Statistics and Data Analysis
Unit 04 – Lecture 09 Notes

Tofik Ali

February 14, 2026

Topic

End-to-end workflow: data -> model -> evaluation -> communication (case style).

Learning Outcomes

- Describe an end-to-end regression workflow
- Choose appropriate regression metrics (RMSE and R^2)
- Check overfitting (train vs test gap)
- Communicate results with plots (predicted vs actual, residuals)

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.

Workflow

- Define target and inputs
- Prepare data and split chronologically if needed
- Fit baseline then iterate

Evaluation

- Use RMSE/MSE/MAE and R^2
- Use plots: predicted vs actual, residuals
- Document limitations

Exercises (with Solutions)

Exercise 1: Metric choice

Target is continuous (price). Should you use accuracy?

Solution

- No; accuracy is for classification.

Exercise 2: Overfitting sign

Train RMSE=5, test RMSE=20. What does it suggest?

Solution

- Overfitting; try simpler model or regularization.

Exercise 3: Communication

Name one plot to communicate regression quality.

Solution

- Predicted vs actual scatter; residual plot.

Exit Question

What would you do first if the case study model performs poorly on the test set?

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.