Statistical Learning and Data Mining

Module 8: Linear Methods for Regression I

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Content structure

1. Statistical foundations.
2. **Regression methods**.
3. Classification methods.
4. Generalised linear models.
5. Nonlinear models.
6. Tree-based methods.
7. Unsupervised learning.

Module 7: Linear Methods for Regression I

1. Introduction

2. Variable selection

3. Regularisation methods

4. Discussion

Introduction

Linear Methods for Regression

In this module we focus again on the linear regression model for prediction. We move beyond OLS to consider other estimation methods.

The motivation for studying these methods is that using many predictors in a linear regression model typically leads to overfitting. We will therefore accept some bias in order reduce variance.

Linear regression (review)

Consider the additive error model

$$Y = f(\boldsymbol{x}) + \varepsilon.$$

The linear regression model is a special case based on a regression function of the form

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

OLS (review)

In the OLS method, we select the coefficient values that minimise the residual sum of squares

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ols}} = \underset{\boldsymbol{\beta}}{\mathsf{argmin}} \; \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

We obtain the formula

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ols}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

In the framework of Module 7, we view OLS as empirical risk minimisation.

MLR model (review)

1. Linearity: if $X = \boldsymbol{x}$, then

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

   for some population parameters $\beta_0, \beta_1, \ldots, \beta_p$ and a random error ε.

2. The conditional mean of ε given X is zero, $E(\varepsilon|X) = 0$.
3. Constant error variance: Var$(\varepsilon|X) = \sigma^2$.
4. Independence: the observations are independent.
5. The distribution of $X_1, \ldots, X_p$ is arbitrary.
6. There is no perfect multicollinearity (no column of $\boldsymbol{X}$ is a linear combination of other columns).

OLS properties (review)

Under Assumptions 1 (the regression function is correctly specified) and 2 (there are no omitted variables that are correlated with the predictors), the OLS estimator is unbiased

$$E(\widehat{\boldsymbol{\beta}}_{\mathsf{ols}}) = \boldsymbol{\beta}.$$

OLS properties (review)

Assumptions 1–4 allow us to show that

$$\mathsf{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{\left(1 - \mathsf{R}_j^2\right)\left(\sum_{i=1}^N (x_{ji} - \overline{x}_j)^2\right)},$$

where R_j^2 is the r-squared of a regression of predictor j on all other predictors.

Additional predictors will tend to increase R_j^2 for almost all predictors, leading to high variance for the coefficient and consequently predictions.

Why we are not satisfied with OLS?

Prediction accuracy. Low bias (if the linearity assumption is approximately correct), but potentially high variance. We can improve performance by setting some coefficients to zero or shrinking them.

Interpretability. A regression estimated with too many predictors and high variance is hard or impossible to interpret. In order to understand the big picture, we are willing to sacrifice some of the small details.

Linear model selection and regularisation

Variable selection. Identify a subset of $k < p$ predictors to use. Estimate the model by using OLS on the reduced set of variables.

Regularisation (shrinkage). Fit a model involving all the p predictors, but shrink the coefficients towards zero relative to OLS. Depending on the type of shrinkage, some estimated coefficients may be zero, in which case the method also performs variable selection.

Dimension reduction. Construct a set of $m < p$ predictors which are are linear combinations of the original predictors. Fit the model by OLS on these new predictors.

Data: Equity Premium Prediction

Quarterly data from Goyal and Welch (2008), updated to 2015.

```
Response: quarterly S&P 500 returns minus treasury bill rate

Predictors (lagged by one quarter):
1. dp        Dividend to price ratio
2. dy        Dividend yield
3. ep        Earnings per share
4. bm        Book-to-market ratio
5. ntis      Net equity expansion
6. tbl       Treasury bill rate
7. ltr       Long term rate of return on US bods
8. tms       Term spread
9. dfy       Default yield spread
10.dfr       Default return spread
11.infl      Inflation
12.ik        Investment to capital ratio

Number of observations: 275 (1947-2015)
```

Variable selection

Best subset selection (key concept)

The **best subset selection** method estimates all possible models and selects the best one according to a model selection criterion (AIC, BIC, or cross validation).

