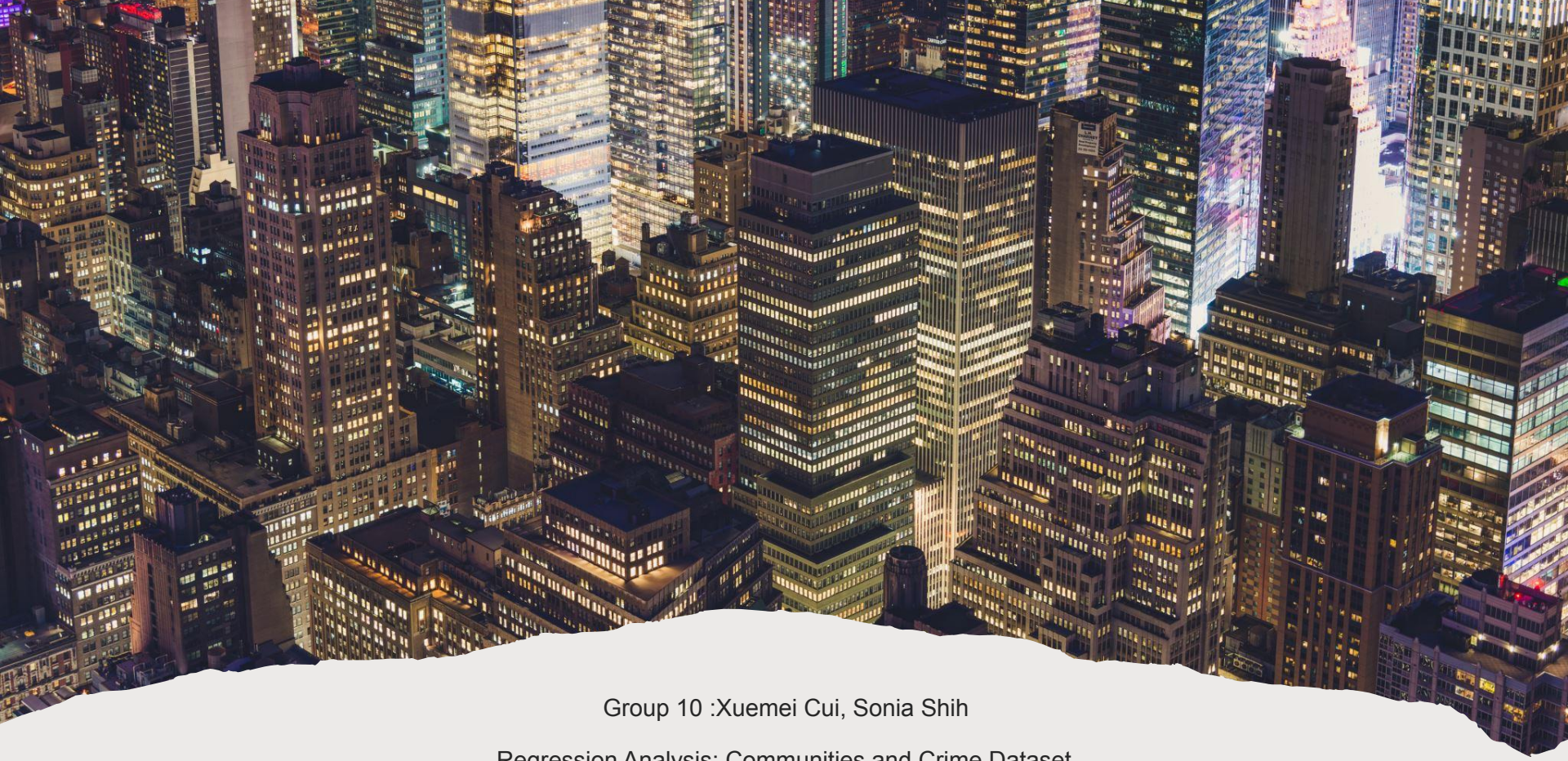


Group 10 :Xuemei Cui, Sonia Shih

Regression Analysis: Communities and Crime Dataset

Communities and Crime Dataset

Source: UCI Machine Learning Repository

After data cleaning:

n = 2215 (observations)

