Introduction to Linear Regression

Sulemana Abdul-Karim

UofG

November 19, 2025

Outline

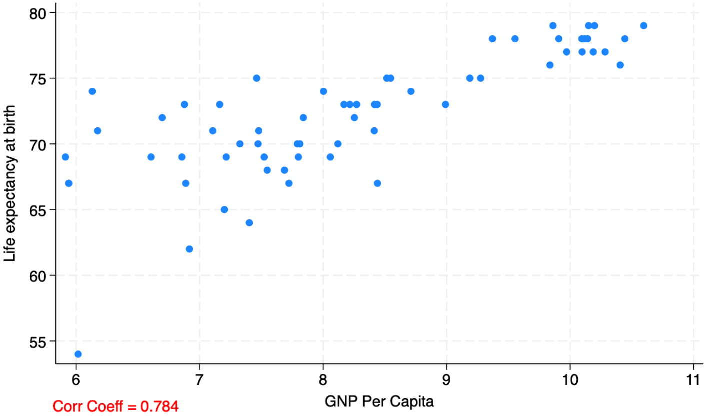
ILO

- Explain the purpose of linear regression
- Interpret regression coefficients
- Conduct and interpret hypothesis tests
- Evaluate the key assumptions and perform diagnostic tests

Intro

- Correlation measures strength and direction of a linear relationship between two variables.
  - Study hours - Students grades
  - Class attendance - grades
  - Parental education - Student achievement
  - Screen time - Academic performance
  - Civic education - Political tolerance
- Correlation is bounded and symmetric.
  - $-1 \leq r \leq 1$ and X with Y = Y with X

Intro



- Can correlation predict one variable from the other?

Outline

ILO

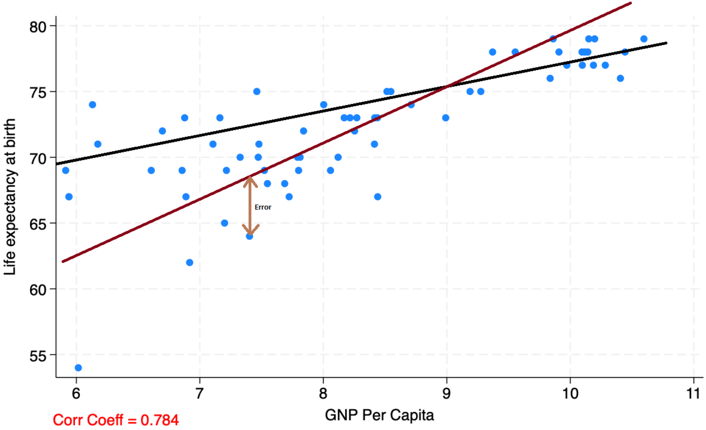
Linear regression

Statistical Aspects of Regression

Regression Diagnostics

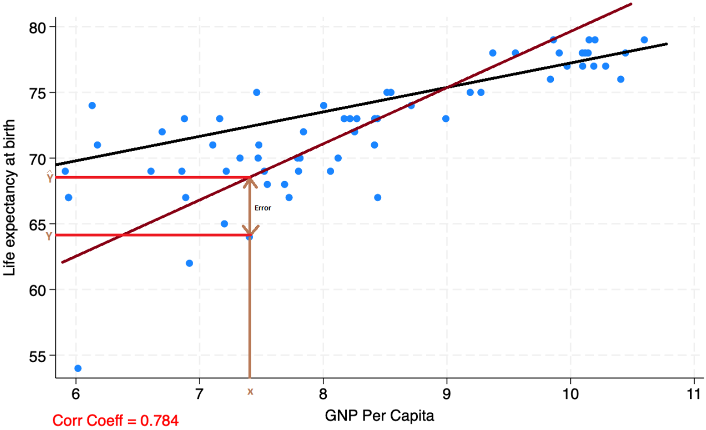
Line fitting

- We know that the correlation coefficient is a measure of how well the points will fit a straight line.
  - But which straight line is best?



