



Galaxies' Redshift Estimates

My Vien

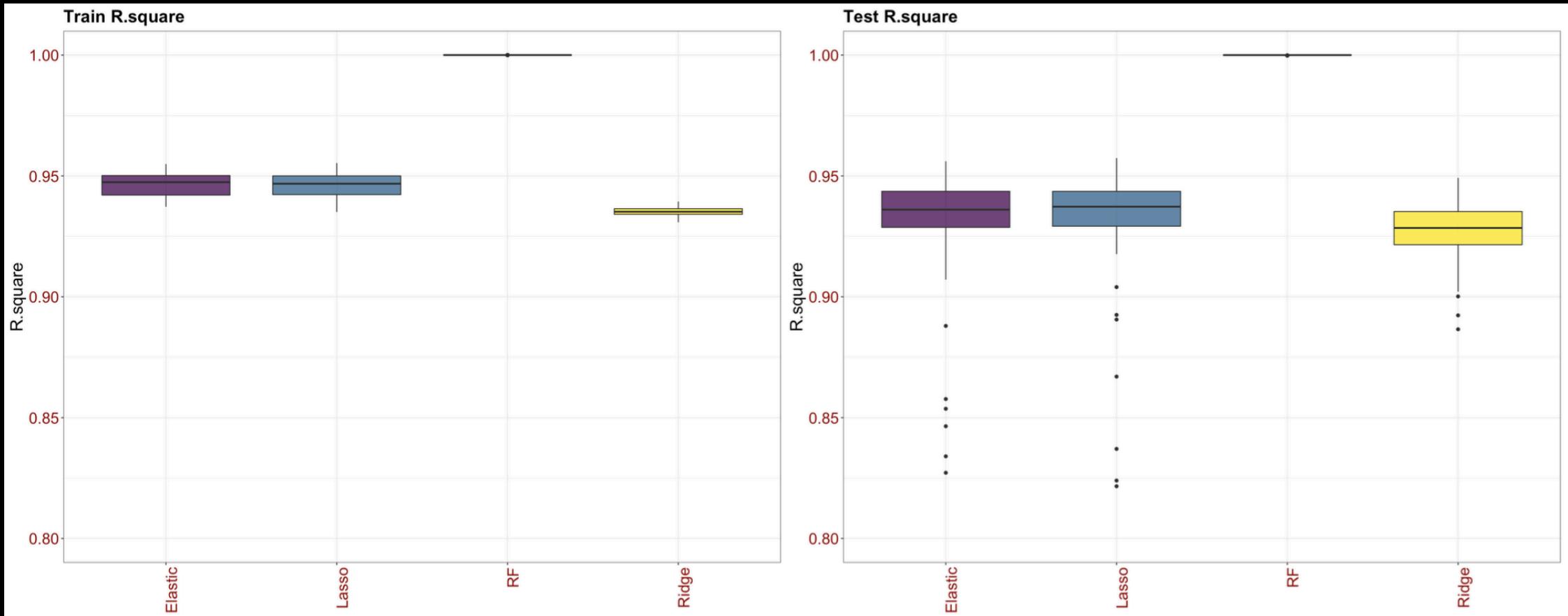
IMAGE SOURCE: [HTTPS://WWW.ESO.ORG/PUBLIC/IMAGES/POTW1830A/](https://www.eso.org/public/images/potw1830a/)

INTRODUCTION AND DATASET

- The target variable is photometric estimates of redshift, a measure of distance from us to an object.
- The predictors variables are mainly the measurements and errors of brightness of color bands of galaxies.
- $n = 3438$, number of galaxies samples (after removing missing values)
- $p = 60$ (number of predictors)
- We will be able to predict the velocity of a galaxy based on the brightness of color bands of a galaxy



