Non-OLS Linear SEL Prediction: Ridge Regression

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Non-OLS linear SEL predictors

There is more to say about the development of a linear predictor

$$\widehat{f}(\mathbf{x}) = \mathbf{x}'\hat{\beta} \quad (1)$$

for an appropriate $\hat{\beta} \in \Re^p$ than what is said in books and courses on ordinary linear models (where ordinary least squares is used to fit the linear form to all p input variables or to some subset of M of them). In what follows we continue the basic notation

$$\underset{N \times p}{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_N' \end{pmatrix} \text{ and } \underset{N \times 1}{\mathbf{Y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

and consider non-OLS choices of $\hat{\beta}$ in (1) based on such training data.

Technical framework

An alternative to seeking a suitable level of complexity in a linear prediction rule through subset selection and least squares fitting of a linear form to the selected variables, is to employ a shrinkage method based on a penalized version of least squares to choose a vector $\hat{\boldsymbol{\beta}} \in \Re^p$. Here we consider several such methods, all of which have parameters that function as complexity measures and allow behavior to range between $\hat{\boldsymbol{\beta}} = \mathbf{0}$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{OLS}}$ depending upon complexity.

The implementation of these methods is not equivariant to the scaling used to express the input variables x_j. So that there is a well-defined scaling, we **assume here that the output variable has been centered** (i.e. that $\langle \mathbf{Y}, \mathbf{1} \rangle = 0$) **and that the columns of $\mathbf{X}$ have been standardized** (and if originally $\mathbf{X}$ had a constant column, it has been removed).

Equivalent optimization formulations

For a $\lambda > 0$ the ridge regression coefficient vector $\widehat{\beta}_\lambda^{\text{ridge}} \in \Re^p$ is

$$\widehat{\beta}_\lambda^{\text{ridge}} = \underset{\beta \in \Re^p}{\arg\min} \left\{ (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta) + \lambda \beta' \beta \right\} \qquad (2)$$

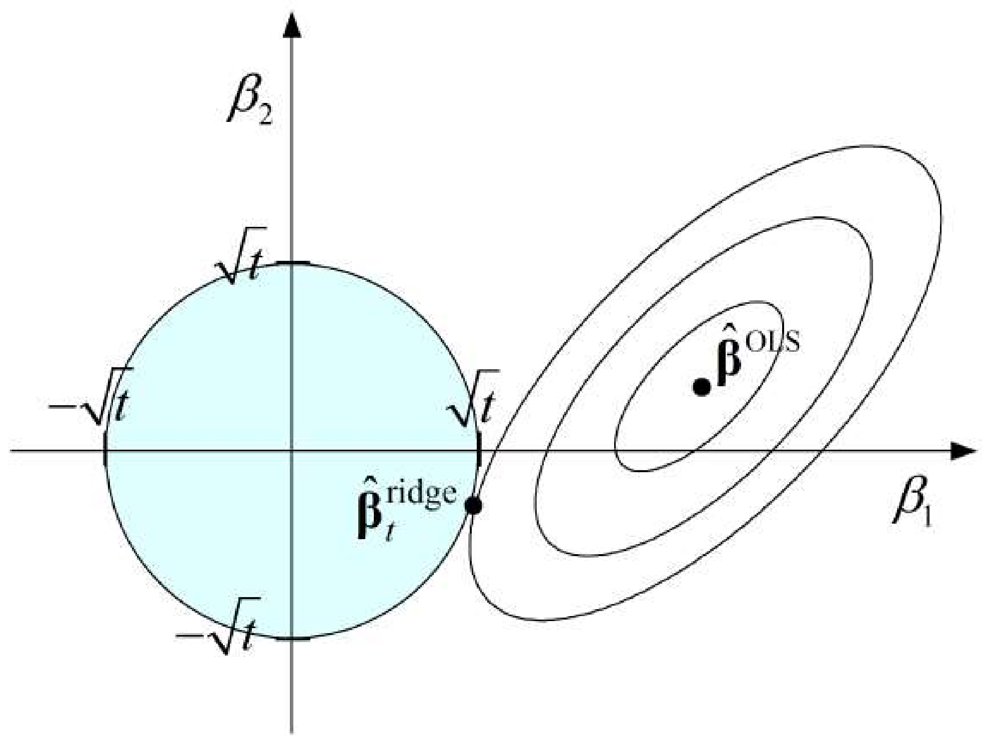
Here λ is a penalty/complexity parameter that controls how much $\widehat{\beta}^{\text{OLS}}$ is shrunken towards $\mathbf{0}$. The unconstrained minimization problem expressed in (2) has an equivalent constrained minimization description as

