

25 Lecture 25: April 7

Last time

- Principal component analysis (JF 13.1.1, RD 8.3.4)

Today

- Presentation
- Biased estimation:
 - Ridge Regression
 - Lasso Regression
- Model selection
- One-way ANOVA

Additional reference

- “A First Course in Linear Model Theory” by Nalini Ravishanker and Kipak K. Dey
- [Lecture notes](#) by Cedric Ginestet

Ridge Regression

Ridge regression and the Lasso regression are two forms of regularized regression. These methods can be used to alleviate the consequences of multicollinearity.

1. When variables are highly correlated, a large coefficient in one variable may be alleviated by a large coefficient in another variable, which is negatively correlated to the former.
2. Regularization imposes an upper threshold on the values taken by the coefficients, thereby producing a more parsimonious solution, and a set of coefficients with smaller variance.

Constrained optimization

Ridge regression is motivated by a constrained minimization problem, which can be formulated as

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge} &= \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \\ \text{subject to } \|\boldsymbol{\beta}\|_2^2 &= \sum_{j=1}^p \beta_j^2 \leq t\end{aligned}$$

for $t \geq 0$.

Use a Lagrange multiplier, we can rewrite the formula as

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

for $\lambda \geq 0$ and where there is a one-to-one correspondence between t and λ . λ is an arbitrary constant usually referred to as the “ridge constant”.

Analytical solutions

The ridge-regression estimator has analytical solution

$$\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

This is obtained by differentiating the objective function with respect to $\boldsymbol{\beta}$ and set it to 0:

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\beta}} \{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \} \\ &= 2(\mathbf{X}^T \mathbf{X})\boldsymbol{\beta} - 2\mathbf{X}^T \mathbf{Y} + 2\lambda \boldsymbol{\beta} \\ &= 0 \end{aligned}$$

Therefore,

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})\boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}$$

Since we are adding a positive constant to the diagonal of $\mathbf{X}^T \mathbf{X}$, we are, in general, producing an invertible matrix, $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$ even if $\mathbf{X}^T \mathbf{X}$ is singular. Historically, this particular aspect of ridge regression was the main motivation behind the adoption of this particular extension of OLS theory.

The ridge regression estimator is related to the classical OLS estimator, $\hat{\boldsymbol{\beta}}^{OLS}$, in the following manner

$$\hat{\boldsymbol{\beta}}^{ridge} = [\mathbf{I} + \lambda(\mathbf{X}^T \mathbf{X})^{-1}]^{-1} \hat{\boldsymbol{\beta}}^{OLS},$$

assuming $\mathbf{X}^T \mathbf{X}$ is non-singular. This relationship can be verified by applying the definition of $\hat{\boldsymbol{\beta}}^{OLS}$,

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{ridge} &= [\mathbf{I} + \lambda(\mathbf{X}^T \mathbf{X})^{-1}]^{-1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y} \end{aligned}$$

using the fact $\mathbf{B}^{-1} \mathbf{A}^{-1} = (\mathbf{AB})^{-1}$.

Moreover, when \mathbf{X} is composed of orthonormal variables, such that $\mathbf{X}^T \mathbf{X} = \mathbf{I}_p$, it then follows that

$$\hat{\boldsymbol{\beta}}^{ridge} = \frac{1}{1 + \lambda} \hat{\boldsymbol{\beta}}^{OLS}$$

Bias and variance of ridge estimator

Ridge estimation produces a biased estimator of the true parameter β . With the definition of $\hat{\beta}^{ridge}$ and the model assumption $\mathbf{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$, we obtain,

$$\begin{aligned}\mathbf{E}(\hat{\beta}^{ridge}|\mathbf{X}) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \lambda\mathbf{I})\beta \\ &= \beta - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\beta\end{aligned}$$

where the bias of the ridge estimator is proportional to λ . The variance of the ridge estimator is

$$\mathbf{Var}(\hat{\beta}^{ridge}|\mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}.$$

When λ increases, the inverted term $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$ is increasingly dominated by $\lambda\mathbf{I}$. The variance of the ridge estimator, therefore, is a decreasing function of λ . This result is intuitively reasonable because the estimator itself is driven toward $\mathbf{0}$.