Line fitting

- A straight line helps. Why?
  - A straight line is best described by $y = \alpha + \beta x$
  - We can therefore predict using only two parameters; α and β.
- Summarizing the relationship by a line causes errors.



Line fitting

- For each x_i, we have $y_i = \hat{y}_i + e_i$ or $y_i = \alpha + \hat{\beta} x_i + e_i$
- The errors is therefore defined as $e_i = y_i - (\alpha + \hat{\beta} x_i)$
  - By saying line fitting, we actually trying to find a line that causes least errors.
  - How do we define the error?
- A first idea would be the sum of all the errors corresponding to all the points:

$$C_0 = \sum_{i=1}^{n} e_i,$$

- However, we dislike positive errors as negative errors, but in the above definition positive and negative errors will cancel with each other.

Line fitting

- These next two are commonly used measures for the error in the full sample

$$C_1(\alpha, \beta) = \sum_{i=1}^{n} |e_i|, \qquad C_2(\alpha, \beta) = \sum_{i=1}^{n} e_i^2.$$

- $C_2(\alpha, \beta)$, the **least squares (LS) criterion** that measures the sum of squared errors, is by far the most frequently used. Also called **ordinary least squares (OLS)**.

- $C_1(\alpha, \beta)$ is called the **least absolute criterion**, which we will not be talking about.

- The solution to least squares yields

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathsf{Cov}(x, y)}{\mathsf{Var}(x)}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

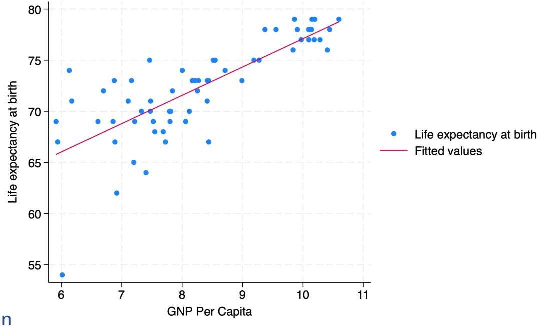
- Note: If we standardize both variables, the regression line passes through the origin, and the slope is the Pearson correlation coefficient.

Line fitting

Source	SS	df	MS
Model	873.264865	1	873.264865
Residual	548.671643	61	8.99461709
Total	1421.93651	62	22.9344598

Number of obs	=	63
F(1, 61)	=	97.09
Prob > F	=	0.0000
R-squared	=	0.6141
Adj R-squared	=	0.6078
Root MSE	=	2.9991

lexp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
llg	2.768349 Slope	.2809566	9.85	0.000	2.206542	3.330157
_cons	49.41502 intercept	2.348494	21.04	0.000	44.71892	54.11113



Multiple Regression

- Same line-fitting intuition holds when we include several variables.
- A multiple regression model with k explanatory variables or predictors is given as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + e$$

- We are fitting a linear model - it is useful to check that scatterplots of Y against each predictor are approximately linear.
- Similarly least squares is used to estimate $\alpha, \beta_1, \beta_2, ..., \beta_k$
- Example: Returns to Education

$$\text{wage} = \alpha + \beta_1 \text{education} + \beta_2 \text{experience} + \beta_3 \text{age} + \beta_4 \text{sex} + e$$

- Similar interpretation for each β_i
  - Caution - partial derivative means holding other variables constant.
  - be mindful about dummy variables too.

OLS Assumptions and Properties

1. **Linearity:** Model is linear in parameters
2. Data are independently and identically distributed (i.i.d.).
3. **No Perfect Multicollinearity:** Regressors are not exact linear combinations of each other.
4. **Zero Conditional Mean:**

$$E(u_i \mid X_1, X_2, \ldots, X_k) = 0$$

5. **Homoskedasticity:**

$$\text{Var}(u_i \mid X_1, X_2, \ldots, X_k) = \sigma^2$$

- **Implications:**
  - Under assumptions (1)–(4), OLS is **unbiased and consistent**:

$$E[\hat{\beta}_j] = \beta_j$$

  - Under all (1)–(5), OLS is **BLUE** (Best Linear Unbiased Estimator): Minimum variance among all linear unbiased estimators (Gauss–Markov Theorem).
- **If normality of errors** ($u_i \sim N(0, \sigma^2)$) also holds → then the OLS estimators are normally distributed even in small samples (t and F tests are exactly valid).

