Statistics and Data Analysis
Unit 04 – Lecture 01 Notes
Correlation and Regression: Concepts

Tofik Ali

February 17, 2026

Topic

Correlation vs regression concepts; causation warning; residual idea.

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

- Concepts
- Causation Warning
- Exercises
- Demo
- Summary

Learning Outcomes

- Differentiate correlation and regression
- Explain why correlation does not imply causation
- Interpret a scatter plot (trend, outliers)
- Define residual and why residuals matter

Why these outcomes matter. **Correlation** measures the strength of a linear association between two variables. It is symmetric (no X/Y direction) and does not imply causation. Outliers can inflate or hide correlation, so always look at the scatter plot. **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.

Concepts: Key Points

- Correlation measures linear association
- Regression models Y as a function of X
- Regression has roles: predictors vs response

Explanation. **Correlation** measures the strength of a linear association between two variables. It is symmetric (no X/Y direction) and does not imply causation. Outliers can inflate or hide correlation, so always look at the scatter plot. **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.

Causation Warning: Key Points

- Confounding can create misleading correlation
- Reverse causality is possible
- Causal claims need design or strong assumptions

Explanation. Always state assumptions clearly. Common assumptions in classical tests include independence of observations, roughly normal errors (or a large-sample justification), and similar variances across groups. Violations do not automatically invalidate a result, but they change how much you should trust the p-value and confidence interval. **Correlation** measures the strength of a linear association between two variables. It is symmetric (no X/Y direction) and does not imply causation. Outliers can inflate or hide correlation, so always look at the scatter plot. A **confounder** is a third variable that influences both X and Y, creating a misleading association. Example: ice-cream sales and drowning both increase in summer; temperature is the confounder. In practice, confounding is the main reason correlation is not causation.

Exercises (with Solutions)

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

Exercise 1: Pick response variable

Predict house price using size and location. What is the response variable?

Solution

- House price is the response (Y).

Exercise 2: Interpret r

If r=0.7 between study hours and score, what does it mean?

Solution

- Strong positive linear association.
- Not proof of causation.

Exercise 3: Residual sign

If y=74 and yhat=80, what is residual?

Solution

- Residual = y - yhat = -6 (over-prediction).

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.

Mini Demo (Python)

Run from the lecture folder:

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

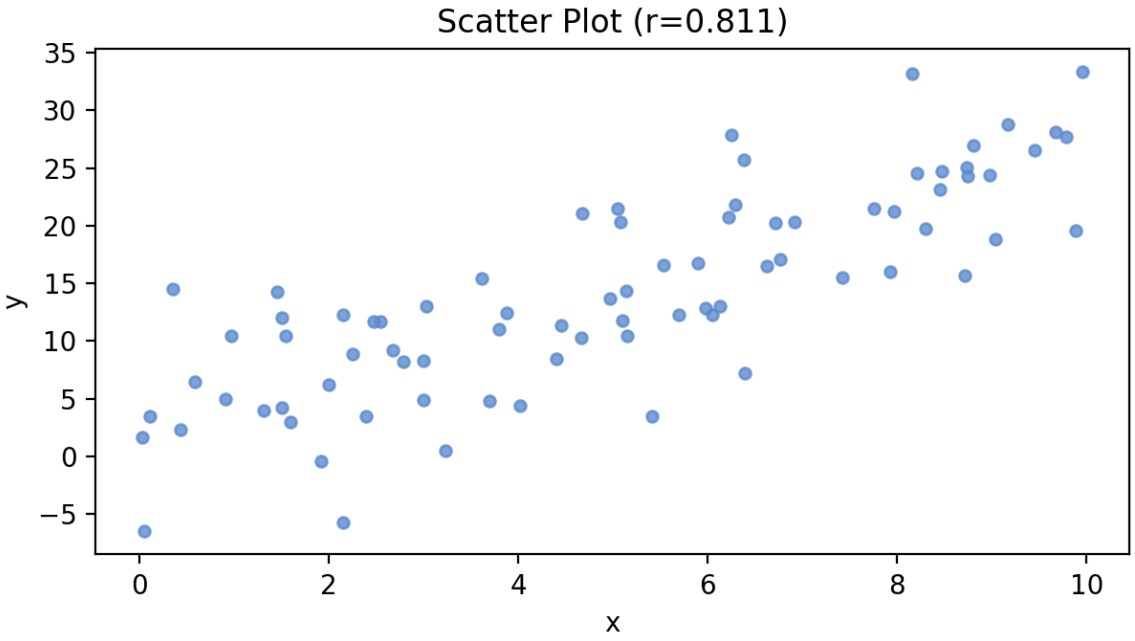
Output files:

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

What to show and say.

