

# Statistics and Data Analysis

## Unit 04 – Lecture 05 Notes

### Multicollinearity

Tofik Ali

February 17, 2026

#### Topic

Multicollinearity: definition, symptoms, detection, and fixes.

#### 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

- What and Why
- Detection
- Exercises
- Demo
- Summary

## Learning Outcomes

- Define multicollinearity (high correlation among predictors)
- Explain why it harms interpretation (unstable coefficients)
- Recognize symptoms (large SEs, unstable signs)
- List common fixes (drop/combine/regularize)

**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. **Multicollinearity** means predictors overlap strongly (high correlation among  $X$ 's). It makes individual coefficients unstable and standard errors large, so interpretation suffers. Prediction can still be good, but explanations like 'feature  $X$  causes  $Y$ ' become unreliable.

## What and Why: Key Points

- Predictors overlap in information
- Coefficients become unstable
- Prediction may still be OK but interpretation suffers

## Detection: Key Points

- Correlation matrix/heatmap (screening)
- VIF (next)
- Condition number (advanced)

**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. **VIF** measures how much the variance of a coefficient is inflated due to multicollinearity. High VIF indicates a predictor can be explained by other predictors, so its coefficient becomes unstable.

## Exercises (with Solutions)

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

### Exercise 1: Identify

If  $\text{corr}(x_1, x_2) = 0.98$ , what risk do you expect?

#### Solution

- High multicollinearity; unstable coefficients.

**Walkthrough.** **Multicollinearity** means predictors overlap strongly (high correlation among  $X$ 's). It makes individual coefficients unstable and standard errors large, so interpretation suffers. Prediction can still be good, but explanations like 'feature  $X$  causes  $Y$ ' become unreliable.

### Exercise 2: Fix

Name one fix for multicollinearity.

#### Solution

- Drop one feature, combine features, or use ridge/PCA.

**Walkthrough.** **Multicollinearity** means predictors overlap strongly (high correlation among  $X$ 's). It makes individual coefficients unstable and standard errors large, so interpretation suffers. Prediction can still be good, but explanations like 'feature  $X$  causes  $Y$ ' become unreliable. **Ridge regression (L2)** shrinks coefficients toward zero, which reduces variance and helps with multicollinearity. It usually keeps all features but with smaller magnitudes. Always scale features before using ridge/lasso so the penalty is fair.

### Exercise 3: Prediction vs interpretation

Can multicollinearity still allow good prediction?

#### Solution

- Yes, but individual coefficients are unreliable.

**Walkthrough.** **Multicollinearity** means predictors overlap strongly (high correlation among  $X$ 's). It makes individual coefficients unstable and standard errors large, so interpretation suffers. Prediction can still be good, but explanations like 'feature  $X$  causes  $Y$ ' become unreliable.

### Mini Demo (Python)

Run from the lecture folder:

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

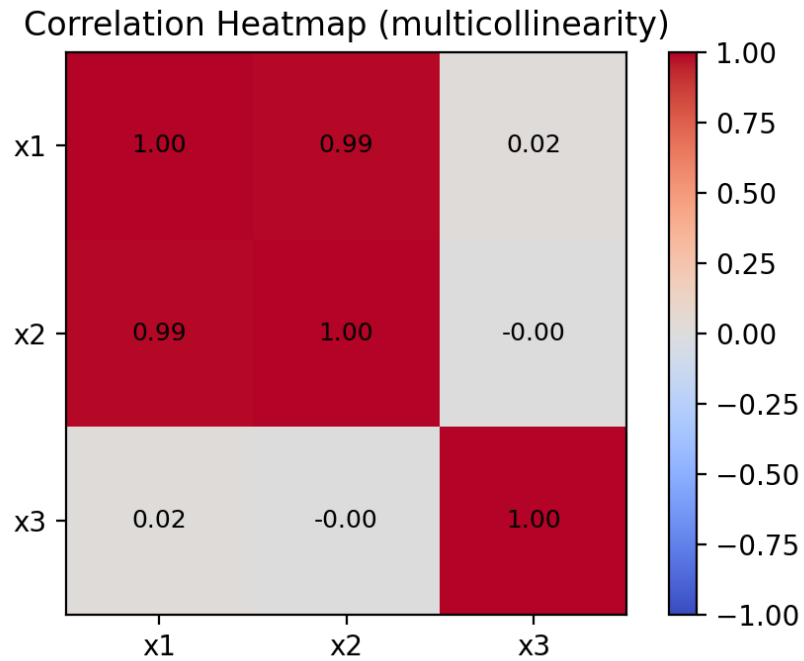
Output files:

