Statistics and Data Analysis
Unit 04 – Lecture 02 Notes
Simple Linear Regression (OLS)

Tofik Ali

February 17, 2026

Topic

Simple linear regression model, interpretation, residuals and R-squared.

How to Use These Notes

These notes are written for students who are seeing the topic for the first time. They follow the slide order, but add the missing 'why', interpretation, and common mistakes. If you get stuck, look at the worked exercises and then run the Python demo.

Course repository (slides, demos, datasets): `https://github.com/tali7c/Statistics-and-Data-Analysis`

Time Plan (55 minutes)

- 0–10 min: Attendance + recap of previous lecture
- 10–35 min: Core concepts (this lecture's sections)
- 35–45 min: Exercises (solve 1–2 in class, rest as practice)
- 45–50 min: Mini demo + interpretation of output
- 50–55 min: Buffer / wrap-up (leave 5 minutes early)

Slide-by-slide Notes

Title Slide

State the lecture title clearly and connect it to what students already know. Tell students what they will be able to do by the end (not just what you will cover).

Quick Links / Agenda

Explain the structure of the lecture and where the exercises and demo appear.

- Overview

- Model
- Fit and Diagnostics
- Exercises
- Demo
- Summary

Learning Outcomes

- Write the simple linear regression model
- Interpret slope and intercept in context
- Compute a prediction and a residual
- Explain R-squared (intuition)

Why these outcomes matter. **Regression** models a response variable Y as a function of predictor(s) X. It has direction (predictors -> response), produces a fitted equation, and lets you predict and explain. Regression is not automatically causal; causality needs design or strong assumptions. A **residual** is $y - \hat{y}$. Residual plots tell you what the model failed to explain. Patterns in residuals (trend, curvature, changing variance) are warnings that your model form is inadequate or assumptions are violated.

Model: Key Points

- y = b0 + b1 x + error
- Slope: expected change in y for 1-unit increase in x
- Intercept: predicted y at x=0 (interpret carefully)

Explanation. In chi-square tests, **expected counts** are what you would expect to see if H_0 were true (e.g., independence). Very small expected counts can break the approximation used by the test; a common rule of thumb is that most expected counts should be at least 5. In simple linear regression, the **slope** is the expected change in Y for a one-unit increase in X (on average). Always state units (e.g., 'marks increase by 2 points per extra hour of study'). The **intercept** is the predicted value when $X = 0$. It may or may not be meaningful depending on whether $X = 0$ is realistic in your context. If $X = 0$ is outside the observed range, do not over-interpret the intercept.

Model: Key Formula

$$y = \beta_0 + \beta_1 x + \epsilon$$

Fit and Diagnostics: Key Points

- Look at residual plots for patterns
- Outliers can dominate the fitted line
- High R^2 does not guarantee a good model

Explanation. A **residual** is $y - \hat{y}$. Residual plots tell you what the model failed to explain. Patterns in residuals (trend, curvature, changing variance) are warnings that your model form is inadequate or assumptions are violated. R^2 is the fraction of variance in Y explained by the model (in-sample). It increases when you add predictors, even useless ones, so it is not a guarantee of a good model. Use residual diagnostics and out-of-sample evaluation to judge model quality.

Exercises (with Solutions)

Attempt the exercise first, then compare with the solution. Focus on interpretation, not only arithmetic.

Exercise 1: Prediction

Model: yhat = 10 + 2x. Predict y when x=7.

Solution

- yhat = 24

Exercise 2: Residual

If actual y=20 at x=7, compute residual.

Solution

- 20-24 = -4

Walkthrough. A **residual** is $y - \hat{y}$. Residual plots tell you what the model failed to explain. Patterns in residuals (trend, curvature, changing variance) are warnings that your model form is inadequate or assumptions are violated.

Exercise 3: Interpret slope

Slope is 5 thousand INR per extra room. Interpret.

Solution

- Each extra room increases predicted price by 5k INR (on average).

Walkthrough. In simple linear regression, the **slope** is the expected change in Y for a one-unit increase in X (on average). Always state units (e.g., 'marks increase by 2 points per extra hour of study').