Variance-bias tradeoff

The mean-squared error of an estimator can be decomposed into the sum of its squared bias and sampling variance.

$$\begin{aligned}MSE(\hat{\theta}) &= \mathbf{E}((\hat{\theta} - \theta)^2) = \mathbf{E}(\hat{\theta}^2) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ Bias^2(\hat{\theta}) &= [\mathbf{E}(\hat{\theta}) - \theta]^2 = \mathbf{E}^2(\hat{\theta}) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ \mathbf{Var}(\hat{\theta}) &= \mathbf{E}(\hat{\theta}^2) - \mathbf{E}^2(\hat{\theta})\end{aligned}$$

Therefore

$$MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + \mathbf{Var}(\hat{\theta})$$

The essential idea here is to trade a small amount of bias in the coefficient estimates for a large reduction in coefficient sampling variance. Hoerl and Kennard (1970) prove that it is always possible to choose a positive value of the ridge constant λ so that the mean-squared error of the ridge estimator is less than the mean-squared error of the least-squares estimator. These ideas are illustrated heuristically in Figure 25.1

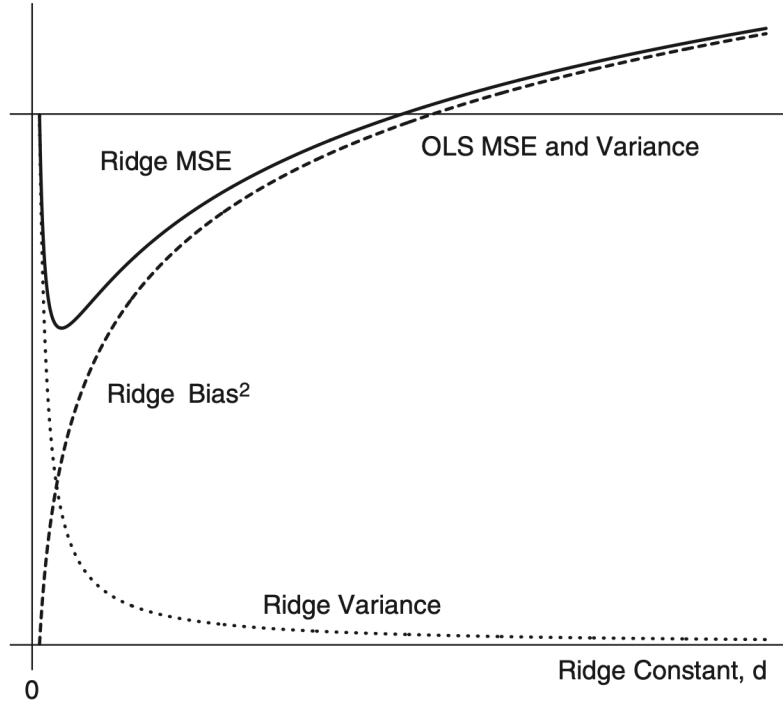


Figure 25.1: Trade-off of bias and against variance for the ridge-regression estimator. The horizontal line gives the variance of the least-squares (OLS) estimator; because the OLS estimator is unbiased, its variance and mean-squared error are the same. The broken line shows the squared bias of the ridge estimator as an increasing function of the ridge constant d (i.e. λ in our notes). The dotted line shows the variance of the ridge estimator. The mean-squared error (MSE) of the ridge estimator, given by the heavier solid line, is the sum of its variance and squared bias. For some values of d , the MSE error of the ridge estimator is below the variance of the OLS estimator. JF Figure 13.9.

Lasso regression

We have seen that ridge regression essentially re-scales the OLS estimates. The lasso, by contrast, tries to produce a *sparse* solution, in the sense that several of the slope parameters will be set to zero.

Constrained optimization

Different from the L_2 penalty for ridge regression, the Lasso regression employs L_1 -penalty.

$$\hat{\beta}^{lasso} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2$$

$$\text{subject to } \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \leq t$$

for $t \geq 0$; which can again be re-formulated using the Lagrangian for the L_1 -penalty,

$$\hat{\beta}^{lasso} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

where $\lambda > 0$ and, as before, there exists a one-to-one correspondence between t and λ .

Parameter estimation

Contrary to ridge regression, the Lasso does not have a closed-form solution. The L_1 -penalty makes the solution non-linear in y_i 's. The above constrained minimization is a quadratic programming problem, for which many solvers exist.

