Course: Mathematical statistics

Week 13: Confidence interval for regression coefficient

Lecturer: Nagulama Moses

Kumi University

June 7, 2025

Outline

1. Introduction
2. Confidence Intervals for Regression Coefficients β_0 and β_1
3. Interval Estimation of the Mean Response $E(Y|x_0)$
4. Prediction Interval for a New Observation Y at x_0
5. Confidence Intervals for β_0, β_1, and σ^2
6. Assumptions for Interval Estimation

Intended learning outcome

- Explain the concept of a confidence interval and its purpose in the context of linear regression.
- Identify and interpret the confidence interval for both the slope and intercept of a simple linear regression model.
- Understand and state the assumptions necessary to construct valid confidence intervals for regression coefficients (e.g., normality, independence, homoscedasticity).

Introduction

In regression analysis, we often estimate how a change in an independent variable x affects a dependent variable y. This relationship is summarized through regression coefficients specifically the slope and intercept.

However, the estimates we obtain from sample data are not exact.
Due to sampling variability, they are subject to uncertainty.
Thus, it is not enough to merely compute the point estimate of a regression coefficient (like $\hat{\beta}_1$) for the slope); we also need to quantify the precision of our estimate.

This is where confidence intervals come into play.

What is a Confidence Interval?

A confidence interval (CI) provides a range of values that is likely to

contain the true (but unknown) population regression coefficient with a specified level of confidence (e.g., 95%).

Confidence Intervals for Regression Coefficients β_0 and β_1

Let:

- $\hat{\beta}_1$: estimated slope
- $SE(\hat{\beta}_1)$: standard error of the slope

Then a $100(1-\alpha)\%$ confidence interval for β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)$$

Similarly, for the intercept β_0

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_0)$$

Where:

- $t_{\alpha/2,n-2}$: critical value from the t-distribution
- n: number of data points

Interval Estimation of the Mean Response $E(Y|x_0)$

To estimate the mean response at $x = x_0$:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Standard error of $\hat{y}_0$:

$$SE(\hat{y}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Then the $100(1 - \alpha)\%$ confidence interval is:

$$\hat{y}_0 \pm t_{\alpha/2, n-2} \cdot SE(\hat{y}_0)$$

Prediction Interval for a New Observation Y at x_0

A prediction interval estimates a single new observed value Y at x_0:

$$SE_{\text{pred}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Prediction interval:

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \cdot SE_{\text{pred}}$$

This includes the additional variability from the individual observation.

Confidence Intervals for β_0, β_1, and σ^2

- In addition to obtaining point estimates of the regression parameters β_0, β_1, and the error variance σ^2, it is essential to construct confidence intervals (CIs) to assess precision and reliability.
- The width of a CI reflects uncertainty in the regression line:
  - Narrower intervals $\Rightarrow$ more precise estimates.

Assumptions for Interval Estimation

To construct valid confidence intervals, assume:

- Errors ε_i are independent.
- Errors are normally distributed: $\varepsilon_i \sim N(0, \sigma^2)$.
- Homoscedasticity (constant error variance).

CI for the Slope β_1

- Let $\hat{\beta}_1$: estimated slope, and $SE(\hat{\beta}_1)$: standard error of slope.
- The $100(1-\alpha)\%$ CI for β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2,\, n-2} \cdot SE(\hat{\beta}_1)$$

- That is,

$$\left[\hat{\beta}_1 - t_{\alpha/2,\, n-2} \cdot SE(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2,\, n-2} \cdot SE(\hat{\beta}_1)\right]$$

CI for the Intercept β_0

- Let $\hat{\beta}_0$: estimated intercept, and $SE(\hat{\beta}_0)$: standard error.
- The $100(1-\alpha)\%$ CI for β_0 is:

$$\hat{\beta}_0 \pm t_{\alpha/2,\, n-2} \cdot SE(\hat{\beta}_0)$$

- That is,

$$\left[\hat{\beta}_0 - t_{\alpha/2,\, n-2} \cdot SE(\hat{\beta}_0),\ \hat{\beta}_0 + t_{\alpha/2,\, n-2} \cdot SE(\hat{\beta}_0)\right]$$

Interpretation

- The confidence intervals have a frequentist interpretation:
- If we repeat sampling and construct 95% CIs each time, then in approximately 95% of those samples, the interval will contain the true value of β_0, β_1, or σ^2.