Mini Demo (Python)

Run from the lecture folder:

```
python demo/demo.py
```

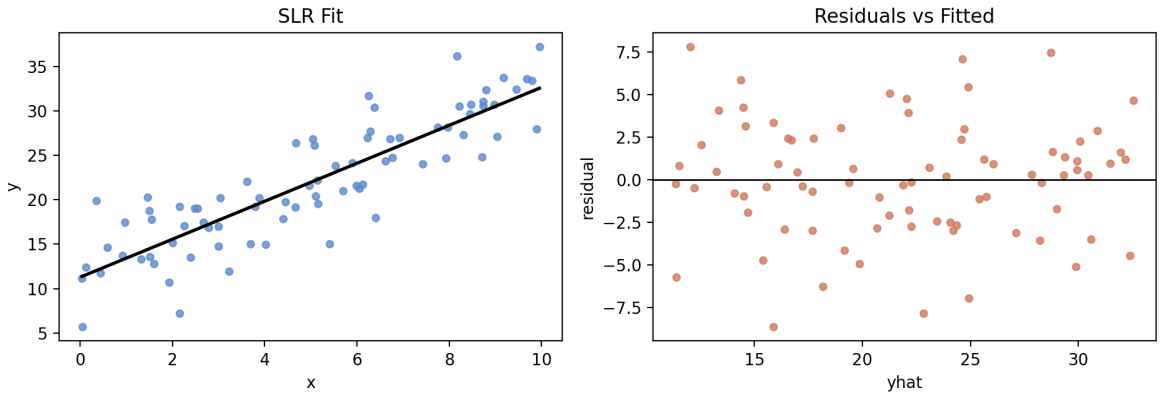
Output files:

- `images/demo.png`
- `data/results.txt`

What to show and say.

- Fits a simple linear regression on synthetic data (one predictor).
- Shows scatter + fitted line and reports slope/intercept and R^2.
- Use residual behavior (in results) to motivate diagnostics.

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

Why do we check residual plots even if R^2 is high?

Suggested answer (for revision). Residual plots reveal nonlinearity, outliers, or changing variance; R^2 alone can look good even when assumptions are violated.

References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley.
- Devore, J. L. *Probability and Statistics for Engineering and the Sciences*, Cengage.
- McKinney, W. *Python for Data Analysis*, O'Reilly.

Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

Title Slide

https://github.com/tali7c/Statistics-and-Data-Analysis

Quick Links

Overview Model Fit and Diagnostics Exercises Demo Summary

Agenda

- Overview
- Model
- Fit and Diagnostics
- Exercises
- Demo
- Summary

Learning Outcomes

- Write the simple linear regression model
- Interpret slope and intercept in context
- Compute a prediction and a residual
- Explain R-squared (intuition)

Model: Key Points

- y = b0 + b1 x + error
- Slope: expected change in y for 1-unit increase in x
- Intercept: predicted y at x=0 (interpret carefully)

Model: Key Formula

$$y = \beta_0 + \beta_1 x + \epsilon$$

Fit and Diagnostics: Key Points

- Look at residual plots for patterns
- Outliers can dominate the fitted line
- High R^2 does not guarantee a good model

Exercise 1: Prediction

Model: yhat = 10 + 2x. Predict y when x=7.

Solution 1

- yhat = 24

Exercise 2: Residual

If actual y=20 at x=7, compute residual.

Solution 2

- 20-24 = -4

Exercise 3: Interpret slope

Slope is 5 thousand INR per extra room. Interpret.

Solution 3

- Each extra room increases predicted price by 5k INR (on average).

Mini Demo (Python)

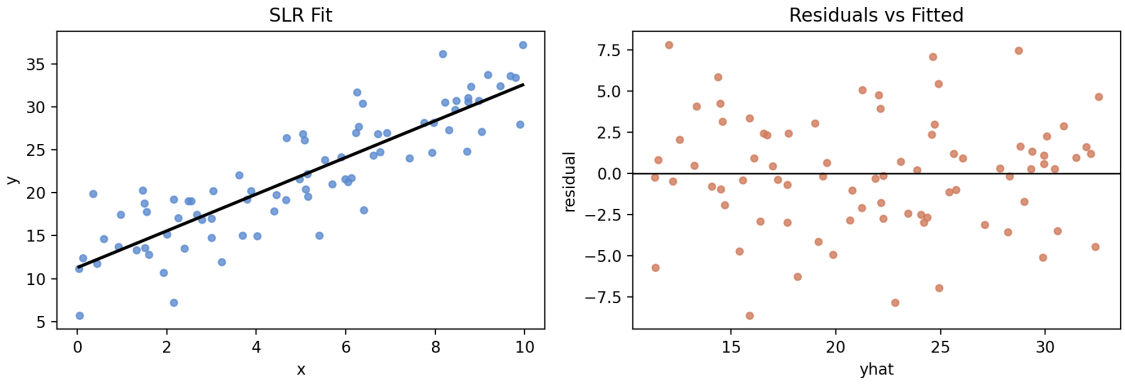
Run from the lecture folder:

```
python demo/demo.py
```

Outputs:

- `images/demo.png`
- `data/results.txt`

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

Why do we check residual plots even if R^2 is high?