Given p predictors, there are 2^p possible models to choose from.

Best subset selection

For example, if $p = 3$ we would estimate $2^3 = 8$ models:

$$\begin{aligned}
k = 0: \quad Y &= \beta_0 + \varepsilon \\
k = 1: \quad Y &= \beta_0 + \beta_1 x_1 + \varepsilon \\
Y &= \beta_0 + \beta_2 x_2 + \varepsilon \\
Y &= \beta_0 + \beta_3 x_3 + \varepsilon \\
k = 2: \quad Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\
Y &= \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon \\
Y &= \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \\
k = 3: \quad Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon
\end{aligned}$$

Best subset selection

Algorithm Best subset selection

```
Estimate the null model $\mathcal{M}_0$, which contains only the constant.
for $k = 1, 2, \ldots, p$ do
    Fit all $\binom{p}{k}$ possible models with exactly $k$ predictors.
    Pick the model with the lowest RSS and call it $\mathcal{M}_k$.
end for
Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ according to cross validation, AIC, or BIC.
```

Computational considerations

The best subset method suffers from a problem of **combinatorial explosion**, since it requires the estimation of 2^p different models. The computational requirement is therefore very high, except in low dimensions.

For example, for $p = 30$ we would need to fit a little over 1 billion models! Best subset selection has a very high computational cost and is infeasible in practice for p larger than around 40.

Stepwise selection

Stepwise selection methods are a family of search algorithms that find promising subsets by sequentially adding or removing regressors, dramatically reducing the computational cost compared to estimating all possible specifications.

Conceptually, they are an approximation to best subset selection, not different methods.

Forward selection

Algorithm Forward selection

```
Estimate the null model $\mathcal{M}_0$, which contains only the constant.
for $k = 1, 2, \ldots, p$ do
    Fit all the $p - k + 1$ models that add one predictor to $\mathcal{M}_{k-1}$.
    Choose the best of $p - k + 1$ models in terms of RSS and call it $\mathcal{M}_k$.
end for
Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ according to cross validation, AIC, or BIC.
```

Backward selection

Algorithm Backward selection

```
Estimate the full model $\mathcal{M}_p$ by OLS.
for $k = p - 1, \ldots, 1, 0$ do
    Fit all the $k + 1$ models that delete one predictor from $\mathcal{M}_{k+1}$.
    Choose the best of the $k + 1$ models in terms of RSS and call it $\mathcal{M}_k$.
end for
Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ according to cross-validation, AIC, or BIC.
```

Stepwise selection

- Compared to best subset selection, the forward and backward stepwise algorithms reduce the number of estimations from 2^p to $1 + p(p+1)/2$. For example, for $p = 30$ the number of fitted models is 466.

- The disadvantage is that the final model selected by stepwise selection is not guaranteed to optimise any selection criterion among the 2^p possible models.

Variable selection

Advantages

- Accuracy relative to OLS. It tends to lead to better predictions compared to estimating a model with all predictors.
- Interpretability. The final model is a linear regression model based a reduced set of predictors.

Disadvantages

- Computational cost.
- By making binary decisions include or exclude particular variables, variable selection may exhibit higher variance than regularisation and dimension reduction approaches.

Illustration: Equity Premium Prediction

We select the following models in the equity premium dataset based on the AIC:

Best subset selection: (dy, bm, tms, dfr)

Forward selection: (ik, tms, dfr)

