Statistics and Data Analysis
Unit 04 – Lecture 05 Notes

Tofik Ali

February 14, 2026

Topic

Multicollinearity: definition, symptoms, detection, and fixes.

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)

Detailed Notes

These notes are designed to be read alongside the slides. They expand each slide bullet into plain-language explanations, small worked examples, and common pitfalls. When a formula appears, emphasize (1) what each symbol means, (2) the assumptions needed to use it, and (3) how to interpret the final number in the problem context.

What and Why

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

Detection

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

Exercises (with Solutions)

Exercise 1: Identify

If corr(x1,x2)=0.98, what risk do you expect?

Solution

- High multicollinearity; unstable coefficients.

Exercise 2: Fix

Name one fix for multicollinearity.

Solution

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

Exercise 3: Prediction vs interpretation

Can multicollinearity still allow good prediction?

Solution

- Yes, but individual coefficients are unreliable.

Exit Question

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

Demo (Python)

Run from the lecture folder:

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

Output files:

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

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.