Statistics and Data Analysis
Unit 04 – Lecture 06 Notes
VIF, AIC/BIC, Ridge and Lasso (Part 1)

Tofik Ali

February 17, 2026

Topic

VIF concept; model selection criteria; ridge and lasso regularization (intro).

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

- VIF

- Ridge/Lasso

- Exercises

- Demo

- Summary

Learning Outcomes

- Compute and interpret VIF (basic)

- Explain AIC/BIC as model selection criteria (intuition)

- Write ridge and lasso objectives

- Explain coefficient shrinkage and feature selection idea

Why these outcomes matter. **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. **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.

VIF: Key Points

- Definition: $\text{VIF}_j = 1/(1 - R_j^2)$

- Higher VIF -> more multicollinearity

- Rule of thumb thresholds (5/10)

Explanation. 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. **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.

VIF: Key Formula

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

How to read the formula. **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.

Ridge/Lasso: Key Points

- Ridge uses L2 penalty (shrinks)

- Lasso uses L1 penalty (can set some to 0)

- Scale features before regularization

Explanation. **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. **Lasso (L1)** can shrink some coefficients exactly to zero, acting like automatic feature selection. This can improve interpretability, but it can be unstable when predictors are highly correlated.

Ridge/Lasso: Key Formula

$$\min \sum (y - \hat{y})^2 + \lambda \sum \beta_j^2 \quad \text{(ridge)}$$

How to read the formula. **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. **Lasso (L1)** can shrink some coefficients exactly to zero, acting like automatic feature selection. This can improve interpretability, but it can be unstable when predictors are highly correlated.

Exercises (with Solutions)

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

Exercise 1: Compute VIF

If $R_j^2 = 0.9$, compute VIF_j.

Solution

- $\text{VIF}_j = 1/(1 - 0.9) = 10$ (high).

Walkthrough. **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.

Exercise 2: Ridge vs lasso

Which can produce exact zero coefficients?

Solution

- Lasso (L1) can set some coefficients to 0.

Walkthrough. 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. **Lasso (L1)** can shrink some coefficients exactly to zero, acting like automatic feature selection. This can improve interpretability, but it can be unstable when predictors are highly correlated.

Exercise 3: AIC/BIC meaning

Lower AIC/BIC means what (conceptually)?

Solution

- Better trade-off between fit and complexity (relative).

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.

- Computes VIF-like behavior and compares ridge vs lasso shrinkage on correlated features.

- Shows how regularization stabilizes coefficients and can improve generalization.

- Use results to connect penalty strength λ to bias-variance trade-off.

Demo Output (Example)
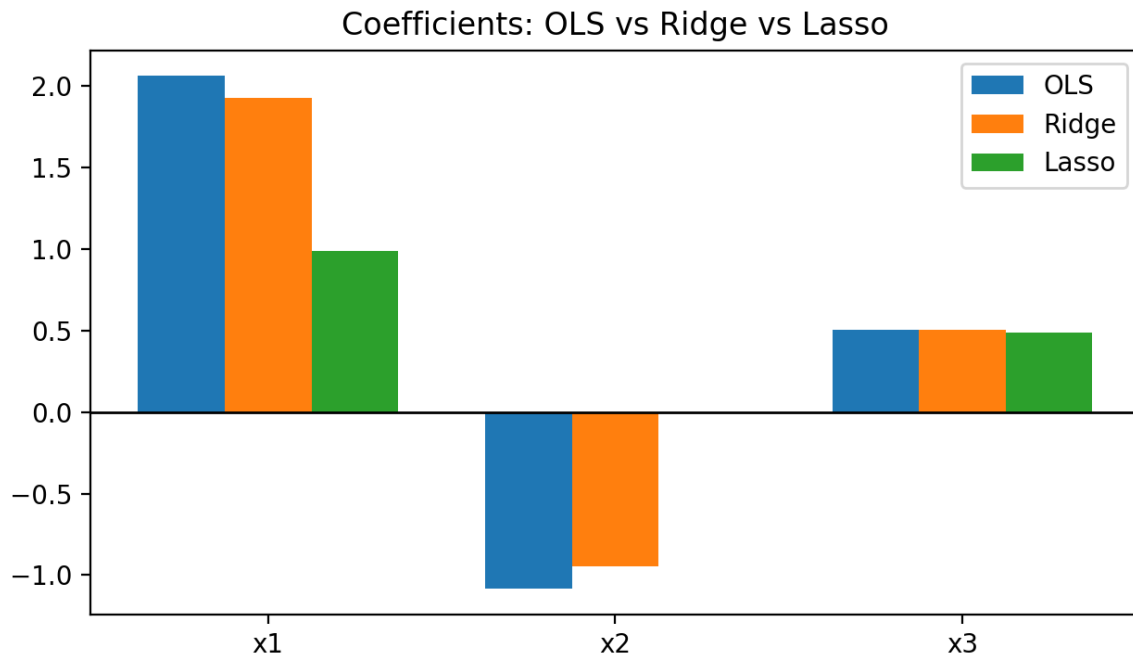
Summary

- Key definitions and the main formula.