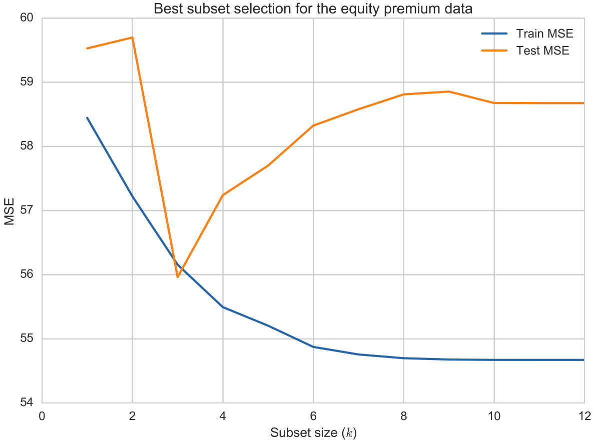
Backward selection: (dy, tms, dfr)

Illustration: Equity Premium Prediction

Table 1: Equity Premium Prediction Results

	Train R^2	Test R^2
OLS	0.108	0.014
Best Subset	0.095	0.038
Forward	0.083	0.042
Backward	0.084	0.060
CSRs	0.078	0.039

Illustration: Equity Premium Prediction



Wrong ways to do variable selection

Adjusted $\mathbf{R}^2$. The adjusted R^2 has no justification as a model selection criterion. It does not sufficiently penalise additional predictors.

Removing statistically insignificant predictors. A statistically significant coefficient means we can reliably say that it is not *exactly* zero. This has almost nothing to do with prediction (see next slide). Furthermore, there are multiple testing issues.