Choice of Hyperparameters

Regularization parameter

The choice of λ in both ridge and lasso regressions is more of an art than a science. This parameter can be constructed as a complexity parameter, since as λ increases, less and less effective parameters are likely to be included in both ridge and lasso regressions. Therefore, one can adopt a model selection perspective and compare different choices of λ using cross-validation or an information criterion. That is, the value of λ should be chosen adaptively, in order to minimize an estimate of the expected prediction error (as in cross-validation), for instance, which is well approximated by AIC. We will discuss model selection in more detail later.

Bayesian perspective

The penalty terms in ridge and lasso regression can also be justified, using a Bayesian framework, whereby these terms arise as a result of the specification of a particular prior distribution on the vector of slope parameters.

1. The use of an L_2 -penalty in multiple regression is analogous to the choice of a Normal prior on the β_j 's, in Bayesian statistics.

$$\begin{aligned} y_i &\stackrel{iid}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_i^T \beta, \sigma^2), \quad i = 1, \dots, n \\ \beta_j &\stackrel{iid}{\sim} \mathcal{N}(0, \tau^2), \quad j = 1, \dots, p \end{aligned}$$

2. Similarly, the use of an L_1 -penalty in multiple regression is analogous to the choice of a Laplace prior on the β_j 's, such that

$$\beta_j \stackrel{iid}{\sim} \text{Laplace}(0, \tau^2), \quad j = 1, \dots, p$$

In both cases, the value of the hyperparameter, τ^2 , will be inversely proportional to the choice of the particular value for λ . For ridge regression, λ is exactly equal to the shrinkage parameter of the hierarchical model, $\lambda = \sigma^2/\tau^2$.

Model selection

Model selection is conceptually simplest when our goal is *prediction* – that is, the development of a regression model that will predict new data as accurately as possible. However, prediction is not often the only desirable characteristic in a statistical model that model interpretation, data summary and explanations are also desired. We discuss several criteria for selecting among m competing statistical models $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$ for n observations of a response variable Y and associated predictors X s.

Adjusted- R^2

The squared multiple correlation “corrected” (or “adjusted”) for degrees of freedom is intuitively reasonable criterion for comparing linear-regression models with different numbers of parameters. Suppose model M_j is one of the models under consideration. If M_j has s_j regression coefficients (including the regression constant) and is fit to a data set with n observations, then the adjusted- R^2 for the model is

$$R_{adj,j}^2 = 1 - \frac{n-1}{n-s_j} \times \frac{RSS_j}{TSS}$$

Models with relatively large numbers of parameters are penalized for their lack of parsimony. The model with the highest adjusted- R^2 value is selected as the best model. Beyond this intuitive rationale, however, there is no deep justification for using R_{adj}^2 as a model selection criterion.

Cross-validation and generalized cross-validation

The key idea in cross-validation (more accurately, leave-one-out cross-validation) is to omit the i th observation to obtain an estimate of $E(Y|x_i)$ based on the other observations as $\hat{Y}_{-i}^{(j)}$ for model M_j . Omitting the i th observation makes the fitted value $\hat{Y}_{-i}^{(j)}$ independent of the observed value Y_i . The cross-validation criterion for model M_j is

$$CV_j \equiv \frac{\sum_{i=1}^n \left[\hat{Y}_{-i}^{(j)} - Y_i \right]^2}{n}$$

We prefer the model with the smallest value of CV_j .

In linear least-squares regression, there are efficient procedures for computing the leave-one-out fitted values $\hat{Y}_{-i}^{(j)}$ that do not require literally refitting the model (recall the discussions of standardized residuals). However, in other applications, leave-one-out cross-validation can be computationally expensive (that requires literally refitting the model n times).

An alternative is to divide the data into a relatively small number of subsets of roughly equal size and to fit the model omitting one subset at a time, obtaining fitted values for all observations in the omitted subset. This method is termed as K -fold cross-validation where K is the number of subsets. The cross-validation criterion is defined the same way as before.

An alternative criterion is to approximate CV by the generalized cross-validation criterion

$$GCV_j \equiv \frac{n \times RSS_j}{df_{res_j}^2}$$

which however is less popular given the increasing computational power we have in the modern era.