- How to interpret results in context.

- How the demo connects to the theory.

Exit Question

Why can ridge help when predictors are highly correlated?

Suggested answer (for revision). Ridge shrinks and stabilizes coefficients, reducing variance caused by correlated predictors and improving numerical stability.

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

Agenda

- Overview
- VIF
- Ridge/Lasso
- Exercises
- Demo
- Summary

Learning Outcomes

- Compute and interpret VIF (basic)
- Explain AIC/BIC as model selection criteria (intuition)
- Write ridge and lasso objectives
- Explain coefficient shrinkage and feature selection idea

VIF: Key Points

- Definition: $\text{VIF}_j = 1/(1 - R_j^2)$
- Higher VIF -> more multicollinearity
- Rule of thumb thresholds (5/10)

VIF: Key Formula

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Ridge/Lasso: Key Points

- Ridge uses L2 penalty (shrinks)
- Lasso uses L1 penalty (can set some to 0)
- Scale features before regularization

Ridge/Lasso: Key Formula

$$\min \sum (y - \hat{y})^2 + \lambda \sum \beta_j^2 \quad \text{(ridge)}$$

Exercise 1: Compute VIF

If $R_j^2 = 0.9$, compute VIF_j.

Solution 1

- $VIF_j = 1/(1 - 0.9) = 10$ (high).

Exercise 2: Ridge vs lasso

Which can produce exact zero coefficients?

Solution 2

- Lasso (L1) can set some coefficients to 0.

Exercise 3: AIC/BIC meaning

Lower AIC/BIC means what (conceptually)?

Solution 3

- Better trade-off between fit and complexity (relative).
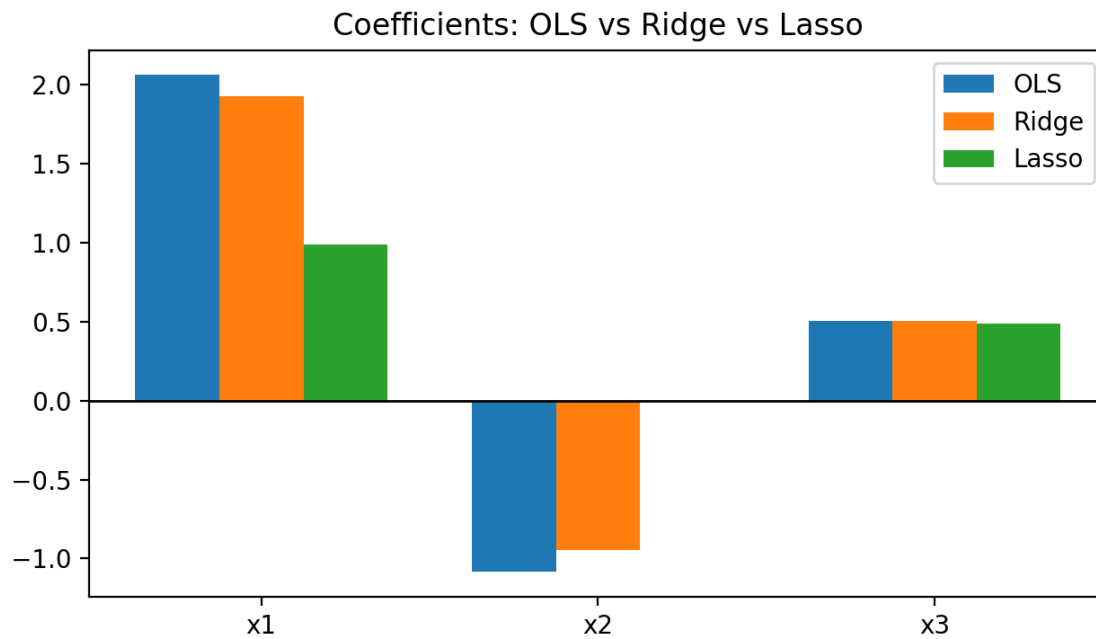
Mini Demo (Python)

Run from the lecture folder:

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

Outputs:

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

Demo Output (Example)

Coefficients: OLS vs Ridge vs Lasso

Summary

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

Exit Question

Why can ridge help when predictors are highly correlated?