Illustration: Equity Premium Prediction

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ret   R-squared:                       0.108
Model:                            OLS   Adj. R-squared:                  0.051
Method:                 Least Squares   F-statistic:                     1.901
Date:                                   Prob (F-statistic):             0.0421
Time:                                   Log-Likelihood:                -629.21
No. Observations:                 184   AIC:                             1282.
Df Residuals:                     172   BIC:                             1321.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     26.1369     14.287      1.829      0.069        -2.064    54.337
dp             0.3280      8.247      0.040      0.968       -15.951    16.607
dy             3.3442      7.941      0.421      0.674       -12.330    19.019
ep             0.3133      2.345      0.134      0.894        -4.315     4.942
bm            -3.2443      6.719     -0.483      0.630       -16.507    10.018
ntis         -46.9566     38.911     -1.207      0.229      -123.762    29.848
tbl           -2.8651     20.922     -0.137      0.891       -44.162    38.432
ltr           10.2432     14.468      0.708      0.480       -18.314    38.800
tms           13.1083     11.129      1.178      0.240        -8.859    35.076
dfy         -156.8202    213.943     -0.733      0.465      -579.111   265.471
dfr           71.0710     29.099      2.442      0.016        13.634   128.508
infl         -36.9489     82.870     -0.446      0.656      -200.521   126.623
ik          -208.4868    242.844     -0.859      0.392      -687.824   270.851
==============================================================================
```

Regularisation methods

Regularisation methods (key concept)

Regularisation or **shrinkage** methods for linear regression follow the general framework of **regularised risk minimisation** introduced in Module 7:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; \left[\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))\right] + \lambda\, C(\boldsymbol{\theta}),$$

Here, the loss function is the squared loss and the complexity function will be the norm of the vector of regression coefficients β. The choice of norm leads to different regularisation properties.

Ridge regression (key concept)

The **ridge regression** method solves the penalised estimation problem

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

for a tuning parameter λ.

The penalty term $\lambda||\boldsymbol{\beta}||_2^2$ has the effect of shrinking the coefficients relative to OLS. We refer to this procedure as $\boldsymbol{\ell_2}$ **regularisation**.

Ridge regression

The ridge estimator has an equivalent formulation as a constrained minimisation problem

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 < t.$$

for some $t > 0$.

Practical details

1. The hyperparameters λ or t control the amount of shrinkage. There is an one-to-one connection between them.

2. We do not penalise the intercept. In practice, we center the response and the predictors before computing the solution and estimate the intercept as $\widehat{\beta}_0 = \overline{y}$.

3. The method is not invariant on the scale of the inputs. We standardise the predictors before solving the minimisation problem.

Ridge regression

We can write the minimisation problem in matrix form as

$$\min_{\beta} \, (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \, \boldsymbol{\beta}^T\boldsymbol{\beta}.$$

