Shrinkage methods

Outline

- Ridge regression
- The lasso
- Elastic net

Ridge regression

Ridge regression optimising function

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

If lambda = 0, ridge regression becomes OLS. When lambda is very large, ridge coefficients are *shrunk toward 0* (but not exactly 0).

Different lambda values will result in different sets of coefficients => using validation to decide the lambda value that minimises test error.

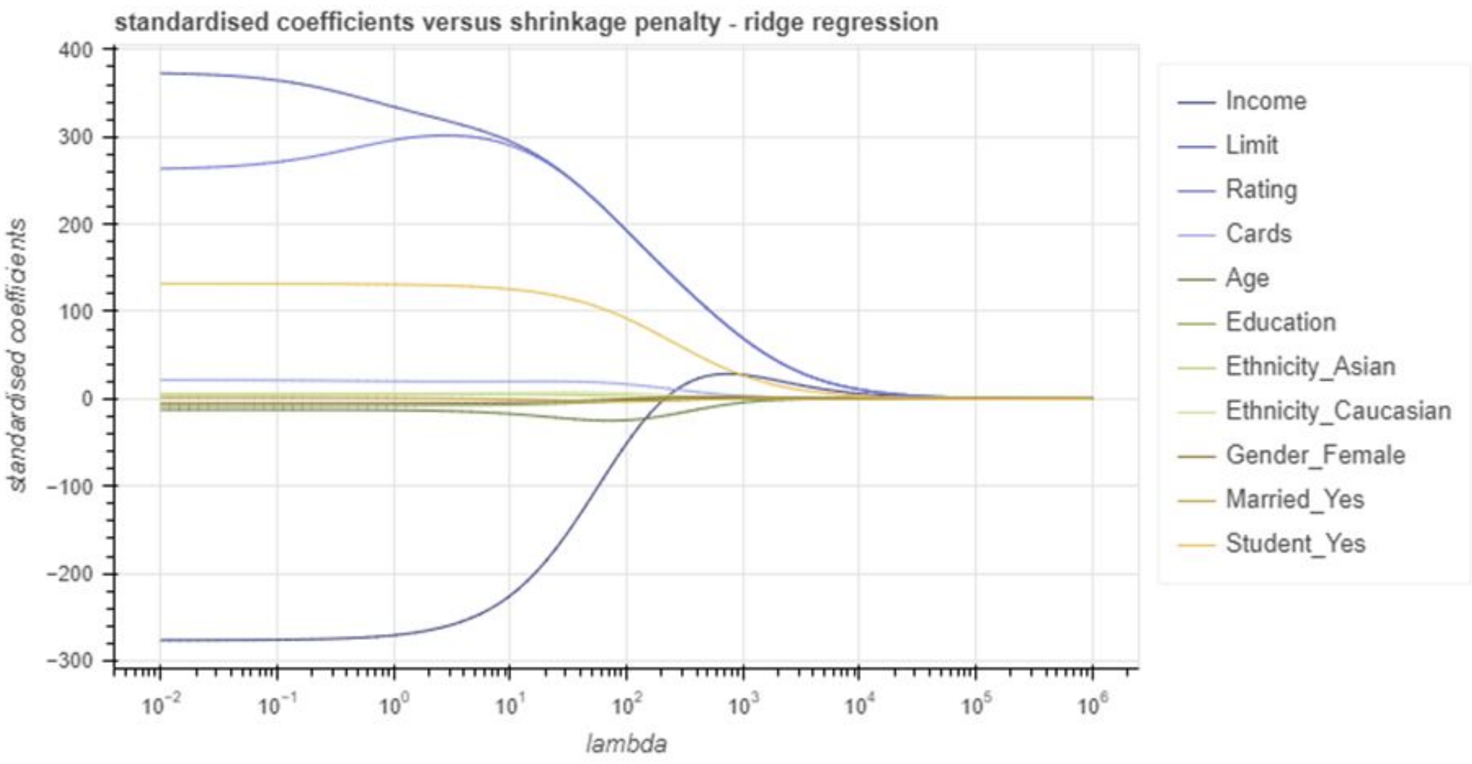
Ridge regression adds bias to the estimation of betas via lambda => reduce variation & improve predictive performance.

