01L

June 19, 2025

1 Lecture Notes: Simple Linear Regression with Time as a Covariate

2 Setup and Model Description

We observe data points:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

where: - x_i: the covariate value (in our case, time, such as months from January 1959 onward), - y_i: the response value (e.g., population at the i-th month).

We are interested in modeling how the response y changes over time. So we apply a linear regression model where time is the covariate.

In the context of a U.S. population dataset from January 1959 to December 2024: - n: total number of months, - $x_i = i$: time index (e.g., January 1959 is 1, February 1959 is 2, and so on), - y_i: observed population in the i-th month.

---

3 The Linear Regression Model

We assume the data follows the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where: - β_0: intercept, - β_1: slope (effect of time), - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$: normally distributed error term.

So, the observed value y_i is modeled as a straight line plus some noise.

Equivalently:

$$y_i \overset{\text{indep}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

---

4 Parameters in the Model

There are three unknowns (parameters) that we want to estimate: - β_0: intercept — value of y when $x = 0$ - β_1: slope — how much y increases when x increases - σ^2: variance — how spread out the data is around the line

---

5 The Likelihood Function (Easy Explanation)

The likelihood function answers this: > **"If I pick some values for β_0, β_1, σ, how likely is it that I'd see the data I observed?"**

Each point y_i has a probability of occurring based on how far it is from the predicted value $\beta_0 + \beta_1 x_i$.

We assume the errors are normally distributed, so:

$$\text{Probability of } y_i = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

For all n points (assuming they are independent), we multiply the probabilities:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

This simplifies to:

$$L(\beta_0, \beta_1, \sigma) = (2\pi)^{-n/2} \cdot \sigma^{-n} \cdot \exp\left(-\frac{S(\beta_0, \beta_1)}{2\sigma^2}\right)$$

where:

$$S(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

This $S(\beta_0, \beta_1)$ is the **sum of squared errors**, and it tells us how far off our model is.

---

6 Assumptions of the Model

1. **Fixed Inputs**: $x_1, \ldots, x_n$ are fixed (not random).
2. **i.i.d. Errors**: The noise terms ϵ_i are independent and normally distributed:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

3. **Linearity**: The response variable y depends linearly on x.