Relying on the same techniques that we used to derive the OLS estimator, we can show the ridge estimator has the formula

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \, \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Orthonormal vectors

We say that two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthonormal when

$$||\boldsymbol{u}|| = \sqrt{\boldsymbol{u}^T\boldsymbol{u}} = 1, \quad ||\boldsymbol{v}|| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}} = 1, \quad \text{and} \quad \boldsymbol{u}^T\boldsymbol{v} = 0.$$

We say that the design matrix $\boldsymbol{X}$ is orthonormal when all its columns are orthonormal.

Ridge regression: shrinkage (key concept)

If the design matrix $\boldsymbol{X}$ was orthonormal, the ridge estimate would just a scaled version of the OLS estimate

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ridge}} = (\boldsymbol{I} + \lambda\, \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{y} = \frac{1}{1+\lambda} \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}$$

In a more general situation, we can say that the ridge regression method will shrink together the coefficients of correlated predictors.

Ridge regression

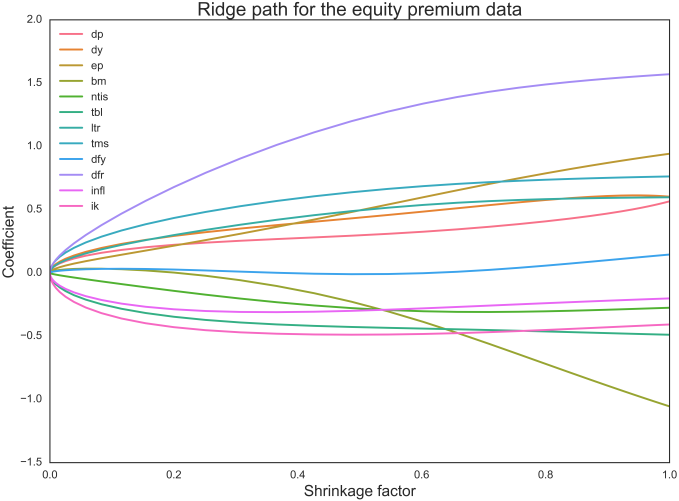
We define the ridge shrinkage factor as

$$s(\lambda) = \frac{||\widehat{\boldsymbol{\beta}}_{\textsf{ridge}}||_2}{||\widehat{\boldsymbol{\beta}}_{\textsf{ols}}||_2},$$

for a given λ or t.

The next slide illustrates the effect of varying the shrinkage factor on the estimated parameters.

Ridge coefficient profiles (equity premium data)

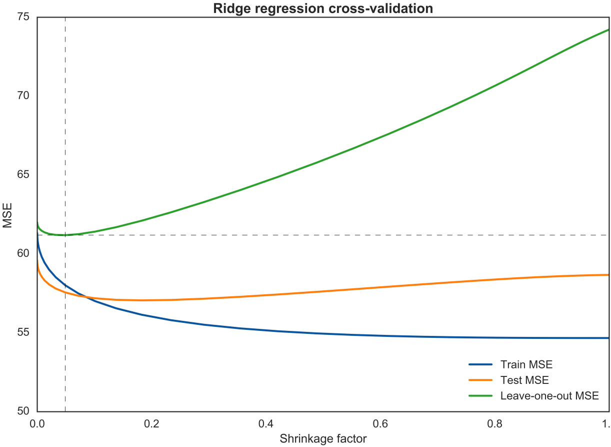


Selecting λ

The ridge regression method leads to a range of models for different values of λ. We select λ by cross validation or generalised cross validation.

GCV is computationally convenient for this model.

Selecting λ (equity premium data)



Ridge regression: Bayesian interpretation

The ridge regression estimator is the posterior mean and mode of a Bayesian Gaussian linear regression model with prior

$$p(\boldsymbol{\beta}) = N(0, \boldsymbol{I}\tau^2)$$

if we define $\lambda = \sigma^2/\tau^2$.

The Lasso

The **Lasso** (least absolute shrinkage and selection operator) method solves the penalised estimation problem

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

for a tuning parameter λ.

The Lasso therefore performs $\boldsymbol{\ell_1}$ **regularisation**.

The Lasso

The equivalent formulation of the lasso as a constrained minimisation problem is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{lasso}} &= \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2 \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| < t.\end{aligned}$$

for some $t > 0$.

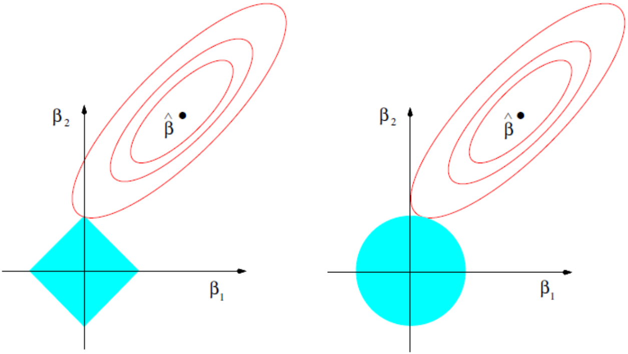
The Lasso: shrinkage and variable selection (key concept)

Shrinkage. As with ridge regression, the lasso shrinks the coefficients towards zero. However, the nature of this shrinkage is different, as we will see below.

Variable selection. In addition to shrinkage, the lasso also performs variable selection. With λ sufficiently large, some estimated coefficients will be exactly zero, leading to sparse models. This is a key difference from ridge.

The Lasso: variable selection property

Estimation picture for the lasso (left) and ridge regression (right):



Practical details

1. We select the tuning parameter λ by cross validation.
2. As with ridge, we center and standardise the predictors before computing the solution.