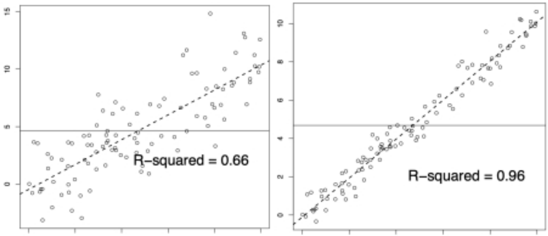
Goodness of Fit: R^2

- How well does the model explain or fit the observed data?
- R^2 (the **coefficient of determination**) measures how well the regression line explains the variation in the dependent variable.

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

where:

  - SSR = Sum of Squared Residuals $= \sum(Y_i - \hat{Y}_i)^2$
  - SST = Total Sum of Squares $= \sum(Y_i - \bar{Y})^2$
- Interpretation:
  - R^2 measures the **proportion of total variation** in Y explained by the model.
  - $0 \leq R^2 \leq 1$ — higher values indicate better fit.
- Limitation:
  - R^2 **always increases** when more variables are added, even if they are irrelevant.


R−squared = 0.66
R−squared = 0.96

Adjusted R^2

- Adjusted R^2 corrects the R^2 for the number of predictors in the model.

  $$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k}$$

  where:
  - $n =$ sample size
  - $k =$ number of explanatory variables
- Interpretation:
  - Penalizes the inclusion of unnecessary variables.
  - Can **decrease** if added variables do not improve model fit.
- Comparison:
  - Use R^2 to describe fit; use $\bar{R}^2$ to compare models with different numbers of predictors.
- Note:
  - $\bar{R}^2$ can be negative (when model fits worse than using the mean of Y).

Outline

Hypothesis Testing

- Hypothesis testing procedures compare a conjecture about a population parameter to the information contained in a sample.
- In every hypothesis test, five ingredients must be present:
  - A **null hypothesis** H_0
  - An **alternative hypothesis** H_1
  - A **test statistic**
  - A **rejection region** (or p-value)
  - A **conclusion**
- Using the sampling distributions of $\hat{\alpha}$, $\hat{\beta}$, and S^2, we can develop tests for hypotheses regarding the unknown population parameters α, β, and σ^2.

Null and Alternative Hypotheses

- The null hypothesis, H_0, specifies a value for a regression parameter.
- Consider the case of β.
- Typically, the null hypothesis is stated as:

  $$H_0 : \beta = \beta_0,$$

  where β_0 is a hypothesized value.
- Every H_0 is paired with a logical alternative hypothesis H_1.
- Three possible alternatives:
  - **Left-tailed:** $H_1 : \beta < \beta_0$
  - **Right-tailed:** $H_1 : \beta > \beta_0$
  - **Two-tailed:** $H_1 : \beta \neq \beta_0$
- The choice of H_1 depends on the research question and theory.

Hypothesis Testing in Regression

- After estimating the model, we often test whether each explanatory variable has a significant effect on Y.
- For a single coefficient β_j:

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

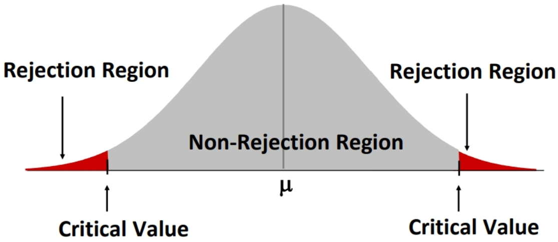
- Test statistic (t-test):

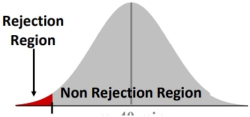
$$t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

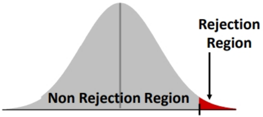
- Decision rule:
  - If $|t| > t_{\alpha/2,\, df}$, reject H_0.
- Equivalently, we coud use the p-value - the probability of observing a test statistic as extreme as the one computed, assuming H_0 is true.
  - A small p-value (< 0.05) $\rightarrow$ strong evidence against H_0.
  - A large p-value (> 0.10) $\rightarrow$ weak evidence; fail to reject H_0.
- Common significance levels:

$$\alpha = 0.10,\ 0.05,\ 0.01$$

  - If we reject H_0, the variable has a statistically significant effect on Y.