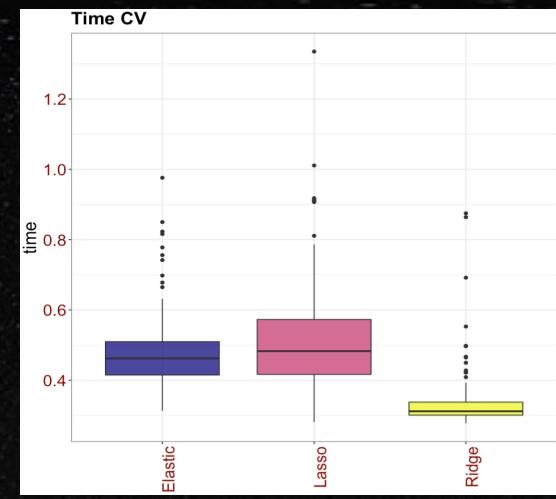
R-SQUARES



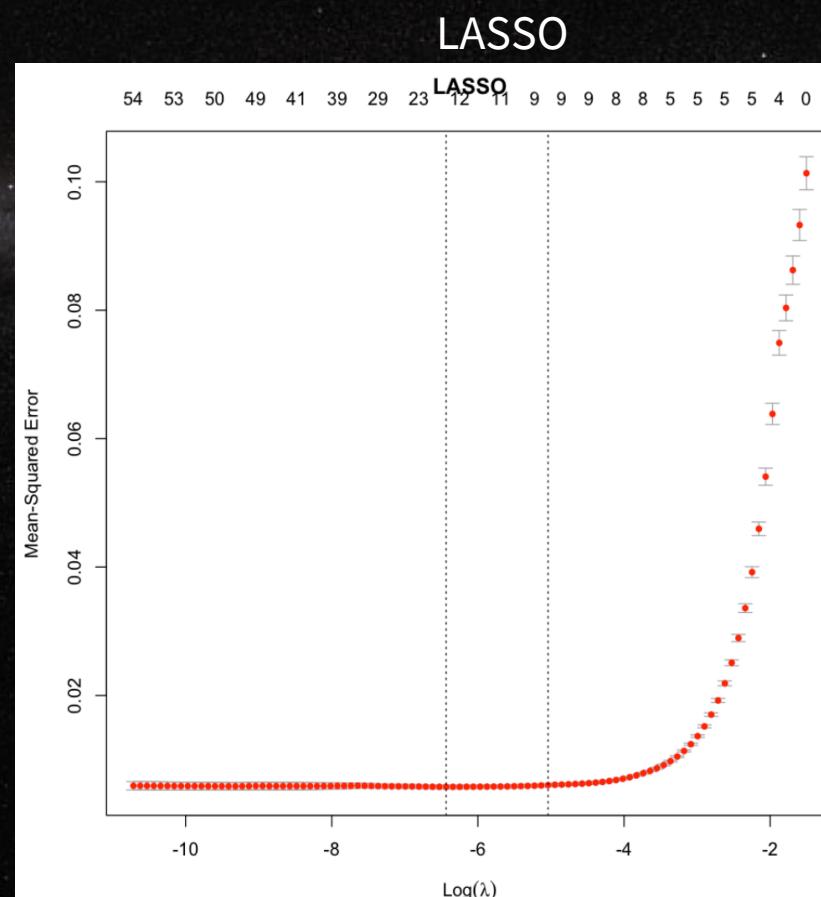
- Random Forest has the best train R-square and test R-square, while Ridge is the lowest on both train and test set.
- Test R-square values are lower and more spread on all models