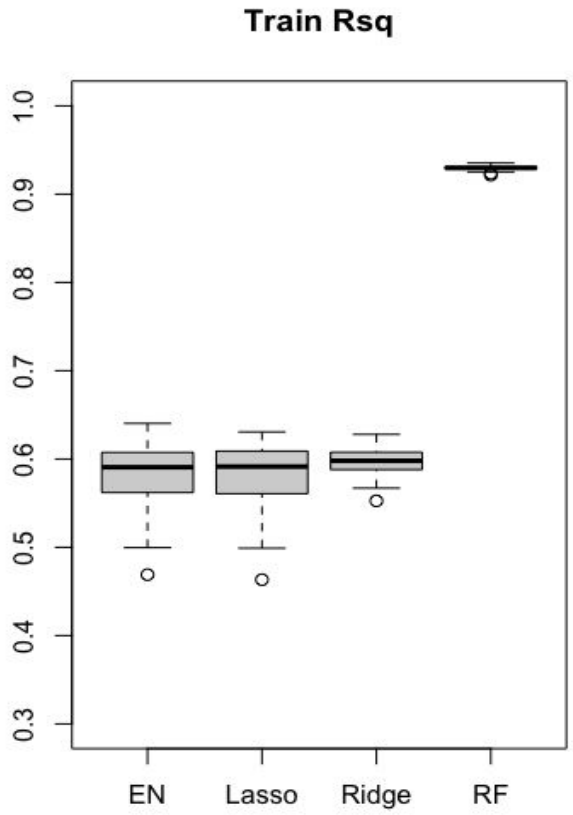
p = 98 (numerical features)

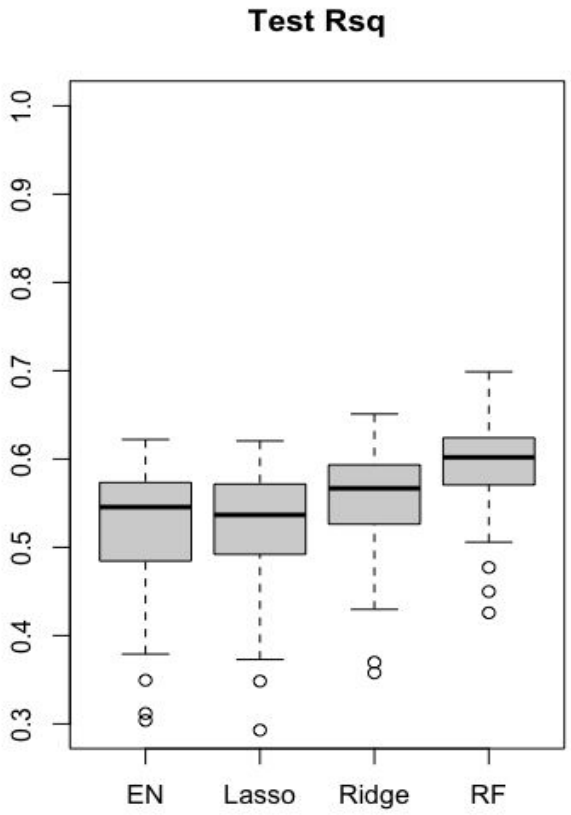
Each observation: a local community in the United States

Features: population for the community, average people per household, median household income, etc.

