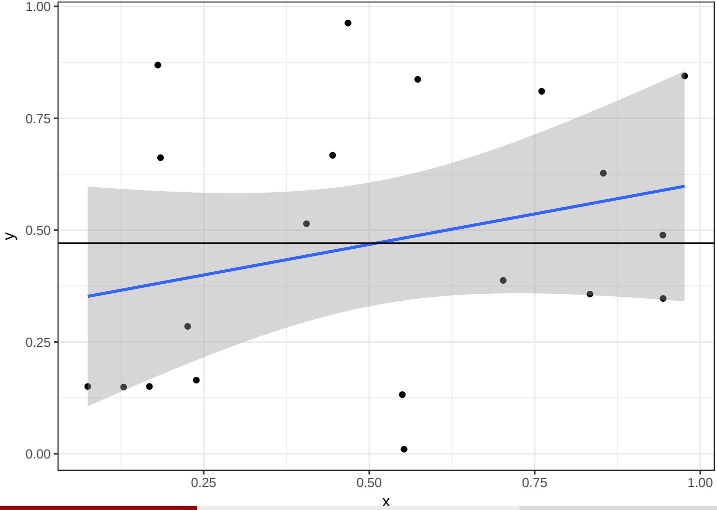
$$t = \frac{\hat{\beta}_1 - 0}{SE[\hat{\beta}_1]} = \frac{b_1 - 0}{SE[b_1]}$$

and a p-value is

$$p\text{-value} = 2P(T_{n-2} > |t|).$$

Why do we care about $\beta_1 = 0$?

If $\beta_1 = 0$, then $y_i = b_0 + \epsilon_i$, i.e. our response variable is independent of our explanatory variable.


1.00
0.75
0.50
0.25
0.00
y
0.25
0.50
0.75
1.00
x