- Creates a noisy linear relationship and plots scatter with fitted line.
- Reports correlation and a simple regression fit to compare the two ideas.
- Use the plot to talk about outliers and why correlation is not causation.

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

Give one example of a confounder that can create a misleading correlation.

Suggested answer (for revision). A confounder can influence both variables (e.g., temperature affects ice-cream sales and swimming), creating correlation without causation.

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 Concepts Causation Warning Exercises Demo Summary

Agenda

- Overview
- Concepts
- Causation Warning
- Exercises
- Demo
- Summary

Learning Outcomes

- Differentiate correlation and regression
- Explain why correlation does not imply causation
- Interpret a scatter plot (trend, outliers)
- Define residual and why residuals matter

Concepts: Key Points

- Correlation measures linear association
- Regression models Y as a function of X
- Regression has roles: predictors vs response

Causation Warning: Key Points

- Confounding can create misleading correlation
- Reverse causality is possible
- Causal claims need design or strong assumptions

Exercise 1: Pick response variable

Predict house price using size and location. What is the response variable?

Solution 1

- House price is the response (Y).

Exercise 2: Interpret r

If r=0.7 between study hours and score, what does it mean?

Solution 2

- Strong positive linear association.
- Not proof of causation.

Exercise 3: Residual sign

If y=74 and yhat=80, what is residual?

Solution 3

- Residual = y - yhat = -6 (over-prediction).

Mini Demo (Python)

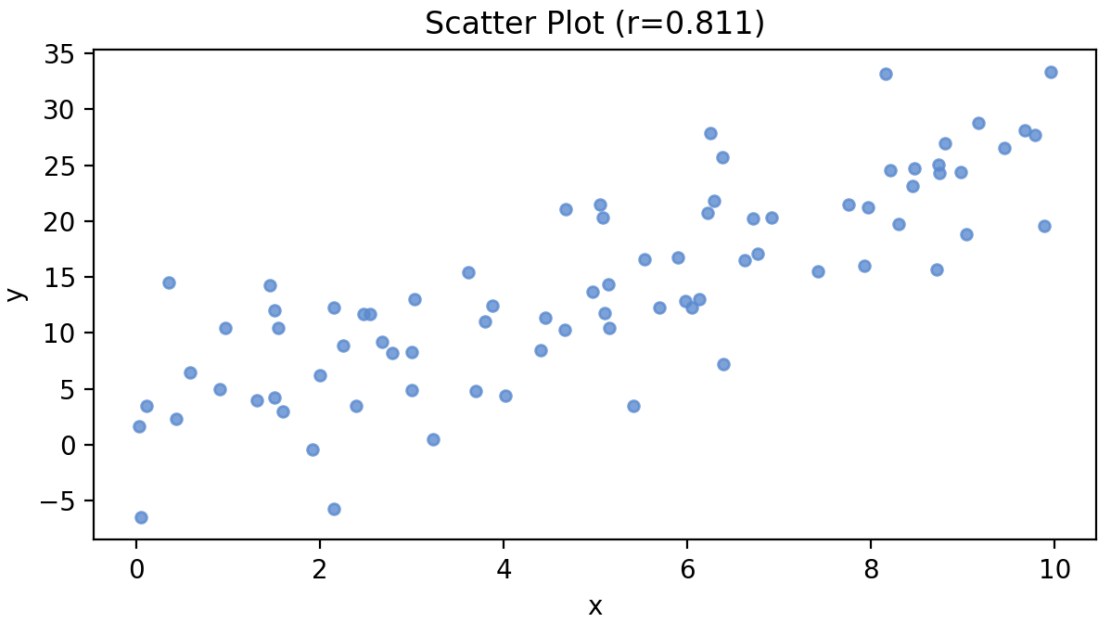
Run from the lecture folder:

```
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

Give one example of a confounder that can create a misleading correlation.