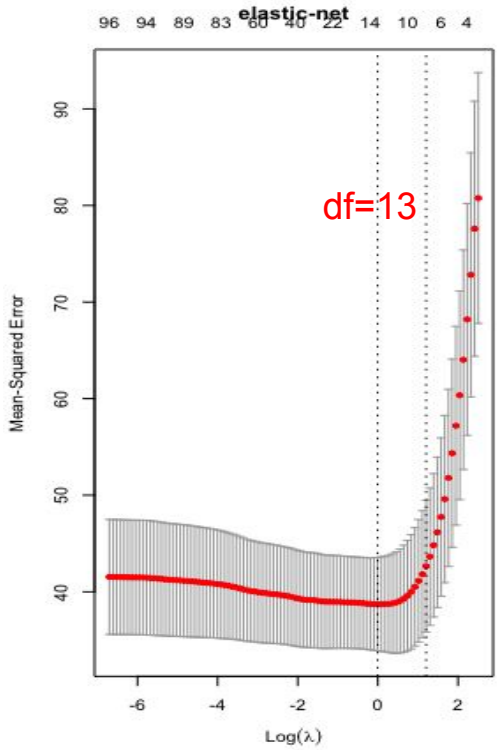
Response variable y: number of murders per 100K population (range: 0 ~ 91.09)

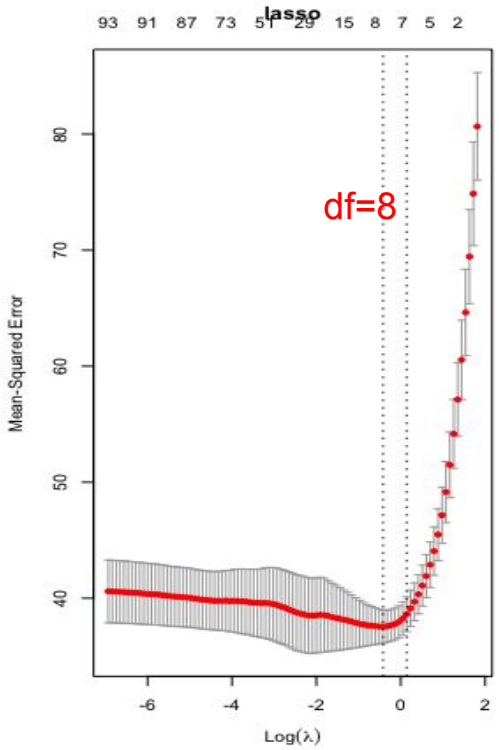
SIDE-BY-SIDE BOXPLOTS OF $\mathbf{R}^2$ (ON 100 SAMPLES)

