Inference in Simple Linear Regression

Residuals and Error Variance Estimation

- **Residual**:

  The difference between the observed value Y_i and the predicted value $A + Bx_i$:

  $\text{Residual}_i = Y_i - (A + Bx_i)$

- **Sum of Squares of Residuals (SSR)**:

  $SSR = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2$

- **Key Point**:

  $\frac{SSR}{\sigma^2} \sim \chi^2_{n-2}$

  (follows chi-square distribution with n−2 degrees of freedom).

✅ Thus:

$E\left[\frac{SSR}{n-2}\right] = \sigma^2$

meaning $\frac{SSR}{n-2}$ is an **unbiased estimator** of σ^2.

Proof Outline: Why Degrees of Freedom = n−2

- Since $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$,

  the standardized errors are independent standard normals.

- Summing squared errors (without estimating parameters) gives χ^2_n.

- Estimating A and B uses up **2 degrees of freedom**, thus:

$$\frac{SSR}{\sigma^2} \sim \chi^2_{n-2}$$

Maximum Likelihood Estimators (MLEs)

- If Y_i are normal, the **joint density**:

$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-\alpha-\beta x_i)^2/2\sigma^2}$

- To maximize likelihood → **Minimize**:

  $\sum(y_i - \alpha - \beta x_i)^2$

✅ **Conclusion**:

Least squares estimators A and B are also the **maximum likelihood estimators**!

Summary Formulas

Define:

- $S_{xY} = \sum(x_i - \bar{x})(Y_i - \bar{Y}) = \sum x_i Y_i - n\bar{x}\bar{Y}$
- $S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$
- $S_{YY} = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$

Then:

$$B = \frac{S_{xY}}{S_{xx}} \qquad A = \bar{Y} - B\bar{x} \qquad SSR = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

Distribution of Estimators A and B

Under the normality assumption:

- $A \sim N\left(\alpha, \sigma^2 \frac{\sum x_i^2}{nS_{xx}}\right)$
- $B \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$

Also:

$$\frac{SSR}{\sigma^2} \sim \chi^2_{n-2} \quad \text{and} \quad SSR \text{ is independent of } A, B$$

Hypothesis Testing about β

Test:

- **Null Hypothesis**: $H_0 : \beta = 0$
- **Alternative Hypothesis**: $H_1 : \beta \neq 0$

Test Statistic:

$TS = \sqrt{(n-2)S_{xx}}\frac{|B|}{\sqrt{SSR}}$

Under $H_0, TS \sim t_{n-2}$.

✅ **Decision Rule**:

Reject H_0 if:

$$|TS| > t_{\gamma/2,n-2}$$

where $t_{\gamma/2,n-2}$ is the critical value from the t-distribution table.

Confidence Interval for β

A $100(1-\alpha)\%$ confidence interval for β is:

$$\left(B - t_{\alpha/2,n-2}\sqrt{\frac{SSR}{(n-2)S_{xx}}},\ B + t_{\alpha/2,n-2}\sqrt{\frac{SSR}{(n-2)S_{xx}}}\right)$$

Inference for α (Intercept)

Similarly, the confidence interval for α can be built:

$$A \pm t_{\alpha/2,n-2}\sqrt{\frac{SSR}{n(n-2)S_{xx}}\sum x_i^2}$$

Prediction at a New Point x_0

- Predict mean response at x_0:

$\text{Predicted value} = A + Bx_0$

- **Distribution**:

$$A + Bx_0 \sim N\left(\alpha + \beta x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

- **Confidence interval**:

$$A + Bx_0 \pm t_{\alpha/2, n-2} \sqrt{\frac{SSR}{n-2} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

✅ This gives a range where the mean value at x_0 is expected to lie.
