Statistics and Data Analysis
Unit 04 – Lecture 09 Notes
Case Study: End-to-End Regression Workflow

Tofik Ali

February 17, 2026

Topic

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

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

- Workflow
- Evaluation
- Exercises
- Demo
- Summary

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)

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.

Workflow: Key Points

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

Evaluation: Key Points

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

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: 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.

Walkthrough. **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.

Mini Demo (Python)

Run from the lecture folder:

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

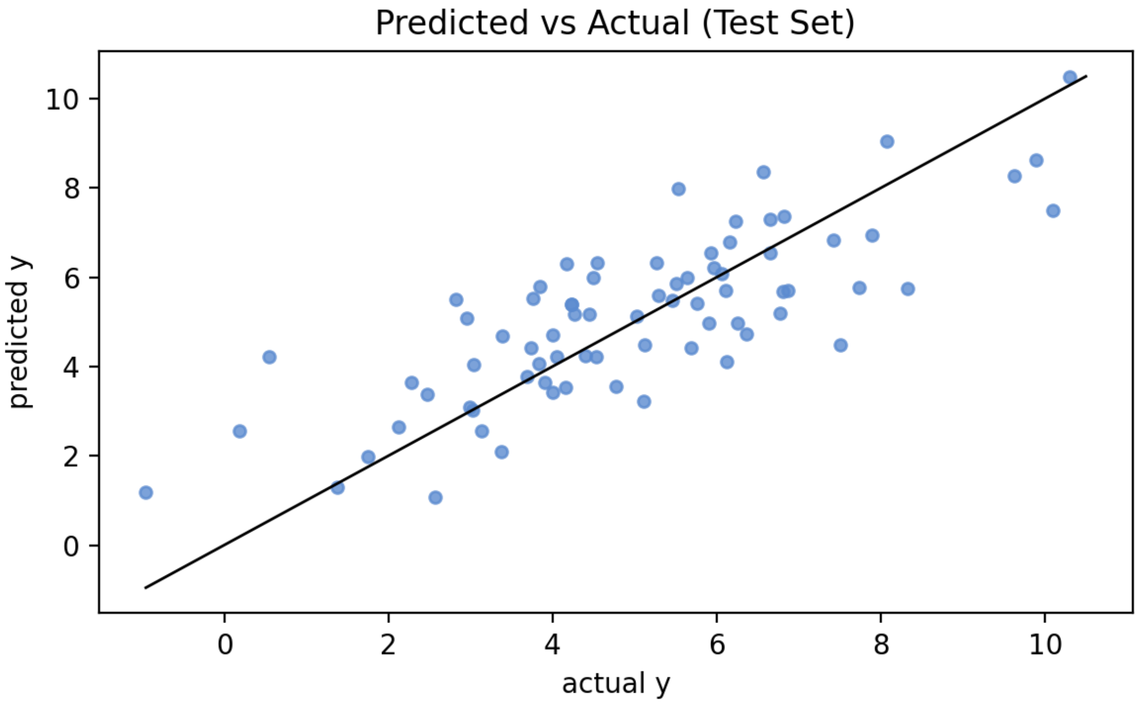
Output files:

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

What to show and say.

- Builds an end-to-end regression mini-case (split -> fit -> evaluate).
- Produces a diagnostic plot and numeric metrics to practice reporting.
- Use it to show what a 'good report' looks like (assumptions + limits).

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

Suggested answer (for revision). Start by checking data leakage and split strategy, then compare to a simple baseline and inspect residuals/feature issues before adding complexity.

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 Workflow Evaluation Exercises Demo Summary

Agenda

- Overview
- Workflow
- Evaluation
- Exercises
- Demo
- Summary

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)

Workflow: Key Points

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

Evaluation: Key Points

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

Exercise 1: Metric choice

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

Solution 1

- No; accuracy is for classification.

Exercise 2: Overfitting sign

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

Solution 2

- Overfitting; try simpler model or regularization.

Exercise 3: Communication

Name one plot to communicate regression quality.

Solution 3

- Predicted vs actual scatter; residual plot.

Mini Demo (Python)

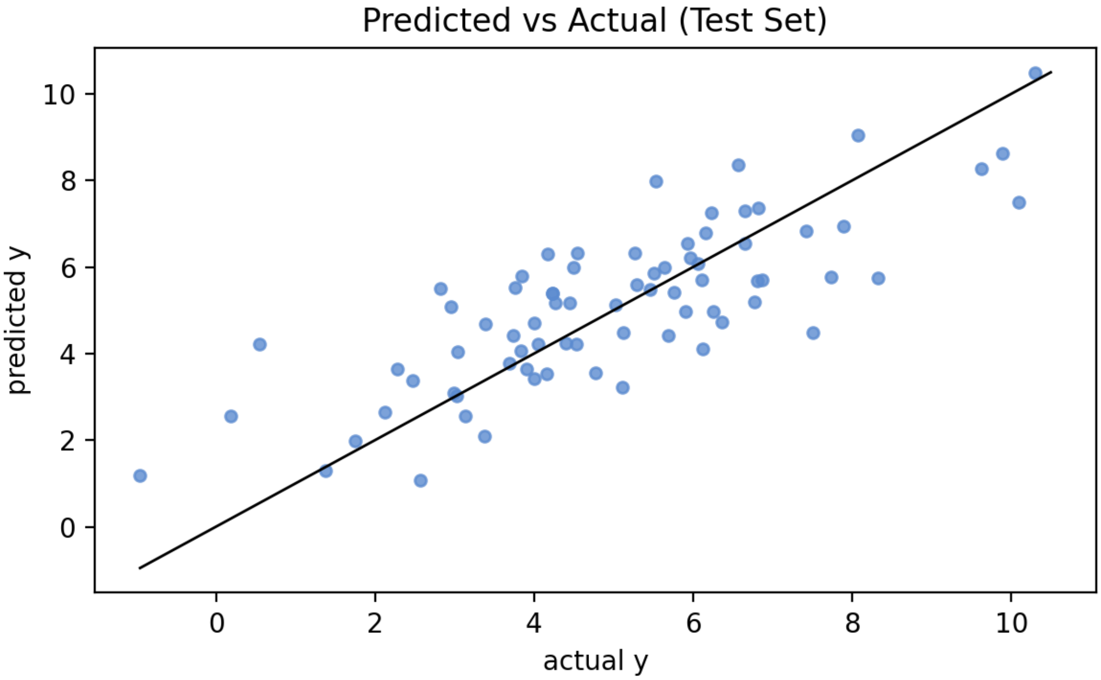
Run from the lecture folder:

```
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

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