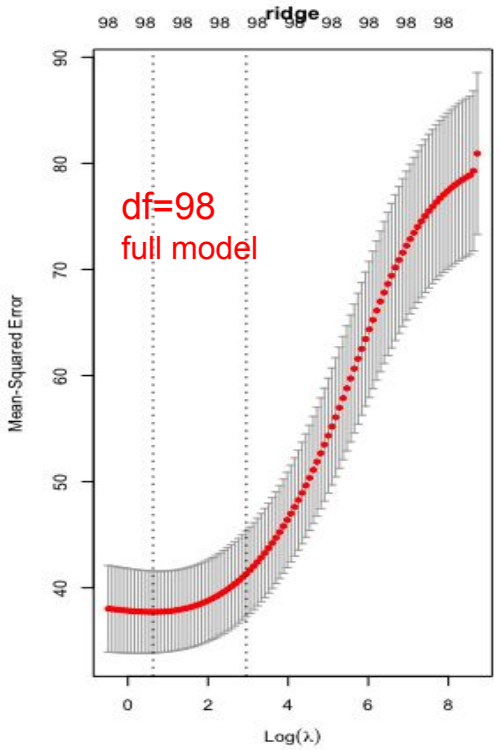




10-FOLD CV CURVES (ON ONE SAMPLE) & AVERAGE TIME





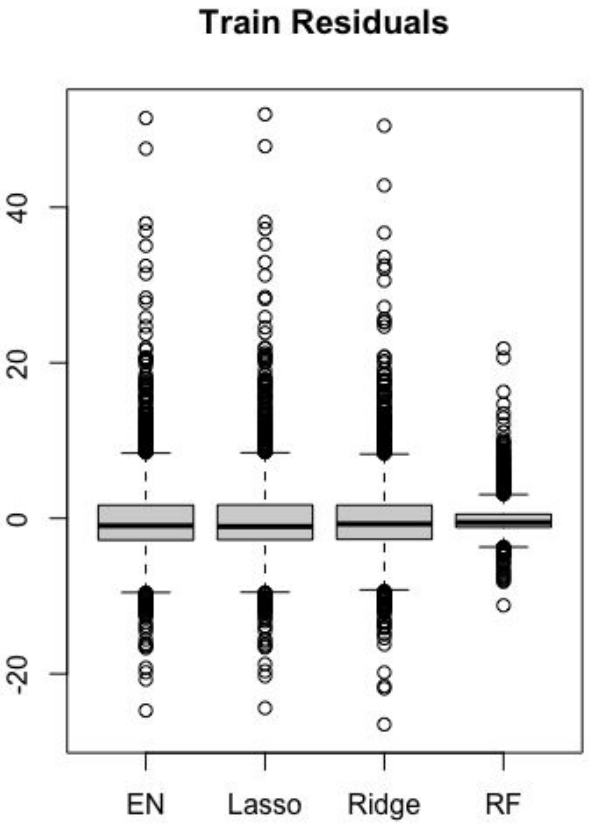


Elastic-Net: 0.74s

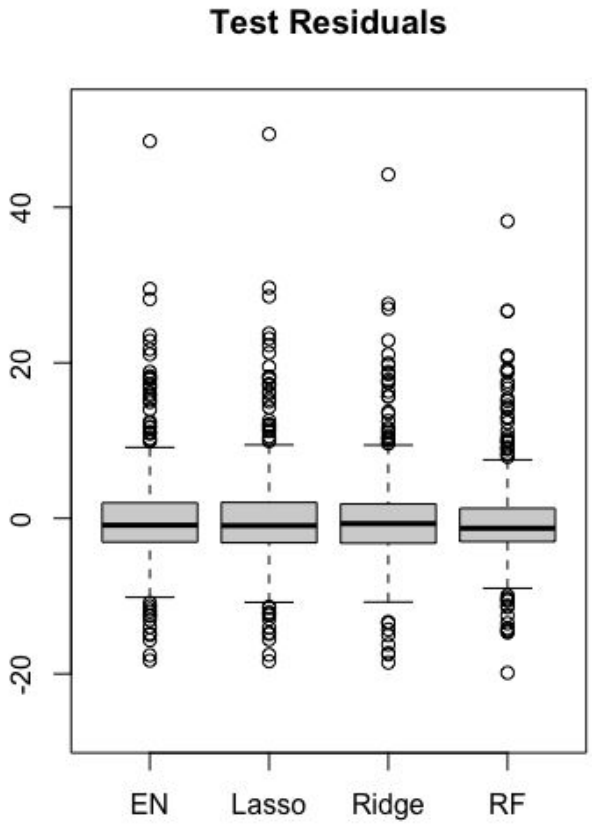
Lasso: 0.76s

Ridge: 0.67s

SIDE-BY-SIDE BOXPLOTS OF RESIDUALS (ON ONE SAMPLE)

