Refresher Course in Calculus, Probability, and Statistics

Day 5a: Linear Regression

Instructor: Dr. Guillaume Wald

11 September 2026

Scatterplot, Sample Covariance and Sample Correlation

- What is the relationship between variables X and Y?
- Different ways to summarise the relationship between variables:
  - **Scatterplot:**

    Plot of the n observations on X_i and Y_i, each represented by a point (X_i, Y_i).

    $\rightarrow$ a good idea to begin the analysis
  - **Sample covariance:**

$$\widehat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y});$$

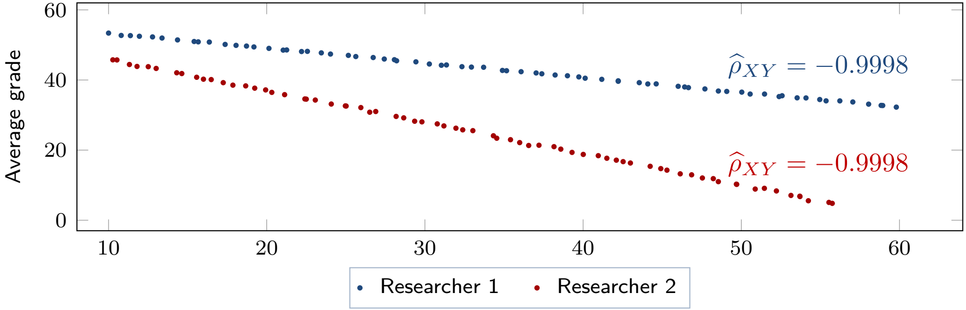
$\rightarrow$ **Sample correlation:**

$$\widehat{\rho}_{xy} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x \, \widehat{\sigma}_y} \qquad \in [-1, 1],$$

A measure of the strength of the *linear* association between X and Y in the sample.

Sample Covariance and Sample Correlation

Hypothetical example: grades and student–teacher ratio.



Same correlation, despite **different slopes**:

- Correlation measures the strength and direction of *linear association*.
- = How tightly the observations follow a **straight-line relationship**, while the sign indicates the direction of the slope.

Linear Regression

Univariate model

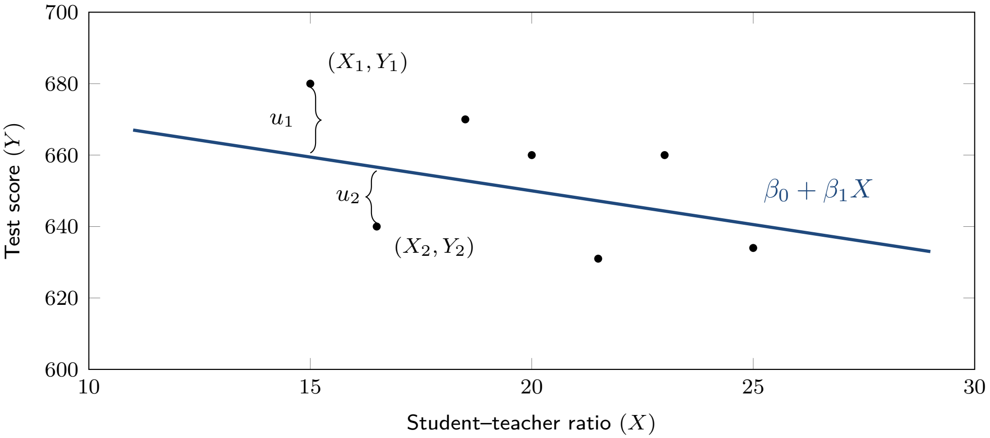
- By how much does Y change if X changes by one unit?

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

  - Y_i: regressand or dependent / explained / left-hand variable;
  - X_i: regressor or independent / explanatory / right-hand variable;
  - β_1: slope of the *population regression line*;
  - β_0: intercept of the *population regression line*;
  - u_i: error term.

Linear Regression

Univariate model (continued)



Stock & Watson (2012)

- How are the coefficients of the linear regression model estimated?

$\rightarrow$ In practice, β_0 & β_1 are **unknown** $\Rightarrow$ we must use data to estimate them.

- **Ordinary Least Squares (OLS) estimator:**

$$\hat{\beta}_1 = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}^2_X}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

  - Predicted values: $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$;
  - Residuals: $\widehat{u}_i = Y_i - \widehat{Y}_i$.

- **Practical interpretation of the coefficients:**
  - $\hat{\beta}_1$: average change in Y associated with a one-unit change in X;
  - $\hat{\beta}_0$: expected value of Y when X is zero.

Linear Regression

Univariate model (continued) — measures of fit

- **Measures of fit:** the R^2.
  - $R^2 \in [0, 1]$;
  - interpretation: the **fraction of the variance** in the dependent variable that is **explained by the variance in the independent variable(s)**;
  - an alternative measure in the multivariate case is the **adjusted** R^2, which accounts for the number of independent variables.