$$\widehat{\beta}_t^{\text{ridge}} = \underset{\beta \text{ with } \|\beta\|^2 \leq t}{\arg\min} (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta) \qquad (3)$$

for an appropriate $t > 0$. (Corresponding to λ used in (2), is $t = \left\| \widehat{\beta}_\lambda^{\text{ridge}} \right\|^2$ used in (3). Conversely, corresponding to t used in (3), one may use a value of λ in (2) producing the same error sum of squares.)

Geometry of ridge optimization

Here is a representation of the constrained optimization problem solved by the ridge coefficient vector, $\widehat{\boldsymbol{\beta}}_t^{\text{ridge}}$ for $p = 2$.



Ridge form and shrinking OLS coefficients

The unconstrained form (2) calls upon one to minimize

$$(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)+\lambda\beta'\beta$$

and some vector calculus leads directly to

$$\widehat{\beta}_{\lambda}^{\text{ridge}}=\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Then, using the singular value decomposition of $\mathbf{X}$ (with $rank = r$) it's possible to argue that

$$\widehat{\mathbf{Y}}_{\lambda}^{\text{ridge}}=\mathbf{X}\widehat{\beta}_{\lambda}^{\text{ridge}}=\sum_{j=1}^{r}\left(\frac{d_j^2}{d_j^2+\lambda}\right)\langle\mathbf{Y},\mathbf{u}_j\rangle\,\mathbf{u}_j$$

Coefficients of the orthonormal basis vectors $\mathbf{u}_j$ producing $\widehat{\mathbf{Y}}_{\lambda}^{\text{ridge}}$ are shrunken version of the coefficients producing $\widehat{\mathbf{Y}}^{\text{OLS}}$. The most severe shrinking is enforced in the directions of the smallest principal components of $\mathbf{X}$ (the $\mathbf{u}_j$ least important in making up low rank approximations to $\mathbf{X}$).

Shrinking OLS prediction vectors

From

$$\widehat{\mathbf{Y}}_\lambda^{\text{ridge}} = \mathbf{X}\widehat{\beta}_\lambda^{\text{ridge}} = \sum_{j=1}^{r} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) \langle \mathbf{Y}, \mathbf{u}_j \rangle \, \mathbf{u}_j \tag{4}$$

the norm of the vector ridge predictions for the N centered responses is

$$\left\| \widehat{\mathbf{Y}}_\lambda^{\text{ridge}} \right\|^2 = \sum_{j=1}^{r} \left(\frac{d_j^2}{d_j^2 + \lambda} \right)^2 \langle \mathbf{Y}, \mathbf{u}_j \rangle^2$$

and is thus decreasing in λ.

Notice also from (4) that

$$\widehat{\mathbf{Y}}_\lambda^{\text{ridge}} = \mathbf{X} \sum_{j=1}^{r} \left(\frac{1}{d_j^2 + \lambda} \right) \langle \mathbf{Y}, \mathbf{X}\mathbf{v}_j \rangle \, \mathbf{v}_j$$

Shrinking OLS coefficient vectors

Thus

$$\widehat{\beta}_{\lambda}^{\text{ridge}} = \sum_{j=1}^{r} \left(\frac{1}{d_j^2 + \lambda} \right) \langle \mathbf{Y}, \mathbf{X}\mathbf{v}_j \rangle \mathbf{v}_j$$

and

$$\left\| \widehat{\beta}_{\lambda}^{\text{ridge}} \right\|^2 = \sum_{j=1}^{r} \left(\frac{1}{d_j^2 + \lambda} \right)^2 \left(\langle \mathbf{Y}, \mathbf{X}\mathbf{v}_j \rangle \right)^2$$

which is also clearly decreasing in λ.