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

### What to show and say.

- Creates correlated predictors and fits regression to show unstable coefficients.
- Reports correlation structure and coefficient variability across runs.
- Use it to motivate VIF and regularization as fixes.

### 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

What does multicollinearity break first: prediction or interpretation (and why)?

**Suggested answer (for revision).** Interpretation breaks first: coefficients become unstable and hard to trust, while prediction can remain reasonable due to redundant information.

### 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**

## Quick Links

[Overview](#) [What and Why](#) [Detection](#) [Exercises](#) [Demo](#) [Summary](#)

## Agenda

- Overview
- What and Why
- Detection
- Exercises
- Demo
- Summary

## Learning Outcomes

- Define multicollinearity (high correlation among predictors)
- Explain why it harms interpretation (unstable coefficients)
- Recognize symptoms (large SEs, unstable signs)
- List common fixes (drop/combine/regularize)

## What and Why: Key Points

- Predictors overlap in information
- Coefficients become unstable
- Prediction may still be OK but interpretation suffers

## Detection: Key Points

- Correlation matrix/heatmap (screening)
- VIF (next)
- Condition number (advanced)

## Exercise 1: Identify

If  $\text{corr}(x_1, x_2) = 0.98$ , what risk do you expect?

## Solution 1

- High multicollinearity; unstable coefficients.

## Exercise 2: Fix

Name one fix for multicollinearity.

### Solution 2

- Drop one feature, combine features, or use ridge/PCA.

## Exercise 3: Prediction vs interpretation

Can multicollinearity still allow good prediction?

### Solution 3

- Yes, but individual coefficients are unreliable.

## Mini Demo (Python)

Run from the lecture folder:

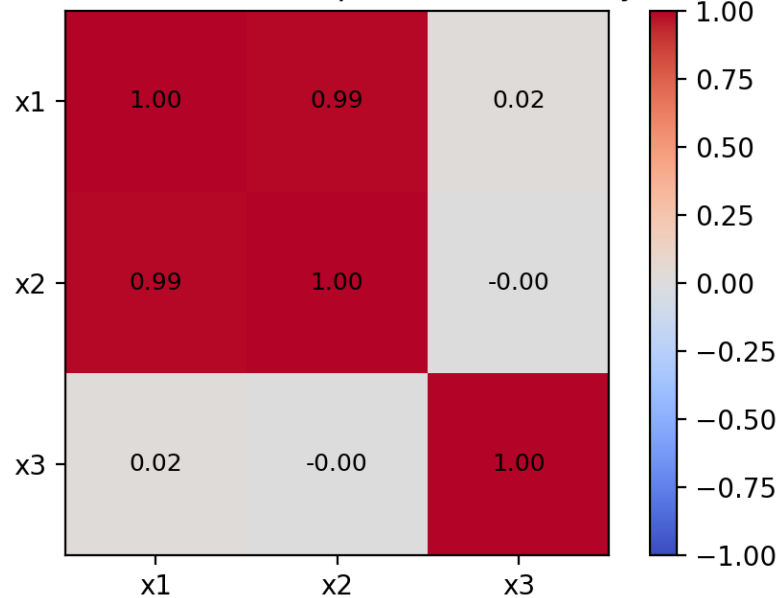
```
python demo/demo.py
```

Outputs:

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

## Demo Output (Example)

Correlation Heatmap (multicollinearity)



## Summary

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

## Exit Question

What does multicollinearity break first: prediction or interpretation (and why)?