Simple Linear Regression

Model Assessment, Confidence Intervals, and Hypothesis Tests

Ravleen Bajaj

Simple Linear Regression Model

- Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma^2)$.
- $\varepsilon_i \sim N(0, \sigma^2)$ and the ε_i's are independent
- Goal: Estimate the association between predictor x_i and response Y_i.

Model Assessment

- Residual analysis:
  - Residuals $\varepsilon_i = y_i - \hat{\mu}_i$.
  - Check assumptions: Independence, Linearity, Homoscedasticity (Common SD), Normality.

Independence Assumption

- One key assumption of simple linear regression: **errors are independent**.
- **Condition:** If sampling occurs at the individual level, individuals are measured only once, and individuals are not ordered in time or space.
- This means that the error for one subject does not provide information about the error for another subject.
- Violations occur when:
  - Repeated measurements on the same individual are treated as independent.
  - Data are ordered in time (time series) or space (spatial data).
  - Clustered or family data are analyzed as if independent.

Assuming Independence

- If data are sampled at the **individual level**:
- Each person (or unit) is measured only once.
- Individuals are not ordered in time.
- Individuals are not ordered in space (not spatially autocorrelated).
- $\Rightarrow$ Under these conditions, we assume **errors are independent**.

Example of Violation

- Measuring blood pressure of patients **every week for 6 months** and analyzing as if each measurement were independent.
- Problem: Observations from the same patient are likely correlated.

Violation: Time Series Data

- Situation: Daily blood glucose measurements on the same patient for 30 days.
- Problem: Consecutive values are correlated (today's level depends on yesterday's).
- Consequence:
  - Errors are autocorrelated.

Violation: Clustered or Family Data

- Situation: Measuring BMI of children from the same family.
- Problem: Children from the same family are more similar than children from different families.
- Consequence:
  - Errors within a family are correlated.

Example of Independence

- Measuring blood pressure **once per patient** on a random sample of patients.
- No temporal or spatial ordering of participants.
- In this case, independence of errors is a reasonable assumption.

Confidence Intervals

- Estimates: $\hat{\beta}_0, \hat{\beta}_1$.
- Standard errors: $SE_{\tilde{\beta}_j}$.
- $100(1-\alpha)\%$ CI for parameter β_j:

$$\hat{\beta}_j \pm t_{n-2,1-\alpha/2} \times SE_{\tilde{\beta}_j}$$

where $j = 0, 1$.

Hypothesis Tests for Parameters

- Test significance of slope β_1:

$$H_0 : \beta_1 = b_1 \quad \text{vs.}$$

$$H_a : \beta_1 \neq b_1 \text{ or } H_a : \beta_1 > b_1 \text{ or } H_0 : \beta_1 < b_1$$

- Test statistic:

$$t = \frac{\hat{\beta}_1 - b_1}{SE_{\tilde{\beta}_1}}$$

  If H_0 is true then t is a random draw from a t_{n-2} distribution.

- Decision rule:
  - Reject H_0 if $|t| > t_{n-2,1-\alpha/2}$
  - Reject H_0 if $t > t_{n-2,\,1-\alpha}$
  - Reject H_0 if $t < -t_{n-2,\,1-\alpha}$

Example: Lung Function Study

- Data collected from 20 adults in a health survey.
- For each participant:
  - Number of years of smoking.
  - Lung capacity (liters) measured by spirometry.
- Question: **Is lung capacity associated with years of smoking?**

Regression Model

- Define:

$$Y_i = \text{Lung capacity (liters)}, \quad x_i = \text{Years of smoking of } i^{th} \text{ individual}$$

- Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and the } \varepsilon_i\text{'s are independent}$$

- Hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{(No relationship between smoking and lung capacity)}$$

$$H_a : \beta_1 \neq 0 \quad \text{(Lung capacity depends on smoking years)}$$

Confidence Interval Example

- Example: If $\hat{\beta}_1 = 2.5$, $SE_{\tilde{\beta}_1} = 0.8$, $n = 20$.
- Test statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE_{\tilde{\beta}_1}}, \quad n = 20$$

- Under H_0, t is a random draw from a t_{n-2} distribution.
-

$$t = \frac{2.5 - 0}{0.8} = 3.125$$

- With $df = n - 2 = 18$, the critical value is $t_{18,0.975} \approx 2.101$.
- Since $3.125 > 2.101$, we reject H_0.

Thank You!

Questions?