Prediction

Kerby Shedden

Department of Statistics, University of Michigan

November 9, 2018

Prediction analysis

In a prediction-oriented analysis, we are interested in fitting a model to capture the mean relationship between indpendent variables x and a dependent variable y, then using the fitted model to make predictions on an independent data set.

The model has the form f_θ, where for each θ, we have a function from $\mathcal{R}^p$ to $\mathcal{R}$. Thus $\{f_\theta\}$ is a family of functions indexed by a parameter θ.

We use the data to obtain an estimate $\hat{\theta}$ of θ, which in turn leads us to an estimate $f_{\hat{\theta}}$ of the regression function.

It is helpful to think in terms of training data $\{(y_i, x_i)\}$ that are used to fit the model, so $\hat{\theta} = \hat{\theta}(\{(y_i, x_i)\})$, and testing data $\{(y_i^*, x_i^*)\}$ on which predictions are made.

Quantifying prediction error

Prediction analysis focuses on prediction errors, for example through the mean squared prediction error (MSPE):

$$E|Y^* - f_{\hat{\theta}}(\mathbf{X}^*)|^2,$$

and its sample analogue

$$\sum_{i=1}^{n^*} \|y_i^* - f_{\hat{\theta}}(\mathbf{x}_i^*)\|^2 / n^*,$$

where n^* is the size of the testing set.

Prediction analysis does not usually focus on properties of the parameter estimates themselves, e.g. bias $E\hat{\theta} - \theta$, or parameter MSE $E(\hat{\theta} - \theta)^2$.

MSPE for OLS analysis

The mean squared prediction error for OLS regression is easy to derive. The testing data follow $y^* = \mathbf{X}^*\beta + \epsilon^*$. Let $\hat{y}^* = \mathbf{X}^*\hat{\beta}$ denote the predicted values in the test set. Then

$$\begin{aligned} E\|y^* - \hat{y}^*\|^2 &= E\|\mathbf{X}^*\beta + \epsilon^* - \mathbf{X}^*\hat{\beta}\|^2 \\ &= E\|\mathbf{X}^*(\beta - \hat{\beta})\|^2 + E\|\epsilon^*\|^2 \\ &= E[(\hat{\beta} - \beta)'(\mathbf{X}^{*\prime}\mathbf{X}^*)(\hat{\beta} - \beta)] + n^*\sigma^2 \\ &= \mathrm{tr}\left(\mathbf{X}^{*\prime}\mathbf{X}^* \cdot E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right)] + n^*\sigma^2 \\ &= \mathrm{tr}\left(\mathbf{X}^{*\prime}\mathbf{X}^* \cdot \Sigma_{\hat{\beta}}\right) + n^*\sigma^2, \end{aligned}$$

where $\Sigma_{\hat{\beta}}$ is the covariance matrix of $\hat{\beta}$ from the training process. Note the requirement for $\hat{Y}^*$ and Y^* to be independent (given $\mathbf{X}$ and $\mathbf{X}^*$).

MSPE for OLS analysis

The MSPE for OLS is

$$\mathrm{tr}\left((\mathbf{X}^{*\prime}\mathbf{X}^{*}/n^{*})\cdot\Sigma_{\hat{\beta}}\right)+\sigma^{2}.$$

If $\mathbf{X}$ is the training set design matrix, then $\Sigma_{\hat{\beta}}=\sigma^{2}(\mathbf{X}'\mathbf{X})^{-1}$, so if $\mathbf{X}=\mathbf{X}^{*}$, then

$$E\|y^{*}-\hat{y}\|^{2}=\sigma^{2}(p+1+n^{*}),$$

and the MSPE in this case is

$$\sigma^{2}(p+1)/n^{*}+\sigma^{2}=\sigma^{2}(p+1)/n+\sigma^{2}.$$

MSPE for OLS analysis

More generally, suppose $\mathbf{X}'\mathbf{X}/n = \mathbf{X}^{*\prime}\mathbf{X}^{*}/n^{*}$. Then

$$\Sigma_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 n^{*}(\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}/n.$$

Thus the MSPE is

$$\mathrm{tr}\left((\mathbf{X}^{*\prime}\mathbf{X}^{*}/n^{*}) \cdot \Sigma_{\hat{\beta}}\right) + \sigma^2 = \sigma^2(p+1)/n + \sigma^2.$$

MSPE in practice

The MSPE discussed here is the primary population quantity of interest for prediction. It is not straightforward to estimate however.

