Why Penalized Regression Should Be Used More To Advance Our Research

Presenter: Merrill Levitt

Background

- Penalized regression techniques, which add a little bias to the model have been shown to create more parsimonious models than ordinary least squares (OLS) regression[1]
- Common types of penalized regression:
  - Ridge regression
  - LASSO
  - Elastic net
- Instances when penalized regression techniques perform particularly well[2]:
  - High variability in predictors
  - High collinearity between predictors
  - Large number of predictors
  - Variable selection is needed
- Research Question: At what point do penalized regression techniques perform better than OLS regression when looking at predictor variability, collinearity, number of cases, and number of predictors?

Method

- Nine simulations were run in R
- Response variable was a random linear combination of some or all variables and error.
- Data frames randomly generated with a range of parameters:
  - Number of predictors, *p* (2-10)
  - Collinearity between predictors, *r* (.1-.8)
  - Number of cases, *n* (20-100)
  - Variance of predictors, σ^2 (1-25)

Conclusion

The adoption of penalized regression into researcher and practitioner toolkits will aid in parameter estimation and the creation of better fitting models.

Table 1. Performance of Competing Regression Models as evidenced by R^2

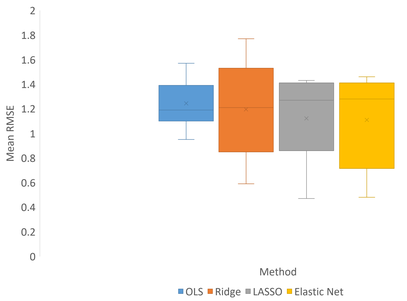
	OLS	Ridge	LASSO	Elastic Net
Sim 1	.89	.99	.99	.99
Sim 2	.99	.94	.95	.95
Sim 3	.94	.93	.95	.95
Sim 4	.90	.97	.88	.94
Sim 5	.74	.60	.77	.77
Sim 6	.60	.52	.75	.74
Sim 7	.87	.87	.83	.83
Sim 8	.88	.92	.93	.93
Sim 9	.99	.99	.99	.99

Note: Sim 1: $p = 2$, $n = 20$, $r = .1$, $\sigma^2 = 1\text{-}3$; Sim 2: $p = 2$, $n = 20$, $r = .3$, $\sigma^2 = 1\text{-}3$; Sim 3: $p = 2$, $n = 20$, $r = .5$, $\sigma^2 = 1\text{-}3$; Sim 4: $p = 2$, $n = 20$, $r = .8$, $\sigma^2 = 1\text{-}3$; Sim 5: $p = 5$, $n = 50$, $r = .3$, $\sigma^2 = 3$; Sim 6: $p = 5$, $n = 50$, $r = .3$, $\sigma^2 = 7$; Sim 7: $p = 5$, $n = 50$, $r = .3$, $\sigma^2 = 15$; Sim 8: $p = 5$, $n = 50$, $r = .5$, $\sigma^2 = 23$; Sim 9: $p = 10$, $n = 100$, $r = .5$, $\sigma^2 = 20\text{-}25$. Where *p* represents the number of parameters (IVs), *n* represents the number of cases, *r* represents the collinearity between variables, σ^2 is the variable variance.

Results

- Penalized models generally obtained greater or equal R-squared values in comparison to OLS solutions.

Figure 1. Box plots of mean-squared prediction error across simulations



Discussion

- Penalized regression techniques won't always outperform OLS models.
- However, the levels of the following characteristics necessary to make penalized regression techniques perform better are modest and variable.
  - The number of predictors
  - The amount of variance within predictors
- Traditional rules of thumbs for when to run penalized regression are insufficient
- While Sim 9 should have been the worst case for OLS, it wasn't. Suggesting the random linear model (which used only 5 of the 10 variables), may not have been complex enough.

Practical Contribution

- Penalized regressions should always be run in addition to you OLS regression analyses

References

1. *Fox, J. (2016). Applied regression analysis & generalized linear models (3rd ed.). Thousand Oaks, CA: Sage.*
2. Hastie, Trevor, Tibshirani, S., & Friedman, H. (2009). *The elements of statistical learning data mining, inference, and prediction* (2nd ed.). Springer.

Corresponding Author can be reached at levit110@umn.edu

UNIVERSITY OF MINNESOTA