Linear Regression

Univariate model (continued) — OLS output

. reg testscr str

Source	SS	df	MS
Model	7794.11004	1	7794.11004
Residual	144315.484	418	345.252353
Total	152109.594	419	363.030056

Number of obs	420
F(1, 418)	22.58
Prob > F	0.0000
R-squared	0.0512
Adj R-squared	0.0490
Root MSE	18.581

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	−2.279808	.4798256	−4.75	0.000	−3.222980	−1.336636
_cons	698.933	9.467491	73.82	0.000	680.3232	717.5428

Population model: $\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$

Estimated model: $\widehat{\text{TestScore}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{STR}_i = 698.93 - 2.28 \cdot \text{STR}_i$

(all other numbers are discussed in the following slides)

Linear Regression

Univariate model (continued) — hypothesis testing on $\hat{\beta}_1$

- Set up the hypotheses:

$$
\begin{aligned}
\text{Two-sided:} \quad & H_0 : \beta_1 = c \quad \text{vs.} \quad H_1 : \beta_1 \neq c, \\
\text{One-sided:} \quad & H_0 : \beta_1 \geq c \quad \text{vs.} \quad H_1 : \beta_1 < c. \\
\text{or:} \quad & H_0 : \beta_1 \leq c \quad \text{vs.} \quad H_1 : \beta_1 > c.
\end{aligned}
$$

- Compute the empirical t-value:

$$
t_{\text{emp}} = \frac{\text{estimator} - \text{value under the null}}{\text{standard error of the estimator}} = \frac{\hat{\beta}_1 - c}{\widehat{\sigma}_{\hat{\beta}_1}}.
$$

- Compare it with the critical value of a t-distribution with $n - k - 1$ degrees of freedom, with k the number of regressors (slope coefficients / explanatory vars.) excluding the intercept:

$\rightarrow$ if $|t_{\text{emp}}| > t_{\text{crit}}(\alpha)$: reject the null at significance level α (error probability $< \alpha$).

- OR/AND: compute the p-value, $p = \Pr_{H_0}(|t| > t_{\text{emp}})$.
- When n is large, $t_{n-k-1} \to N(0,1)$, so with the standard normal Φ:
  - two-sided $(H_1 : \beta_1 \neq c)$: $p = \Pr_{H_0}(|Z| > t_{\text{emp}}) = 2\big[1 - \Phi(|t_{\text{emp}}|)\big]$;
  - one-sided $(H_1 : \beta_1 < c)$: $p = \Pr_{H_0}(Z < t_{\text{emp}}) = \Phi(t_{\text{emp}})$;
  - one-sided $(H_1 : \beta_1 > c)$: $p = \Pr_{H_0}(Z > t_{\text{emp}}) = 1 - \Phi(t_{\text{emp}})$.
- Decision rule: reject H_0 if $p < \alpha$.

- **Definition of a $(1-\alpha)$ confidence interval:**
  - the set of values that cannot be rejected by a two-sided test at significance level α;
  - an interval that contains the true value of β_1 in a fraction $1-\alpha$ of all samples.
- **Computing the $(1-\alpha)$ interval:**
  - two-sided: $\mathrm{CI}_{1-\alpha} = \left[\hat{\beta}_1 - t_{n-k-1,\,(\alpha/2)} \cdot \widehat{\sigma}_{\hat{\beta}_1},\ \ \hat{\beta}_1 + t_{n-k-1,\,(\alpha/2)} \cdot \widehat{\sigma}_{\hat{\beta}_1}\right]$;
  - one-sided $(H_1 : \beta_1 > c)$: $\mathrm{CI}_{1-\alpha} = \left[\hat{\beta}_1 - t_{n-k-1,\,\alpha} \cdot \widehat{\sigma}_{\hat{\beta}_1},\ +\infty\right)$;
  - one-sided $(H_1 : \beta_1 < c)$: $\mathrm{CI}_{1-\alpha} = \left(-\infty,\ \hat{\beta}_1 + t_{n-k-1,\,\alpha} \cdot \widehat{\sigma}_{\hat{\beta}_1}\right]$.
- For large n these critical values approach those of the standard normal distribution:
  - 10%: 1.645
  - 5%: 1.96
  - 1%: 2.58.

Linear Regression

Univariate model (continued) — robust standard errors

. reg testscr str, robust

Number of obs	420
F(1, 418)	19.26
Prob > F	0.0000
R-squared	0.0512
Root MSE	18.581

testscr	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	−2.279808	.5194892	−4.39	0.000	−3.300945	−1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3058

Estimated model (robust SEs in parentheses):

$$\widehat{\text{TestScore}}_i = \underset{(10.36)}{698.93^{***}} - \underset{(0.52)}{2.28^{***}} \cdot \text{STR}_i.$$

- More general, with k regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \qquad i = 1, \ldots, n.$$

  - predicted values: $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$;
  - residuals: $\widehat{u}_i = Y_i - \widehat{Y}_i$.
- Estimation, interpretation, testing and confidence intervals extend directly from the univariate case (with $n - k - 1$ degrees of freedom).