CI for the Error Variance σ^2

- Estimate: $\hat{\sigma}^2 = MSE = \dfrac{RSS}{n-2}$
- Based on normal errors:

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

- The $100(1-\alpha)\%$ CI for σ^2 is:

$$\left[\frac{(n-2)\cdot MSE}{\chi^2_{1-\alpha/2,\, n-2}},\ \frac{(n-2)\cdot MSE}{\chi^2_{\alpha/2,\, n-2}}\right]$$

Example 1: Confidence Interval for the Slope β_1

Question:

A study investigates the relationship between hours studied (X) and exam score (Y) for 5 students. The regression equation is:

$$\hat{Y} = 2.1 + 0.85X$$

The standard error of the slope $\hat{\beta}_1$ is $SE(\hat{\beta}_1) = 0.22$. Construct a 95% confidence interval for the true slope β_1.

Solution

Given:

- $\hat{\beta}_1 = 0.85$
- $SE(\hat{\beta}_1) = 0.22$
- $n = 5 \Rightarrow df = 3$
- $\alpha = 0.05, \; t_{0.025,3} = 3.182$

CI formula:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)$$

$$0.85 \pm 3.182 \cdot 0.22 = 0.85 \pm 0.700$$

$$\Rightarrow [0.15, 1.55]$$

The 95% confidence interval for β_1 is [0.15, 1.55].

Example 2: Confidence Interval for the Intercept β_0

Question:

Using the same regression model:

$$\hat{Y} = 2.1 + 0.85X$$

Suppose the standard error of the intercept is $SE(\hat{\beta}_0) = 0.95$. Construct a 95% confidence interval for β_0.

Solution

Given:

- $\hat{\beta}_0 = 2.1$
- $SE(\hat{\beta}_0) = 0.95$
- $t_{0.025,3} = 3.182$

CI formula:

$$2.1 \pm 3.182 \cdot 0.95 = 2.1 \pm 3.023$$

$$\Rightarrow [-0.923, 5.123] \approx [-0.92, 5.12]$$

The 95% confidence interval for β_0 is [-0.92, 5.12].

Example 3: Confidence Interval for Error Variance σ^2

Question:

In a regression analysis with $n = 10$ observations, the residual sum of squares is $RSS = 18.0$. Construct a 95% confidence interval for the error variance σ^2.

Solution

Given:

- $RSS = 18.0, \; n = 10$
- $df = 8$

Mean Squared Error:

$$MSE = \frac{RSS}{n-2} = \frac{18}{8} = 2.25$$

From chi-square table:

- $\chi^2_{0.025,8} = 17.535$
- $\chi^2_{0.975,8} = 2.180$

Confidence interval:

$$\left[\frac{8 \cdot 2.25}{17.535}, \frac{8 \cdot 2.25}{2.180}\right] = [1.027, 8.257] \approx [1.03, 8.26]$$

The 95% confidence interval for σ^2 is [1.03, 8.26].

Interval Estimation of the Mean Response

One key application of simple linear regression is estimating the mean value of the response variable $E(y)$ at a specific value of the predictor variable x.

This is referred to as **predicting the mean response**.

Suppose we want to estimate the mean shear strength of a propellant bond in a rocket motor made from a batch that is 10 weeks old.

- $x_0 = 10$ (weeks)
- Estimate $E(y \mid x_0)$, the mean response at $x = x_0$
- Assume x_0 lies within the range of observed x values

Point Estimator

From the fitted simple linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The point estimator of $\mu_{y|x_0} = E(y \mid x_0)$ is:

$$\hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Substitute x_0 into the regression equation.