An upshot of these facts about "shrinking" is that one can think of (the penalty parameter) λ as a complexity parameter that defines paths in $\Re^N$ and $\Re^p$ from OLS predictions and coefficients to degenerate ($\mathbf{0}$) ones passing through a spectrum of plausible (ridge) linear predictors.

Coefficient grouping effect

There is a "grouping effect" associated with ridge regression. Highly correlated inputs, say x_j and $x_{j'}$, (being standardized so they both have standard deviation 1) have ridge regression coefficients of essentially the same magnitude. This can be understood as follows. Without loss of generality, assume that x_j and $x_{j'}$ are highly positively correlated (so they are essentially the same variable). For any regression coefficients β_j and $\beta_{j'}$ and number α (including $\beta_j / (\beta_j + \beta_{j'})$) the contribution of x_j and $x_{j'}$ to $\hat{y}$ (and thus the error sum of squares) is

$$\beta_j x_j + \beta_{j'} x_{j'} \approx \alpha (\beta_j + \beta_{j'}) x_j + (1 - \alpha) (\beta_j + \beta_{j'}) x_{j'}$$

But the contribution of $\alpha (\beta_j + \beta_{j'})$ and $(1 - \alpha) (\beta_j + \beta_{j'})$ to the sum of squared regression coefficients is

$$\alpha^2 (\beta_j + \beta_{j'})^2 + (1 - \alpha)^2 (\beta_j + \beta_{j'})^2 = \left(\alpha^2 + (1 - \alpha)^2\right) (\beta_j + \beta_{j'})^2$$

minimized at $\alpha = 1/2$, where the coefficients for x_j and $x_{j'}$ are the same.

Ridge "effective degrees of freedom"

The function

$$\mathrm{df}(\lambda) = \mathrm{tr}\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\right) = \sum_{j=1}^{r}\left(\frac{d_j^2}{d_j^2 + \lambda}\right)$$

is called the "effective degrees of freedom" associated with the ridge regression. In regard to this choice of nomenclature, note again that if $\lambda = 0$ ridge regression is ordinary least squares and this is r, the usual degrees of freedom associated with projection onto $C(\mathbf{X})$, i.e. trace of the projection matrix onto this column space and that as $\lambda \to \infty$, the effective degrees of freedom goes to 0 and (the centered) $\widehat{\mathbf{Y}}_\lambda^{\text{ridge}}$ goes to $\mathbf{0}$ (corresponding to a constant predictor).

More general forms for effective df

Notice that since $\widehat{\mathbf{Y}}_\lambda^{\text{ridge}} = \mathbf{X}\widehat{\beta}_\lambda^{\text{ridge}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{MY}$ for $\mathbf{M} = \mathbf{X}\left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'$, if one assumes that

$$\text{Cov}\mathbf{Y} = \sigma^2\mathbf{I}$$

(conditioned on the $\mathbf{x}_i$ in the training data, the outputs are uncorrelated and have constant variance σ^2) then

$$\text{effective degrees of freedom} = \text{tr}\left(\mathbf{M}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{N}\text{Cov}\left(\hat{y}_i, y_i\right) \qquad (5)$$

This suggests that tr($\mathbf{M}$) is a plausible definition for effective degrees of freedom for any *linear* fitting method $\widehat{\mathbf{Y}} = \mathbf{MY}$, and that more *generally*, the last form in (5) might be used in situations where $\widehat{\mathbf{Y}}$ is other than a linear form in $\mathbf{Y}$. The last form is a measure of how strongly the outputs in the training set can be expected to be related to their predictions.

Another alternative form for effective df

Some additional insight into the notion of effective degrees of freedom is this. In the linear case, with $\widehat{\mathbf{Y}} = \mathbf{MY}$,

$$\text{effective degrees of freedom} = \text{tr}\left(\mathbf{M}\right) = \sum_{i=1}^{N} \frac{\partial \hat{y}_i}{\partial y_i}$$

and we see that the effective degrees of freedom is some total measure of how sensitive predictions are at the training inputs $\mathbf{x}_i$ to the corresponding training values y_i.

This raises at least the possibility that in nonlinear cases, an approximate/estimated value of the general effective degrees of freedom (5) might be the random variable

$$\sum_{i=1}^{N} \frac{\partial \hat{y}_i}{\partial y_i}\bigg|_{\mathbf{Y}}$$