Applied Statistical Analysis I

Bivariate regression, inference, and prediction

Elena Karagianni, PhD Candidate
karagiae@tcd.ie

October 22, 2025

Department of Political Science, Trinity College Dublin

Model Fit and Goodness of Fit

Explained vs. unexplained variation

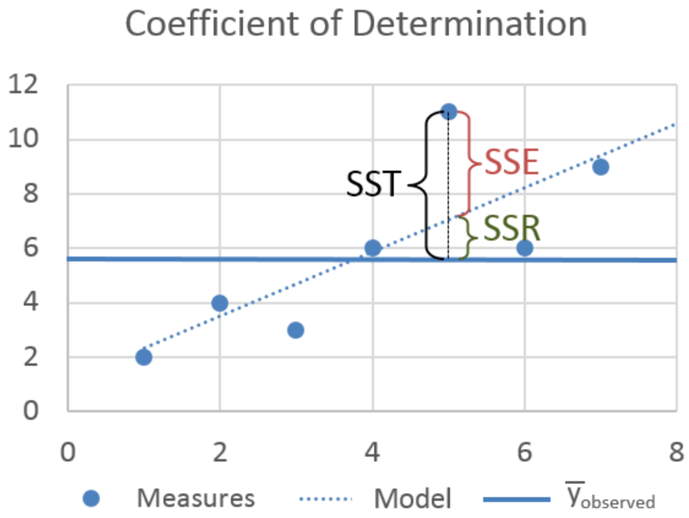
$$Y_i = \hat{Y}_i + e_i$$

- **Total variation:** $SST = \sum(Y_i - \bar{Y})^2$
- **Explained variation:** $SSR = \sum(\hat{Y}_i - \bar{Y})^2$
- **Unexplained variation:** $SSE = \sum(Y_i - \hat{Y}_i)^2$

Relationship:

$$SST = SSR + SSE$$

Visualization



R^2: Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Measures the proportion of variance in Y explained by X.
- $R^2 = 0$ means the model explains none of the variation.
- $R^2 = 1$ means perfect fit (rarely observable in real data).
- A higher R^2 indicates better fit, but not necessarily a better or causal model.

In political science and social data, R^2 values around 0.3–0.6 are typical — relationships are often probabilistic, not deterministic.

Adjusted R^2 (for multiple regression)

When we include more variables:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

- Penalizes adding variables that don't improve model fit much.
- Only increases if new variable improves explanatory power beyond chance.

Adjusted R^2 is preferred when comparing models with different numbers of predictors.

Summary

Putting it all together

1. Estimate coefficients $(\hat{\alpha}, \hat{\beta})$ by OLS.
2. Check assumptions and residual plots.
3. Compute standard errors.
4. Perform hypothesis test:
$$t = \frac{\hat{\beta}}{se(\hat{\beta})}$$
5. Compute and interpret confidence intervals.
6. Evaluate overall model fit (R^2, adjusted R^2). (We'll talk about it in later lectures).
7. Translate results into substantive conclusions.