Rejection Region
Rejection Region
Non-Rejection Region
μ
Critical Value
Critical Value


Rejection Region
Non Rejection Region


Rejection Region
Non Rejection Region

Joint Significance Testing (F-test)

- Tests whether a group of coefficients are jointly equal to zero.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \text{At least one } \beta_j \neq 0$$

- Test statistic:

$$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(n-k)} \quad \sim F(q, n-k)$$

  where:
  - R_U^2: from the unrestricted model
  - R_R^2: from the restricted model (imposing H_0)
  - q: number of restrictions
- Decision rule:
  - If $F > F_{\alpha,q,n-k}$ or p-value $< \alpha$, reject H_0.
- Interpretation:
  - The model (or set of regressors) is jointly significant in explaining Y.

Outline

ILO

Linear regression

Statistical Aspects of Regression

Regression Diagnostics

Linearity

- The true model is:

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

- The model is correctly specified.
- Linearity refers to the way the parameters (β_0, β_1) and the error term u enter the equation — not necessarily to the relationship between Y and X themselves.
- **Other examples of linearity:**
  - $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
  - $\ln(Y_i) = \beta_0 + \beta_1 / X_i + u_i$
- **Examples of non-linearity:**
  - $Y_i = \beta_0 + \exp(\beta_1 X_i) + u_i$
  - $Y_i = \beta_0 + 1/(1 + \exp(\beta_1 X_i)) + u_i$

Model	Interpretation of $\widehat{\beta}_1$
Level-level $Y_i = \beta_0 + \beta_1 X_i + u_i$	An increase in X by 1 unit is associated with a change in Y by $\widehat{\beta}_1$ units on average
Log-level $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	An increase in X by 1 unit is associated with a change in Y by $(100 \times \widehat{\beta}_1)$% on average
Level-log $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	An increase in X by 1% is associated with a change in Y by $(\widehat{\beta}_1/100)$ units on average
Log-log $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	An increase in X by 1% is associated with a change in Y by $\widehat{\beta}_1$% on average

Multicollinearity

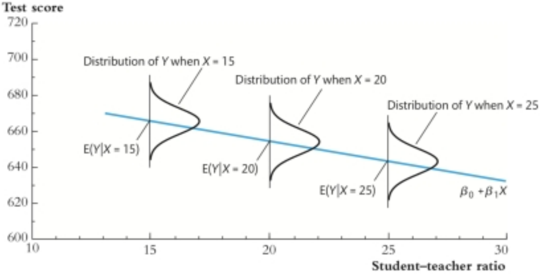
- Multicollinearity occurs when two or more highly correlated variables does have an "effect" on the response variable, but in the regression output some or all of them show insignificance.
- Caution - If you do not see any insignificance in the regression, you don't have to worry about multicollinearity problem.
- Whether a group of regressors have an effect can be tested using the F test
- Detection
  - High pairwise correlations among regressors.
- Remedies, should be guided by practical background or meaning of these variables
  - Drop one of the correlated variables.
  - Combine them (e.g., create an index or ratio).
  - Collect more data to reduce sampling variation.

Heteroskedasticity

- Heteroskedasticity occurs when the variance of the error term is not constant across observations.