AIC and BIC

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also popular model selection criteria. Both are members of a more general family of *penalized* model-fit statistics (in the form of “*IC”), applicable to regression models fit by maximum likelihood, that take the form

$$*IC_j = -2 \log_e L(\hat{\theta}_j) + cs_j$$

where $L(\hat{\theta}_j)$ is the maximized likelihood under model M_j ; $\hat{\theta}_j$ is the vector of parameters of the model (including, for example, regression coefficients and an error variance); s_j is the number of parameters in $\hat{\theta}_j$; and c is a constant that differs from one model selection criterion to another. The first term, $-2 \log_e L(\hat{\theta}_j)$, is the residual deviance under the model; for a linear model with normal errors, it is simply the residual sum of squares.

The model with the smallest *IC is the one that receives most support from the data (the selected model). The AIC and BIC are defined as follows:

$$\begin{aligned} AIC_j &\equiv -2 \log_e L(\hat{\theta}_j) + 2s_j \\ BIC_j &\equiv -2 \log_e L(\hat{\theta}_j) + s_j \log_e(n) \end{aligned}$$

The lack-of-parsimony penalty for the BIC grows with the sample size, while that for the AIC does not. When $n \geq 8$ the penalty for the BIC is larger than that for the AIC resulting in BIC tends to nominate models with fewer parameters. Both AIC and BIC are based on deeper statistical considerations, please refer to JF 22.1 sections **A closer look at the AIC** and **A closer look at the BIC** for more details.

Sequential procedures

Besides the ranking systems above, there is another class loosely defined as sequential procedures for model selection.

1. Forward selection
2. Backwards elimination
3. Stepwise selection

Forward selection :

1. Choose a threshold significance level for adding predictors, “SLENTRY” (SL stands for significance level). For example, $SLENTRY = 0.10$.
2. Initialize with $y = \beta_0 + \epsilon$.
3. Form a set of candidate models that differ from the working model by addition of one new predictor
4. Do any of the added predictors have $p - value \leq SLENTRY$?
 - Yes: add predictor with smallest p -value to working model + repeat steps 3 to 4.
 - No: stop. Final model = working model.

Backwards elimination

1. Choose threshold level for removing predictors. For example, $SLSTAY = 0.05$.
2. Initialize with most general model (biggest possible): $y = \beta_0 + \beta_1 x_1 + \dots + \epsilon$.
3. Form a set of candidate models that differ from working model by deletion of one term
4. Do any $p - value > SLSTAY$ (from fitting the current working model)?
 - Yes: remove the term with largest p -value and repeat steps 3 and 4.
 - No: stop. Final model = working model.

Stepwise Alternate forwards + backwards steps. Initialize with $y = \beta_0 + \epsilon$. Stop when consecutive forward + backward steps do not change working model. ($SLENTRY \leq SLSTAY$)

Some examples

- [Model selection by AIC](#)
- [Model selection by AIC and Lasso](#)

Additional reference

[Course notes](#) by Dr. Jason Osborne.

Analysis of Variance

The term analysis of variance is used to describe the partition of the response-variable sum of squares into “explained” and “unexplained” components, noting that this decomposition applies generally to linear models. For historical reasons, analysis of variance (abbreviated ANOVA) also refers to procedures for fitting and testing linear models in which the explanatory variables are categorical.

One-way ANOVA

Suppose that there are *no* quantitative explanatory variables, but only a single factor (categorical data). For example, for a three-category classification, we have the model

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i \quad (1)$$

employing the following coding for the dummy regressors:

Group	D_1	D_2
1	1	0
2	0	1
3	0	0

The expectation of the response variable in each group (i.e. in each category or level of the factor) is the population group mean, denoted by μ_j for the j th group. Equation 1 produces the following relationship between group means and model parameters:

$$\text{Group 1: } E(Y_i | D_{i1} = 1, D_{i2} = 0) = \alpha + \gamma_1 \times 1 + \gamma_2 \times 0 = \alpha + \gamma_1$$

$$\text{Group 2: } E(Y_i | D_{i1} = 0, D_{i2} = 1) = \alpha + \gamma_1 \times 0 + \gamma_2 \times 1 = \alpha + \gamma_2$$

$$\text{Group 3: } E(Y_i | D_{i1} = 0, D_{i2} = 0) = \alpha + \gamma_1 \times 0 + \gamma_2 \times 0 = \alpha$$