Ridge regression

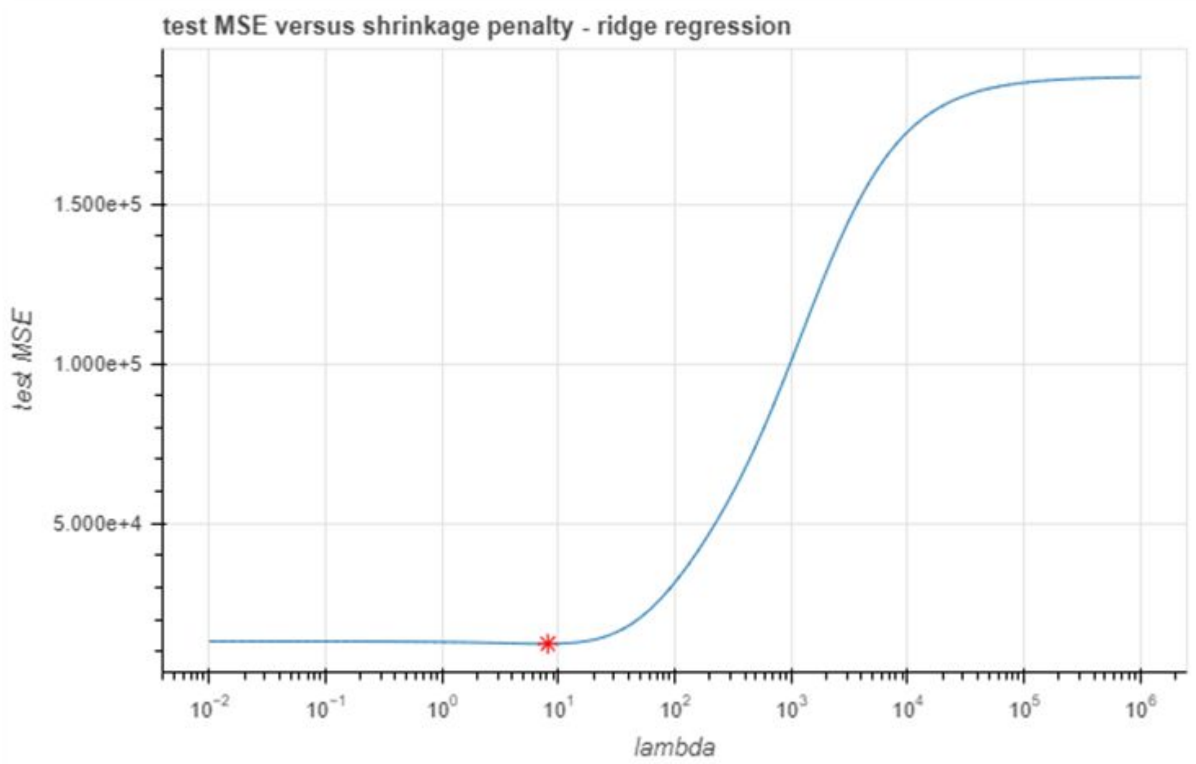
It is best to standardise the predictors before ridge regression

- OLS regression coefficients are *scale equivariant*. Ridge regression coefficients are not.
- Ridge regression coefficients are shrunk toward zero and toward each other.
- The shrinking between coefficients not on the same scale would be unequal.
- Standardisation brings predictors to the same scale => allow us to rank relative importance of the predictors. More important predictors have higher standardised coefficients.