CV CURVE

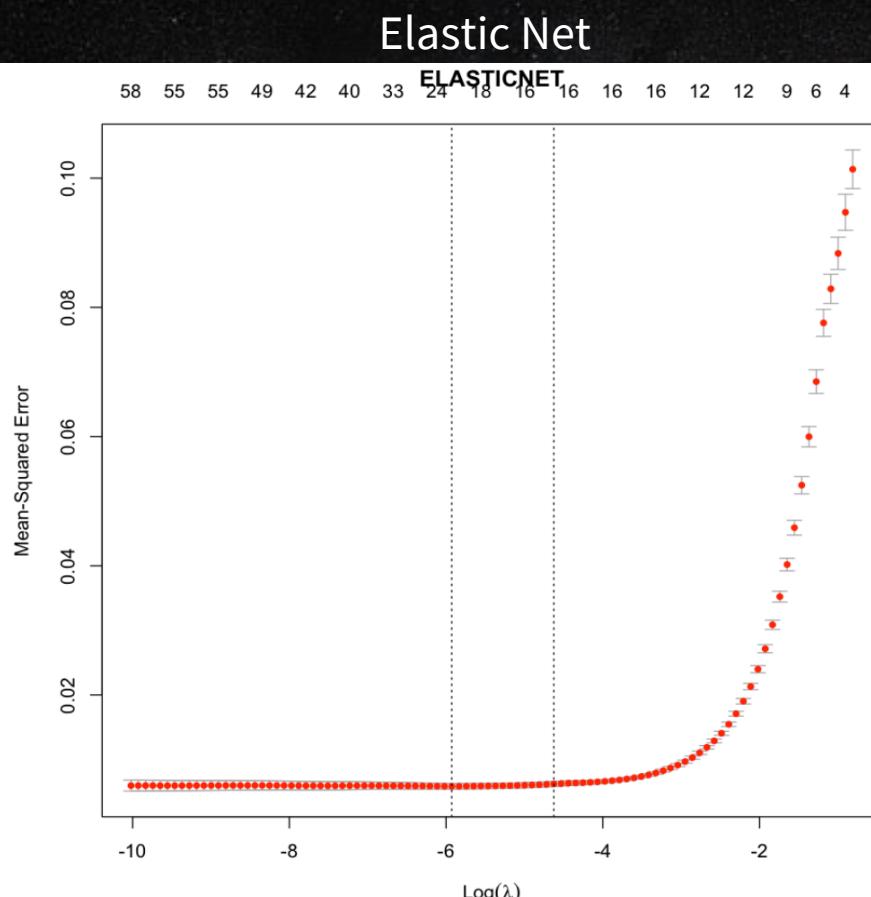
- In overall, the variance of CV MSE does not look so different between all 3 models
- Optimized Lasso has around 12-23 non-zero coefficients
- Optimized ElasticNet has around 23 non-zero coefficients
- As expected, Optimized Ridge has maintained all predictors coefficients