There are three parameters (α , γ_1 and γ_2) and three group means, so we can solve uniquely for the parameters in terms of the group means:

$$\alpha = \mu_3$$

$$\gamma_1 = \mu_1 - \mu_3$$

$$\gamma_2 = \mu_2 - \mu_3$$

Not surprisingly, α represents the mean of the baseline category (Group 3) and that γ_1 and γ_2 captures differences between the other group means and the mean of the baseline category.

notations

Because observations are partitioned according to groups, it is convenient to let Y_{jk} denote the k th observation within the j th of m groups. The number of observations in the j th group is n_j , and the total number of observations is $n = \sum_{j=1}^m n_j$. Let $\mu_j \equiv E(Y_{jk})$ be the population mean in group j .

The one-way ANOVA model is

$$Y_{jk} = \mu + \alpha_j + \epsilon_{jk}$$

where μ represents the general level of response variable in the population; α_j represents the effect on the response variable of membership in the j th group; ϵ_{jk} is an error variable that follows the usual linear-model assumptions: $\epsilon_{jk} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

By taking expectations, we have

$$\mu_j = \mu + \alpha_j$$

The parameters of the model are, therefore, underdetermined, for there are $m + 1$ parameters (including μ) but only m population group means (recall the dummy variable trap introduced in collinearity). To produce easily interpretable parameters and that estimates and generalizes usefully to more complex models, we impose the sum-to-zero constraint

$$\sum_{j=1}^m \alpha_j = 0$$

With the sum-to-zero constraint, we solve for the parameters

$$\begin{aligned}\hat{\mu} &= \frac{\sum \tilde{\mu}_j}{m} \\ \hat{\alpha}_j &= \tilde{\mu}_j - \hat{\mu}\end{aligned}$$

where $\tilde{\mu}_j$ represents the sample group mean for group j .

The fitted Y values are the group means for the one-way ANOVA model:

$$\hat{Y}_{jk} = \hat{\mu} + \hat{\alpha}_j$$

and the regression and residual sums of squares therefore take particularly simple forms in one-way ANOVA:

$$\begin{aligned}RegSS &= \sum_{j=1}^m \sum_{k=1}^{n_j} (\hat{Y}_{jk} - \bar{Y})^2 = \sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2 \\ RSS &= \sum_{j=1}^m \sum_{k=1}^{n_j} (Y_{jk} - \hat{Y}_{jk})^2 = \sum_{j=1}^m \sum_{k=1}^{n_j} (Y_{jk} - \bar{Y}_j)^2\end{aligned}$$

and can be presented in an ANOVA table.

Table 1: General one-way ANOVA table

Source	Sum of Squares	df	Mean Square	F	H_0
Groups	$\sum n_j (\bar{Y}_j - \bar{Y})^2$	$m - 1$	$\frac{RegSS}{m-1}$	$\frac{RegMS}{RMS}$	$\alpha_1 = \dots = \alpha_m = 0$
Residuals	$\sum \sum (Y_{jk} - \bar{Y}_j)^2$	$n - m$	$\frac{RSS}{n-m}$		
Total	$\sum \sum (Y_{jk} - \bar{Y})^2$	$n - 1$			

Sometimes, the column of Source can also be denoted with Treatments (for Groups) and Error (for Residuals). And a balanced one-way ANOVA model has the same number of observations in one group (or treatment), in other words, $n_1 = \dots = n_m = \frac{n}{m}$.

one-way ANOVA example

The following data come from study investigating binding fraction for several antibiotics using $n = 20$ bovine serum samples:

Antibiotic	Binding Percentage	Sample mean
Penicillin G	29.6 24.3 28.5 32.0	28.6
Tetracyclin	27.3 32.6 30.8 34.8	31.4
Streptomycin	5.8 6.2 11.0 8.3	7.8
Erythromycin	21.6 17.4 18.3 19	19.1
Chloramphenicol	29.2 32.8 25.0 24.2	27.8

Question: Are the population means for these 5 treatments plausibly equal?

Answer:

How do we obtain standard errors of parameter estimates? (HW)