Ridge regression results - The Credit dataset



Ridge regression results - The Credit dataset



The Lasso

The Lasso optimising function

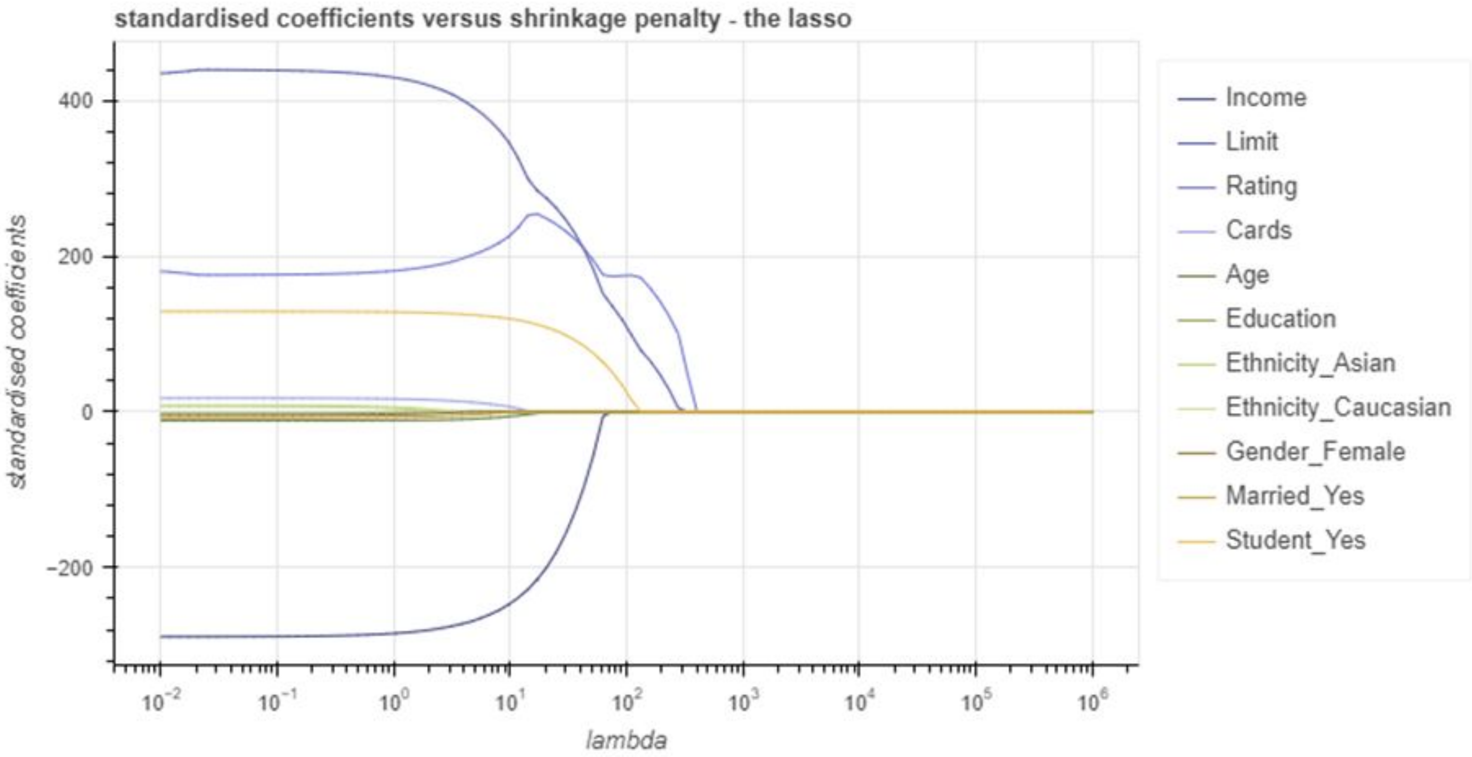
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

Unlike ridge regression, the shrinkage penalty in the lasso can force some regression coefficient estimates *to be exactly 0*.

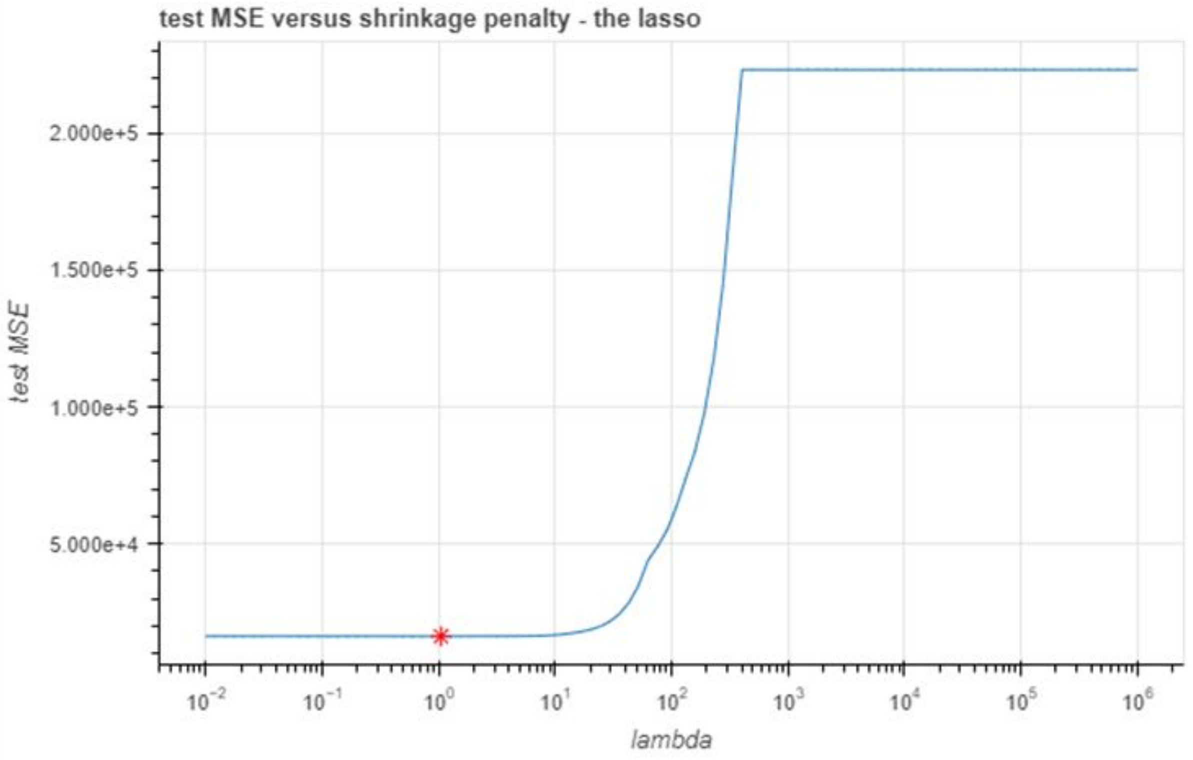
=> The lasso effectively performs variable selection, results in simpler models & improves model interpretability (like subset selection methods but *faster*!)