The task of model selection, discussed later, can be viewed as aiming to identify the model with the lowest MSPE (among a set of candidate models under consideration).

Note that the candidate models we fit to data may not be correctly-specified, so the usual estimate of $\hat{\sigma}^2$ may be biased.

PRESS residuals

One way to estimate the MSPE with few theoretical conditions is using cross validation. We will briefly introduce this idea here, then return to it and cover it in more detail when we talk about model selection.

If case i is deleted and a prediction of y_i is made from the remaining data, we can compare the observed and predicted values to get the prediction residual:

$$r_{(i)} \equiv y_i - \hat{y}_{(i)i}.$$

where $\hat{y}_{(i)i}$ is the prediction of y_i based on a data set in which case i was removed.

PRESS residuals

A simple formula for the prediction residual in OLS is given by

$$\begin{aligned} r_{(i)} &= y_i - \mathbf{x}_i \hat{\beta}_{(i)} \\ &= y_i - \mathbf{x}_i (\hat{\beta} - r_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i / (1 - P_{ii})) \\ &= r_i / (1 - P_{ii}). \end{aligned}$$

where $\mathbf{X}$ is the design matrix, $\mathbf{x}_i$ is row i of the design matrix, and P is the projection matrix (for the full sample).

The sum of squares of the prediction residuals is called PRESS (prediction residual error sum of squares). It is equivalent to using leave-one-out cross validation to estimate the "generalization error rate".

Bias and variance in prediction

If we are using a function $f_{\hat{\theta}}$ to predict y from x, we can view the prediction error as arising from contributions of bias and variance.

The bias is

$$b(x) \equiv E[f_{\hat{\theta}}(x)|x] - E[y|x].$$

The variance is $v(x) \equiv \mathrm{var}[f_{\hat{\theta}}(x)|x]$.

The MSPE is

$$E[(y - f_{\hat{\theta}}(x))^2] = b(x)^2 + v(x).$$

Bias and variance in prediction

While having zero bias is an important consideration in some statistical analyses, arguably the overall accuracy, as measured by MSPE, should be the dominant consideration.

The MSPE results from a combination of squared bias and variance. If we want to minimize the MSPE we should consider using a biased estimator, if by doing so we attain better MSPE (due to it having much smaller variance).

The relationship between bias and variance discussed here is often referred to as the bias/variance tradeoff.

Ridge regression

Ridge regression uses the minimizer of a penalized squared error loss function to estimate the regression coefficients:

$$\hat{\beta} \equiv \text{argmin}_{\beta} \|y - \mathbf{X}\beta\|^2 + \lambda \beta' D \beta.$$

Typically D is a diagonal matrix with 0 in the 1,1 position and ones on the rest of the diagonal. In this case,

$$\beta' D \beta = \sum_{j \geq 1} \beta_j^2.$$

This makes most sense when the covariates have been standardized, so it is reasonable to penalize the β_j equally.

Ridge regression

Ridge regression is a compromise between fitting the data as well as possible (by making $\|y - \mathbf{X}\beta\|^2$ small), while not allowing any one fitted coefficient to get very large (which causes $\beta' D \beta$ to get large).

Ridge regression and colinearity

Suppose x_1 and x_2 are standardized vectors with a substantial positive correlation (i.e. $x_1' x_2$ is large), and the population slopes are β_1 and β_2, i.e. $E[y|x_1, x_2] = \beta_1 x_1 + \beta_2 x_2$.

Fits of the form

$$(\beta_1 + \gamma)x_1 + (\beta_2 - \gamma)x_2 = E[y|x_1, x_2] + \gamma(x_1 - x_2)$$

have similar MSE values as γ varies, since $x_1 - x_2$ is small when x_1 and x_2 are strongly positively associated.

In other words, OLS can't easily distinguish among these fits.

For example, if $x_1 \approx x_2$, then $3x_1 + 3x_2$, $4x_1 + 2x_2$, $5x_1 + x_2$, etc. all produce similar fitted values.