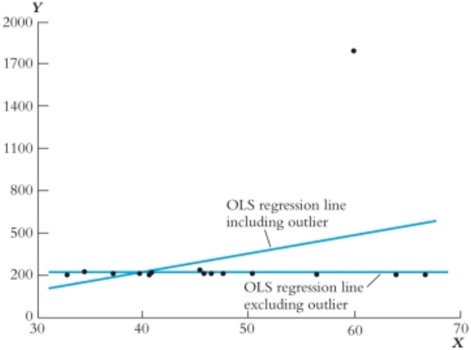
$$\mathsf{Var}(e_i \mid X_i) = \sigma_i^2 \neq \sigma^2$$

- Consequences
  - OLS estimates remain unbiased and consistent.
  - But standard errors are biased $\Rightarrow$ invalid t and F tests.
  - Confidence intervals and p-values become incorrect.
- Detection
  - Plot residuals $\hat{e}_i$ vs. fitted values $\hat{Y}_i$
  - Formal tests:
    - Breusch–Pagan test
    - White test
- Remedies
  - Use robust standard errors(White's correction).
  - Transform the model (e.g., use logs).


Test score
720
700
680
660
640
620
600
10
15
20
25
30
Student–teacher ratio
Distribution of Y when X = 15
Distribution of Y when X = 20
Distribution of Y when X = 25
E(Y|X = 15)
E(Y|X = 20)
E(Y|X = 25)
β0 + β1X

Outliers

- Outliers are observations with unusually large or small values relative to the rest of the data.
- They may arise from data entry errors, unusual events, or genuine extreme cases.
- Why Outliers Matter
  - Can distort regression estimates and predicted values.
  - May pull the regression line toward them, affecting slope and intercept.
  - Often inflate residual variance, reducing precision.
- Detection
  - Inspect scatterplots or boxplots.
  - Examine residuals / standardized residuals.
- Remedies
  - Verify data accuracy and correct errors.
  - Consider robust regression or transforming variables (e.g., log scale).
  - Remove outlier only if justified (document reasoning).


Y
2000
1700
1400
1100
800
500
200
0
30
40
50
60
70
X
OLS regression line
including outlier
OLS regression line
excluding outlier

Omitted/Redundant Variables

- Omitted variable bias occurs when a relevant explanatory variable is left out of the regression model.
- Consequence
  - The estimated coefficients become **biased and inconsistent**.
  - The direction of bias depends on the correlation between the omitted and included variables.
- Illustration

$$\text{True model: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

$$\text{Estimated model: } Y = \beta_0 + \tilde{\beta}_1 X_1 + e$$

  Then,

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

  $\Rightarrow$ If X_1 and X_2 are correlated, $\tilde{\beta}_1$ is biased.
- Remedies
  - Include all relevant variables that affect Y.
  - Start with the largest possible model, and then use significance test and/or F test to remove some variables, if any.

Dummy Variables in Regression

- **Definition:**
  - Dummy variables represent categorical information numerically.
  - They take values:

$$D_i = \begin{cases} 1, & \text{if category is present} \\ 0, & \text{otherwise} \end{cases}$$

- **Interpretation:**
  - The coefficient on a dummy variable measures the **difference in the intercept** between groups.

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

    If $D_i = 1$: $E(Y|D_i = 1) = \beta_0 + \beta_1$; If $D_i = 0$: $E(Y|D_i = 0) = \beta_0$

- **Multiple Categories:**
  - For m groups, use $m - 1$ dummies to avoid the **dummy variable trap** (perfect multicollinearity).
- **Interactions:**
  - Combine dummies with other variables to test different slopes:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

  - Allows slope and intercept to differ across groups.

Dummy Variable

- Consider the wage equation:

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{female}_i + \beta_3 (\text{educ}_i \times \text{female}_i) + u_i$$

  where:
  - educ_i = years of education
  - $\text{female}_i = 1$ if female, 0 if male
  - Interaction term allows education to affect wages differently by sex
- **Expected values:**

$$E(\text{wage}|\text{male}) = \beta_0 + \beta_1 \text{educ}$$
$$E(\text{wage}|\text{female}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{educ}$$

- **Interpretation Example:**
  - If $\beta_2 < 0$: females earn less than males at zero education (intercept gap).
  - If $\beta_3 < 0$: females have a smaller return to each additional year of education.
  - If $\beta_3 > 0$: education reduces the gender wage gap.