

23 Lecture 23: March 24

Last time

- Midterm exam review
- Biased estimation (JF 13.2.3, CG's notes)
 - Ridge regression

Today

- Midterm evaluation suggestions
- Poll on alternative grade path
- Lasso regression (CG's notes)
- Model selection (JF 22)

Additional reference

[Lecture notes](#) by Cedric Ginestet

Midterm evaluation suggestions

1. What has been most helpful for your learning in this class so far?
 - lecture notes
 - lab session
2. What has caused you the most difficulty in terms of learning in this class so far?
 - Prior knowledge
 - R
 - More explanation in writing besides talking
 - prewritten slides
 - disconnection between lecture notes and homework (XJ: not really sure)
3. What suggestion(s) do you have that would enhance your learning experience in this class?
 - more examples
 - more R practice
 - partially filled slides and questions
 - more comments on homework grading (XJ: probably redundant with in-class reviews)

4. Please share any additional comments about the content, format or instructor that you have about the course. Your comments are appreciated!

- appreciate TA and instructor's effort
- It's cool that we get to see student presentations throughout the semester instead of all at once during finals week like other classes.
- I know that the class is meant to move at a fast pace, but I feel I am struggling to stay on top of everything.

Alternative grade path proposal

The alternative grade path would be an addition to the original grading scheme. Let $S_{original} = w_{homework}S_{homework} + w_{mid-term}S_{mid-term} + w_{final}S_{final} + w_{presentation}S_{presentation}$ represent the final grade from the original grading scheme. Then the proposed alternative grading scheme has $S_{alternative} = w_{homework}S_{homework} + (w_{mid-term} + w_{final})S_{final} + w_{presentation}S_{presentation}$. And the total score will be $S_{total} = \max\{S_{original}, S_{alternative}\}$.

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.

$$\begin{aligned}\hat{\beta}^{lasso} &= \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 \\ \text{subject to } \|\beta\|_1 &= \sum_{j=1}^p |\beta_j| \leq t\end{aligned}$$

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.

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

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](#)