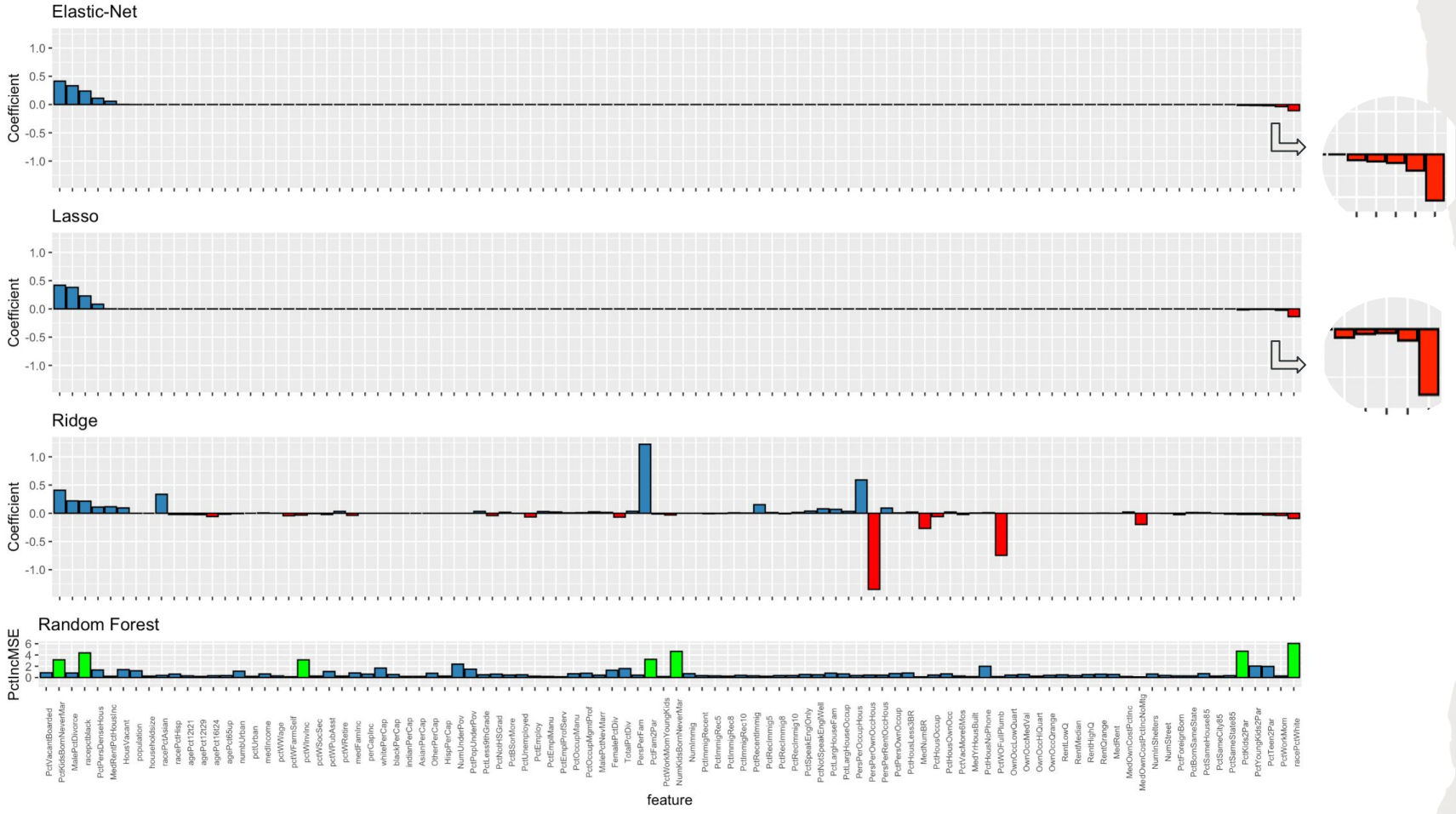


n=1772



n=443

COEFFICIENT & IMPORTANCE PLOT



CI & TIME

Method	90% CI for test R^2	Median	Time to fit All data (sec)
Elastic-Net	(0.3438580 0.6112454)	0.5438	1.458
Lasso	(0.3210890 0.6042081)	0.5349	1.610
Ridge	(0.4579196 0.6211772)	0.5627	1.628
Random Forest	(0.5295845 0.6519314)	0.5965	35.979

CONCLUSION

Trade-off between time and performance? Yes

- Random Forest performed the best → took longer time
- **Decided to choose Ridge:** performed similar to Random Forest, relatively high R^2 and took less time

- All 4 models selected 4 variables to predict Murder per 100K populaion

  + PctKidsBornNeverMar:percentage of kids born to never married women (nonmarital childbearing)
  + racepctblack:percentage of population that is african american

  - racePctWhite percentage of population that is caucasian
  - PctKids2Par:percentage of kids in family housing with two parents