Variance of the Estimator

- $\hat{\mu}_{y|x_0}$ is normally distributed.
- Its variance is:

$$\text{Var}(\hat{\mu}_{y|x_0}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

Where:

- n = sample size
- $\bar{x}$ = sample mean of x_i
- $S_{xx} = \sum (x_i - \bar{x})^2$

Estimating σ^2

Since σ^2 is unknown, estimate it with:

$$\hat{\sigma}^2 = MS_{\text{Res}} = \frac{RSS}{n-2}$$

Where RSS is the residual sum of squares.

Confidence Interval for $E(y \mid x_0)$

A $100(1-\alpha)\%$ confidence interval for $E(y \mid x_0)$:

$$\hat{\mu}_{y|x_0} \pm t_{\alpha/2,n-2} \cdot \sqrt{MS_{\text{Res}}\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

Where $t_{\alpha/2,n-2}$ is the critical value from the t-distribution with $n-2$ degrees of freedom.

Interpretation and Properties

- CI gives the range in which true $E(y \mid x_0)$ lies with specified confidence.
- Width depends on:
    - Sample size n
    - Data spread S_{xx}
    - Distance of x_0 from $\bar{x}$
- Narrowest when $x_0 = \bar{x}$
- Wider as x_0 moves away from $\bar{x}$

Example 1: Estimating the Mean Response at a Given Point

Problem:
Given the model:

$$\hat{y} = 2.5 + 0.8x$$

- $n = 6$, $\bar{x} = 4$, $S_{xx} = 10$
- $MS_{Res} = 1.2$
- Estimate 95% CI at $x_0 = 6$

Example Solution

Point estimate
$\hat{\mu}_{y|6} = 2.5 + 0.8(6) = 7.3$
Standard Error (SE)

$$SE = \sqrt{1.2\left(\frac{1}{6} + \frac{(6-4)^2}{10}\right)}$$
$$= \sqrt{1.2\,(0.167 + 0.4)} = \sqrt{1.2 \times 0.567}$$
$$= \sqrt{0.6804} \approx 0.825$$

Example 1: Continued

Critical value

$t_{0.025,4} = 2.776$

Margin of Error

$ME = 2.776 \times 0.825 = 2.29$

CI

$CI = 7.3 \pm 2.29 = (5.01, 9.59)$

95% CI for mean response at $x = 6$ is **(5.01, 9.59)**.

Example 2: CI When $x_0 = \bar{x}$

Problem:

Model: $\hat{y} = 1.2 + 3.5x$

- $n = 8,\ \bar{x} = 10,\ x_0 = 10$
- $S_{xx} = 40,\ MS_{Res} = 2.0$

Example 2: Solution

Point estimate

$\hat{\mu}_{y|10} = 1.2 + 3.5(10) = 36.2$

SE (since $x_0 = \bar{x}$)

$SE = \sqrt{2.0\left(\frac{1}{8}\right)} = \sqrt{0.25} = 0.5$

$t_{0.05,6} \approx 1.943$

CI

$CI = 36.2 \pm (1.943 \times 0.5) = 36.2 \pm 0.972 = (35.23, 37.17)$

90% CI for mean response at $x = 10$ is **(35.23, 37.17)**.

Example 3: Wider Interval for a Distant x_0

Problem

Model: $\hat{y} = 10 - 1.5x$

- $n = 12$, $\bar{x} = 4$, $S_{xx} = 36$
- $MS_{Res} = 3.6$, $x_0 = 8$

Example 3: Solution

Point estimate
$\hat{\mu}_{y|8} = 10 - 1.5(8) = -2$
SE

$$SE = \sqrt{3.6\left(\frac{1}{12} + \frac{(8-4)^2}{36}\right)}$$
$$= \sqrt{3.6\,(0.083 + 0.444)} = \sqrt{3.6 \times 0.527}$$
$$= \sqrt{1.897} \approx 1.378$$

Example

$t_{0.025,10} = 2.228$

CI

$CI = -2 \pm 2.228 \times 1.378 = -2 \pm 3.07 = (-5.07, 1.07)$ 95% CI for mean response at $x = 8$ is **(-5.07, 1.07)**.

References

- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*.
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models*.
- Any standard textbook on regression or applied statistics.

Thank You!
Test of hypothesis as regression coefficient