3. There is no closed form solution for the lasso coefficients. Computing the lasso solution is a quadratic programming problem.
4. There are efficient algorithms for computing an entire path of solutions for a range of λ values.

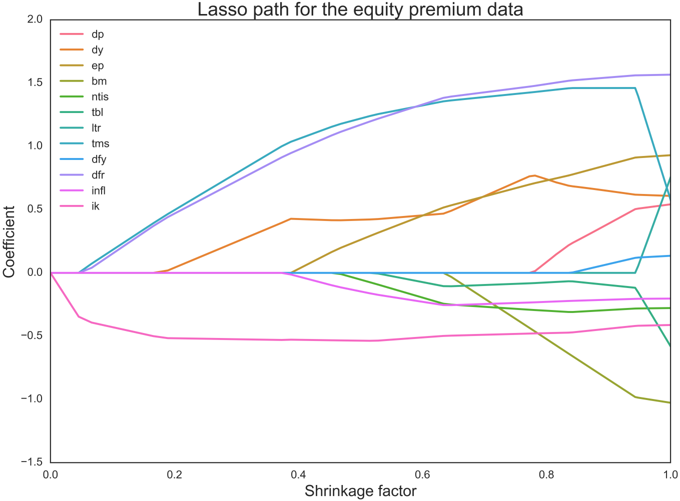
The Lasso

We define the shrinkage factor for a given value of λ (or t) as

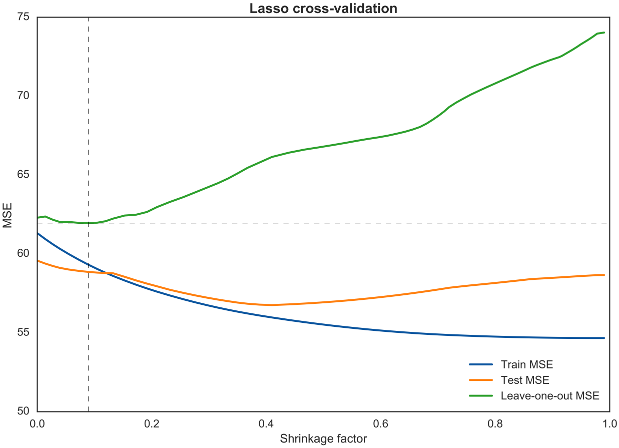
$$s(\lambda) = \frac{\sum_{j=1}^{p} \left|\widehat{\beta}_j^{\text{lasso}}\right|}{\sum_{j=1}^{p} \left|\widehat{\beta}_j^{\text{ols}}\right|}.$$

The next slide illustrates the effect of varying the shrinkage factor on the estimated parameters.

Lasso coefficient profiles (equity premium data)



Model selection for the equity premium data



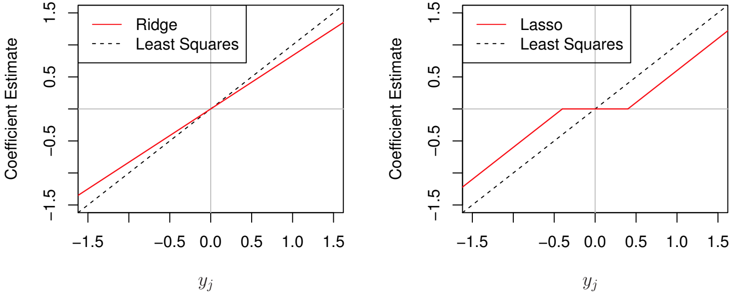
Discussion

Subset selection, ridge, and lasso: comparison in the orthonormal case

Estimator	Formula
Best subset (size k)	$\widehat{\beta}_j \cdot I(\lvert\widehat{\beta}_j\rvert > \lvert\widehat{\beta}_{(k)}\rvert)$
Ridge	$\widehat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\widehat{\beta}_j)(\lvert\widehat{\beta}_j\rvert - \lambda)_+$

Estimators of β_j in the case of orthonormal columns of $\boldsymbol{X}$.

Ridge and Lasso: comparison in the orthonormal case



Which method to use?

- Recall the **no free lunch theorem**: neither ridge regression or the lasso universally outperform the other. The choice of method should be data driven.

- In general terms, we can expect the lasso to perform better when a small subset of predictors have important coefficients, while the remaining predictors having small or zero coefficients (sparse problems).

- Ridge regression will tend to perform better when the predictors all have similar importance.

- The lasso may have better interpretability since it can lead to a sparse solution.

Elastic Net

The **elastic net** is a compromise between ridge regression and the lasso:

$$\widehat{\boldsymbol{\beta}}_{\mathsf{EN}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \Big(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \Big),$$

for $\lambda \geq 0$ and $0 < \alpha < 1$.

The elastic net performs variable selection like the lasso, and shrinks together the coefficients of correlated predictors like ridge regression.

Illustration: equity premium data

Estimated coefficients (tuning parameters selected by leave-one-out CV)

	OLS	**Ridge**	**Lasso**	**EN**
dp	0.566	0.159	0.000	0.111
dy	0.602	0.197	0.000	0.153
ep	0.942	0.116	0.000	0.048
bm	-1.055	0.033	0.000	0.000
ntis	-0.276	-0.067	-0.000	-0.000
tbl	-0.489	-0.248	-0.000	-0.178
ltr	0.597	0.186	0.000	0.124
tms	0.762	0.286	0.161	0.239
dfy	0.145	0.031	0.000	0.000
dfr	1.570	0.377	0.131	0.294
infl	-0.202	-0.214	-0.000	-0.150
ik	-0.408	-0.318	-0.422	-0.282

Illustration: equity premium data

Prediction results

	Train R^2	Test R^2
OLS	0.108	0.014
Ridge	0.054	0.033
Lasso	0.033	0.011
Elastic Net	0.050	0.029

Comparison with variable selection

Regularisation methods have two important advantages over variable selection.

1. They are continuous procedures, generally leading to lower variance.
2. The computational cost is not much larger than OLS.

Review questions

- What is best subset selection?
- What are stepwise methods?
- What are the advantages and disadvantages of variable selection?
- What are the penalty terms in the ridge and lasso methods?
- What are the key differences between the ridge and lasso methods?
- In what situations would we expect the ridge or lasso methods to perform better?