Like ridge regression, the lasso adds bias to the estimation of betas via lambda => reduce variation & improve predictive performance.

Ridge regression results - The Credit dataset



Ridge regression results - The Credit dataset



Ridge regression vs the lasso - estimating optimum lambda

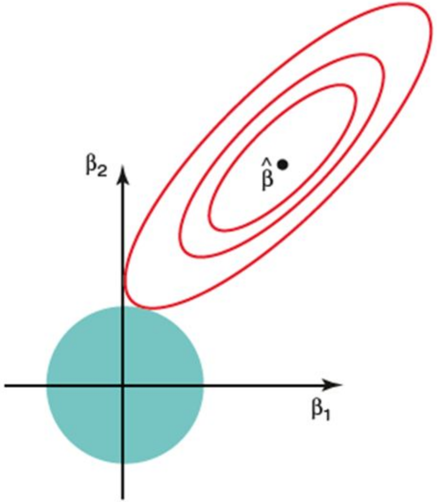
Estimating lambda that results in smallest test error for ridge regression and the lasso using the Credit dataset example.

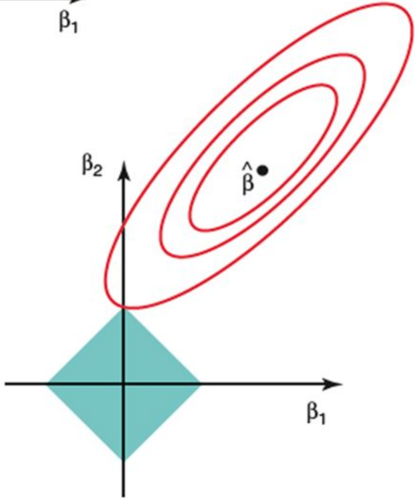
[See the attached notebook for sample Python codes and the results]

Another formulation for ridge regression and the lasso

Ridge regression

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$





The lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

How are ridge regression and the lasso better than OLS?

OLS has no means to control bias (and variance) => unique set of coefficients.

Ridge regression and the lasso allows for the introduction of bias into the estimation of coefficients (via lamba).

Higher lambda => higher bias but lower variance => better generalisation => better predictive performance.

Ridge and the lasso can handle cases where $p > n$.

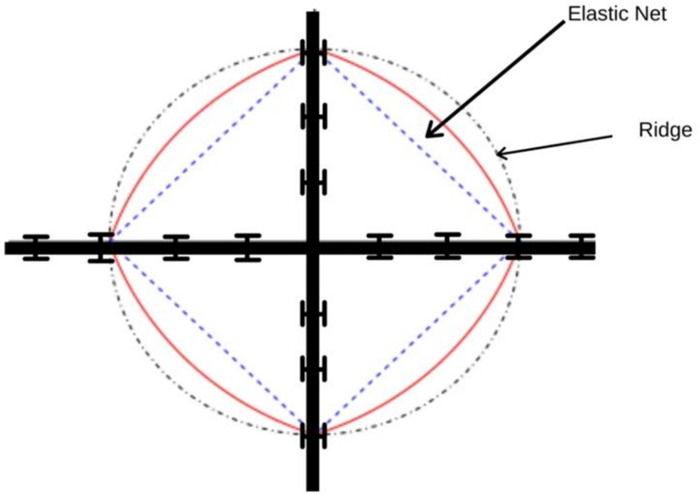
Multicollinearity:

- Ridge regression: coefficients of correlated predictors are equal.
- The lasso: coefficient of some of the correlated predictors become 0.

Elastic net

- Ridge regression - better when the predictors have relatively equal importance
- The lasso - better when some predictors are dominant over the others.
- But we don't know any of those => best way is to combine them!
- Elastic net - the shrinkage penalty in the optimising function replaced by

$$\lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right)$$



Exercise

Perform ridge regression and the lasso with the Boston dataset, using crime rate (column ‘crim’) as the response and others as the predictors.