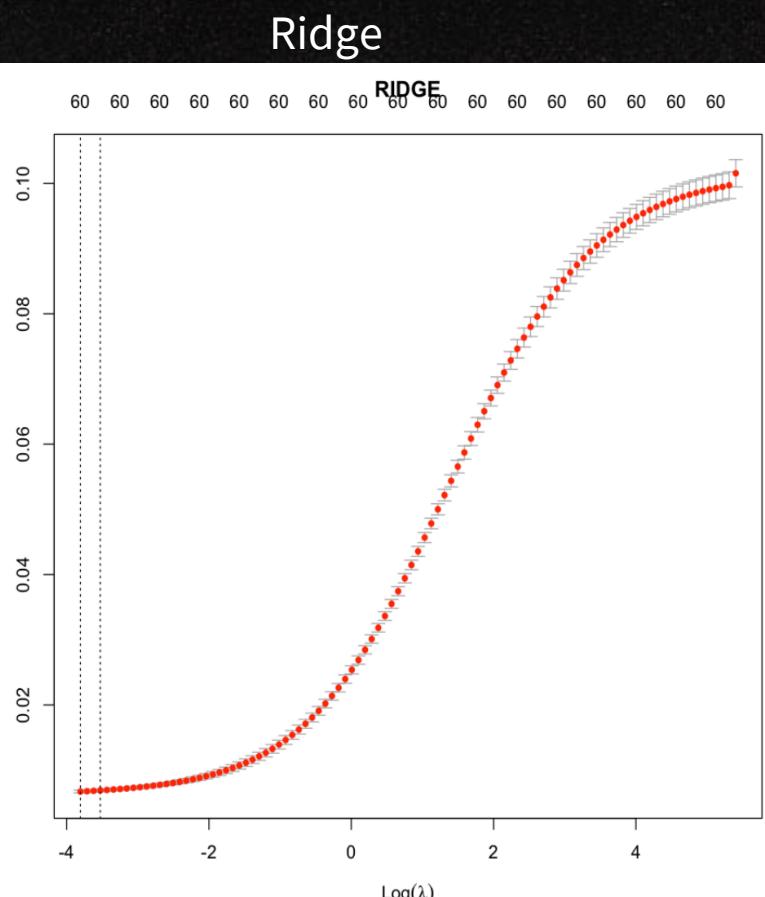
LASSO



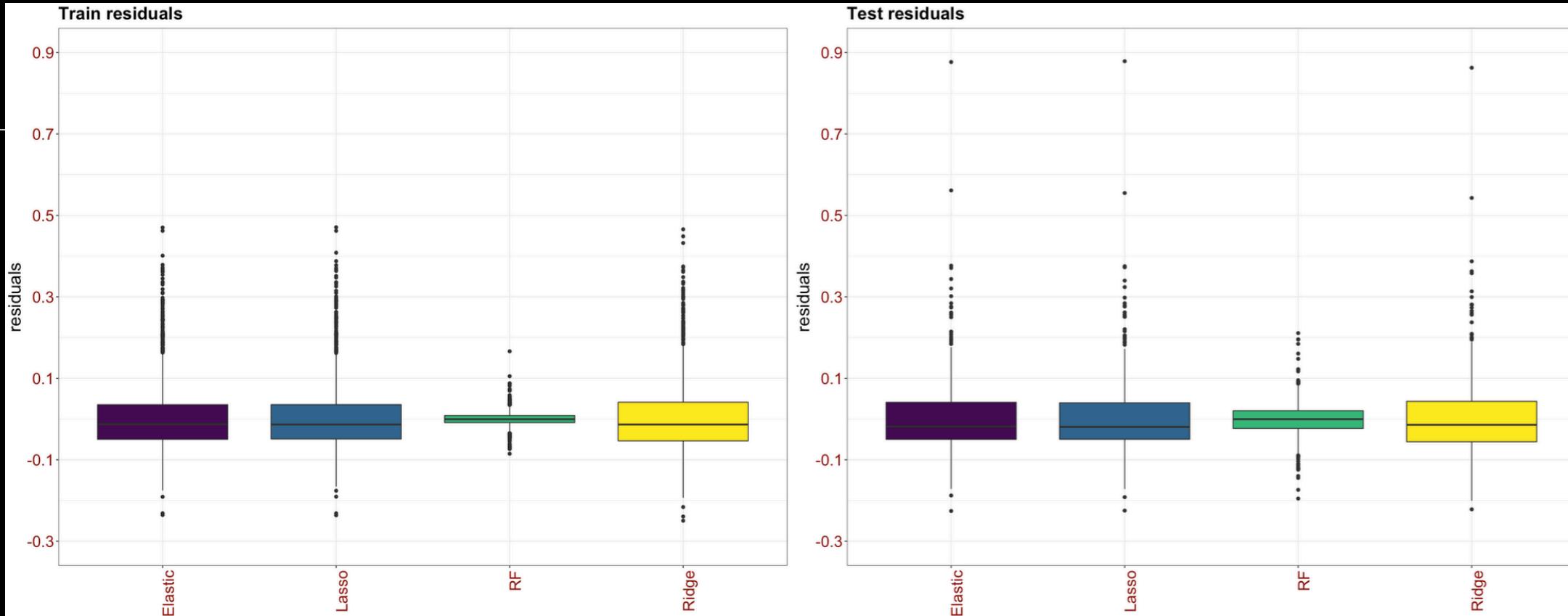
Elastic Net



Ridge



RESIDUALS



- Train size is 2750 and test size is 688.
- Random Forest has smallest spread while all constrained models have larger spread.
- Extreme residual outliers are detected on test set of all 3 constrained models.