Ridge regression and colinearity

For large λ, ridge regression favors the fits that minimize

$$(\beta_1 + \gamma)^2 + (\beta_2 - \gamma)^2.$$

This expression is minimized at $\gamma = (\beta_2 - \beta_1)/2$, giving the fit

$$(\beta_1 + \beta_2)x_1/2 + (\beta_1 + \beta_2)x_2/2.$$

$\Rightarrow$ Ridge regression favors coefficient estimates for which strongly positively correlated covariates have similar estimated effects.

Calculation of ridge regression estimates

For a given value $\lambda > 0$, ridge regression is no more difficult computationally than ordinary least squares, since

$$\frac{\partial}{\partial \beta}\|y - \mathbf{X}\beta\|^2 + \lambda \beta' D \beta \quad = \quad -2\mathbf{X}'y + 2\mathbf{X}'\mathbf{X}\beta + 2\lambda D\beta,$$

so the ridge estimate $\hat{\beta}$ solves the system of linear equations

$$(\mathbf{X}'\mathbf{X} + \lambda D)\beta = \mathbf{X}'y.$$

This equation can have a unique solution even when $\mathbf{X}'\mathbf{X}$ is singular. Thus one application of ridging is to produce regression estimates for singular design matrices.

Ridge regression bias and variance

Ridge regression estimates are biased, but may be less variable than OLS estimates. If $\mathbf{X}'\mathbf{X}$ is non-singular, the ridge estimator can be written

$$\begin{aligned}
\hat{\beta}_\lambda &= (\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'y \\
&= (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \\
&= (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) \\
&= (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}\beta + (I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon.
\end{aligned}$$

Thus the bias is

$$E[\hat{\beta}_\lambda|\mathbf{X}] - \beta = ((I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1} - I)\beta$$

Ridge regression bias and variance

The variance of the ridge regression estimates is

$$\mathrm{var}\hat{\beta}_\lambda = \sigma^2(I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-1}(\mathbf{X}'\mathbf{X})^{-1}(I + \lambda(\mathbf{X}'\mathbf{X})^{-1}D)^{-T}.$$

Ridge regression bias and variance

Next we will show that $\mathrm{var}[\hat{\beta}] \geq \mathrm{var}[\hat{\beta}_\lambda]$, in the sense that

$$\mathrm{var}[\hat{\beta}] - \mathrm{var}[\hat{\beta}_\lambda]$$

is a non-negative definite matrix.

First let $M = \lambda(\mathbf{X}'\mathbf{X})^{-1}D$, and note that

Ridge regression bias and variance

$$
\begin{aligned}
v'(\mathrm{var}\hat{\beta} - \mathrm{var}\hat{\beta}_\lambda)v &\propto v'\left((\mathbf{X}'\mathbf{X})^{-1} - (I+M)^{-1}(\mathbf{X}'\mathbf{X})^{-1}(I+M)^{-T}\right)v \\
&= u'\left((I+M)(\mathbf{X}'\mathbf{X})^{-1}(I+M)' - (\mathbf{X}'\mathbf{X})^{-1}\right)u \\
&= u'\left(M(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}M' + M(\mathbf{X}'\mathbf{X})^{-1}M'\right)u \\
&= u'\left(2\lambda(\mathbf{X}'\mathbf{X})^{-1}D(\mathbf{X}'\mathbf{X})^{-1} + \right. \\
&\qquad \left. \lambda^2(\mathbf{X}'\mathbf{X})^{-1}D(\mathbf{X}'\mathbf{X})^{-1}D(\mathbf{X}'\mathbf{X})^{-1}\right)u
\end{aligned}
$$

where $u = (I+M)^{-T}v$.

We can conclude that for any fixed vector θ,

$$\mathrm{var}(\theta'\hat{\beta}_\lambda) \leq \mathrm{var}(\theta'\hat{\beta}).$$

Ridge regression effective degrees of freedom

Like OLS, the fitted values under ridge regression are linear functions of the observed values

$$\hat{Y}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'Y$$

In OLS regression, the degrees of freedom is the number of free parameters in the model, which is equal to the trace of the projection matrix P that satisfies $\hat{Y} = PY$.

Fitted values in ridge regression are not a projection of Y, but the matrix

$$\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'.$$

plays an analogous role to P.