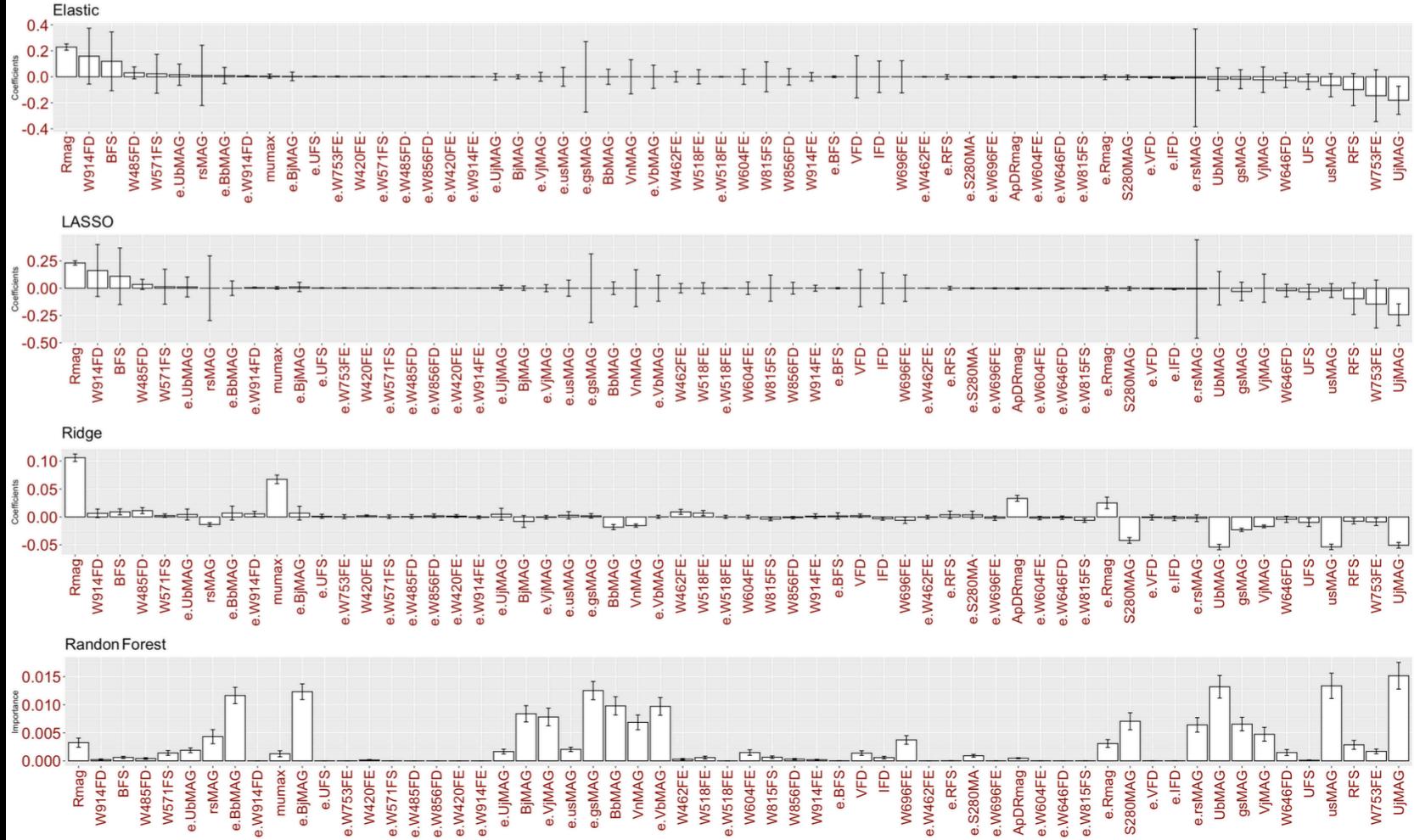
Performance

- There is a trade off between time and model performance comparing between 3 constrained linear models and RF.
- Elastic Net is considered the best amongst linear constrained models considering its test R2 and time.
- Lasso takes longest time to run but does not perform better than Lasso.
- Ridge is fastest, but its score is less comparable.

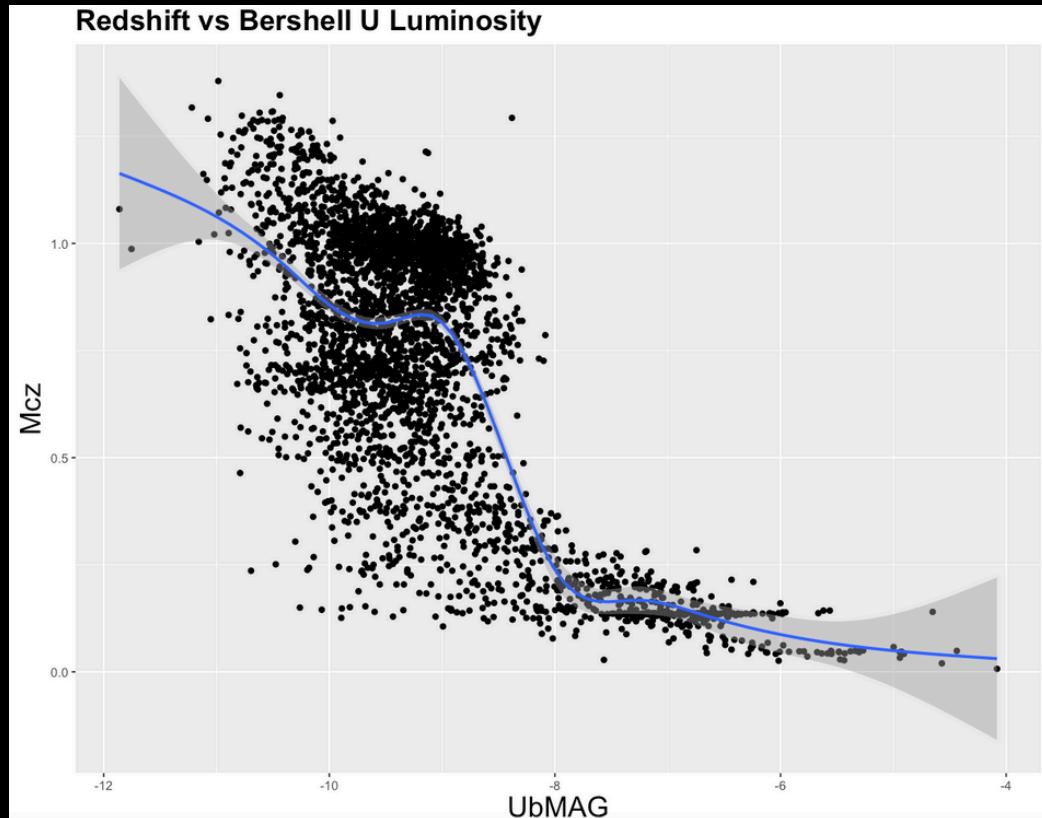
Models	90% test R2	Time (s)
Lasso	(0.92043, 0.94697)	0.683
ElasticNet	(0.92049, 0.94722)	0.549
Ridge	(0.91169, 0.93955)	0.551
RF	(0.999941, 0.999999)	44.316

Coefficients (100 Bootstrap Samples)

- All linear models have similar feature importance: UjMAG (Johnson U luminosity) and Rmag (red band magnitude) are the two largest coefficients in 3 linear models.
- With Random Forest, it is UjMAG (Johnson U luminosity) and UbMAG (Bessell U luminosity)
- Lasso has the largest coefficients in all linear models and Ridge has the smallest.
- All linear models have more spread in features importance than Random Forest.
- Some features in Elastic Net and Lasso have very wide standard deviation.



Conclusion:



- There is not a big difference between train and test R-squared, train and test residuals which proves that all our predictive models have good performance.
- Taking time and test R-squared into account, elastic net has the best performance between 3 linear models.
- Even though RF takes longest time to run, but RF is still the best model considering its 90% test R-squared.
- The features importance between RF and linear models are different, there might be underlying non-linear processes so that RF can achieve significantly better score than the linear models. The Bersell U luminosity could be the non-linear feature.