Ridge regression effective degrees of freedom

The effective degrees of freedom for ridge regression is defined as

$$\mathrm{EDF}_\lambda = \mathrm{tr}\left[\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'\right].$$

The trace can be easily computed using the identity

$$\mathrm{trace}\left(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'\right) = \mathrm{trace}\left((\mathbf{X}'\mathbf{X} + \lambda D)^{-1}\mathbf{X}'\mathbf{X}\right).$$

Ridge regression effective degrees of freedom

EDF_λ is monotonically decreasing in λ. To see this we will use the following fact about matrix derivatives

$$\partial \text{tr}(A^{-1}B)/\partial A = -A^{-T}B'A^{-T}.$$

By the chain rule, letting $A = \mathbf{X}'\mathbf{X} + \lambda D$, we have

$$\begin{aligned}
\partial \text{tr}\left(A^{-1}\mathbf{X}'\mathbf{X}\right)/\partial\lambda &= \sum_{ij} \frac{\partial \text{tr}\left(A^{-1}\mathbf{X}'\mathbf{X}\right)}{\partial A_{ij}} \cdot \frac{\partial A_{ij}}{\partial \lambda} \\
&= -\sum_{ij} [A^{-T}(\mathbf{X}'\mathbf{X})A^{-T}]_{ij} \cdot D_{ij} \\
&= -\sum_{i} [A^{-T}(\mathbf{X}'\mathbf{X})A^{-T}]_{ii} \cdot D_{ii} \\
&\leq 0.
\end{aligned}$$

Ridge regression effective degrees of freedom

EDF_λ equals $\text{rank}(\mathbf{X})$ when $\lambda = 0$. To see what happens as $\lambda \to \infty$, we can apply the Sherman-Morrison-Woodbury identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1}.$$

Let $G = \mathbf{X}'\mathbf{X}$, and write $D = FF'$, where F has independent columns (usually F will be $p + 1 \times p$ as we do not penalize the intercept).

Ridge regression effective degrees of freedom

Applying the SMW identity and letting $\lambda \to \infty$ we get

$$\begin{aligned}
\mathrm{tr}\left[(G+\lambda D)^{-1}G\right] &= \mathrm{tr}\left[\left(G^{-1}-G^{-1}F(I/\lambda+F'G^{-1}F)^{-1}F'G^{-1}\right)G\right] \\
&= \mathrm{tr}\left[I_{p+1}-G^{-1}F(I/\lambda+F'G^{-1}F)^{-1}F'\right] \\
&\to \mathrm{tr}I_{p+1}-\mathrm{tr}\left[G^{-1}F(F'G^{-1}F)^{-1}F'\right] \\
&\to \mathrm{tr}I_{p+1}-\mathrm{tr}\left[(F'G^{-1}F)^{-1}F'G^{-1}F\right] \\
&= p+1-\mathrm{rank}(F).
\end{aligned}$$

Therefore in the usual case where F has rank p, EDF_λ converges to 1 as λ grows large, reflecting the fact that all coefficients other than the intercept are forced to zero.

Ridge regression and the SVD

Suppose we are fitting a ridge regression with $D = I$, and we factor $\mathbf{X} = USV'$ using the singular value decomposition (SVD), so that U and V are orthogonal matrices, and S is a diagonal matrix with non-negative diagonal elements.

The fitted coefficients are

$$\begin{aligned}\hat{\beta}_\lambda &= (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'Y \\ &= (VS^2V' + \lambda VV')^{-1}VSU'Y \\ &= V(S^2 + \lambda I)^{-1}SU'Y\end{aligned}$$

Note that for OLS ($\lambda = 0$), we get $\hat{\beta} = VS^{-1}U'Y$. The effect of ridging is to replace S^{-1} in this expression with $(S^2 + \lambda I)^{-1}S$, which are uniformly smaller values when $\lambda > 0$.

Ridge regression tuning parameter

There are various ways to set the ridge parameter λ.

Cross-validation can be used to estimate the MSPE for any particular value of λ. Then this estimated MSPE could be minimized by checking its value at a finite set of λ values.

Generalized cross validation, which minimizes the following over λ, is a simpler, and more commonly used approach.

$$\mathrm{GCV}(\lambda) = \frac{\|Y - \hat{Y}_\lambda\|^2}{(n - \mathrm{EDF}_\lambda)^2}.$$