Ridge regression

Patrick Breheny

February 2

Introduction

- Large-scale testing is, of course, a big area and we could keep talking about it
- However, for the rest of the course we will take up the issue of high-dimensional regression: using the features to predict/explain the outcome
- As we saw in our first lecture, ordinary least squares is problematic in high dimensions
- Reducing the dimensionality through model selection allows for some progress, but has several shortcomings

Likelihood and loss

- More broadly speaking, this can be seen as a failure of likelihood-based methods
- In this course, we will use the notation L to refer to the negative log-likelihood:

$$\begin{aligned} L(\theta|\text{Data}) &= -\log \ell(\theta|\text{Data}) \\ &= -\log p(\text{Data}|\theta) \end{aligned}$$

- Here, L is known as the *loss function* and we seek estimates with a low loss; this is equivalent to finding a value (or interval of values) with a high likelihood

Likelihood for linear regression

- In the context of linear regression, the loss function is

$$L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

- It is only the difference in loss functions between two values, $L(\boldsymbol{\beta}_1|\mathbf{X}, \mathbf{y}) - L(\boldsymbol{\beta}_2|\mathbf{X}, \mathbf{y})$, i.e., the likelihood ratio, that is relevant to likelihood-based inference; thus, the first term may be ignored
- For the purposes of finding the MLE, the $1/(2\sigma^2)$ factor in the second term may also be ignored, although we must account for it when constructing likelihood-based intervals

Penalized likelihood

Given the aforementioned problems with likelihood methods, consider instead the following modification:

$$Q(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) + P_\lambda(\boldsymbol{\beta}),$$

where

- P is a *penalty function* that penalizes what one would consider less realistic values of the unknown parameters
- λ is a *regularization parameter* that controls the tradeoff between the two components
- The combined function Q is known as the *objective function*

Meaning of the penalty

- What exactly do we mean by "less realistic" values?
- The most common use of penalization is to impose the belief that small regression coefficients are more likely than large ones; i.e., that we would not be surprised if β_j was 1.2 or 0.3 or 0, but would be very surprised if β_j was 9.7×10^4
- Later in the course, we consider other uses for penalization to reflect beliefs that the true coefficients may be grouped into hierarchies, or display a spatial pattern such that β_j is likely to be close to β_{j+1}

Remarks

- Some care is needed in the application of the idea that small regression coefficients are more likely than large ones
- First of all, it typically does not make sense to apply this line of reasoning to intercept; hence β_0 is not included in the penalty
- Second, the size of the regression coefficient depends on the scale with which the associated feature is measured; depending on the units $\mathbf{x}_j$ is measured in, $\beta_j = 9.7 \times 10^4$ might, in fact, be realistic

Standardization

- This is a particular problem if different features are measured on different scales, as the penalty would not have an equal effect on all coefficient estimates
- To avoid this issue and ensure invariance to scale, features are usually *standardized* prior to model fitting to have mean zero and standard deviation 1:

$$\sum_{i=1}^{n} x_{ij} = 0$$

$$\sum_{i=1}^{n} x_{ij}^2 = n \qquad \text{for all } j$$

- This can be accomplished without any loss of generality, as any location shifts for $\mathbf{X}$ are absorbed into the intercept and scale changes can be reversed after the model has been fit

Added benefits of standardization

Centering and scaling the explanatory variables has added benefits in terms of computational savings and conceptual simplicity:

- The features are now orthogonal to the intercept term, meaning that in the standardized covariate space, $\widehat{\beta}_0 = \bar{y}$ regardless of the rest of the model
- Also, standardization simplifies the solutions; to illustrate with simple linear regression,

$$\begin{aligned} \widehat{\beta}_0 &= \bar{y} - \widehat{\beta}\bar{x} \\ \widehat{\beta}_1 &= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \end{aligned}$$

However, if we center and scale $\mathbf{x}$ and center $\mathbf{y}$, then we get the much simpler expression $\widehat{\beta}_0 = 0$, $\widehat{\beta}_1 = \mathbf{x}^\top \mathbf{y}/n$

Ridge regression: Penalty

- If penalized regression is to impose the assumption that small regression coefficients are more likely than large ones, we should choose a penalty that discourages large regression coefficients
- A natural choice is to penalize the sum of squares of the regression coefficients:

$$P_\tau(\boldsymbol{\beta}) = \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2$$

- Applying this penalty in the context of penalized regression is known as *ridge regression*, and has a long history in statistics, dating back to 1970

Objective function

- The ridge regression objective function is

$$Q(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \frac{1}{2\sigma^2}\sum_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \frac{1}{2\tau^2}\sum_{j=1}^p \beta_j^2$$

- It is often convenient to multiply the above objective function by σ^2/n; as we will see, doing so tends to simplify the expressions involved in penalized regression:

$$Q(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \frac{1}{2n}\sum_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \frac{\lambda}{2}\sum_{j=1}^p \beta_j^2,$$

where $\lambda = \sigma^2/(n\tau^2)$

Solution

- For linear regression, the ridge penalty is particularly attractive to work with because the maximum penalized likelihood estimator has a simple closed form solution
- This objective function is differentiable, and it is straightforward to show that its minimum occurs at

$$\widehat{\boldsymbol{\beta}} = (n^{-1}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1} n^{-1}\mathbf{X}^\top\mathbf{y}$$

- The solution is similar to the least squares solution, but with the addition of a "ridge" down the diagonal of the matrix to be inverted
- Note that the ridge solution is a simple function of the marginal OLS solutions $n^{-1}\mathbf{X}^\top\mathbf{y}$ and the correlation matrix $n^{-1}\mathbf{X}^\top\mathbf{X}$

Orthonormal solutions

- To understand the effect of the ridge penalty on the estimator $\widehat{\beta}$, it helps to consider the special case of an orthonormal design matrix ($\mathbf{X}^\top \mathbf{X}/n = \mathbf{I}$)
- In this case,

$$\widehat{\beta}_j = \frac{\widehat{\beta}_j^{\mathrm{OLS}}}{1+\lambda}$$

- This illustrates the essential feature of ridge regression: *shrinkage*; i.e., the primary effect of applying ridge penalty is to shrink the estimates toward zero

Simple example

- The benefits of ridge regression are most striking in the presence of multicollinearity
- Consider the following very simple simulated example:

```
> x1 <- rnorm(20)
> x2 <- rnorm(20, mean=x1, sd=.01)
> y <- rnorm(20, mean=3+x1+x2)
> lm(y ~ x1 + x2)
...
(Intercept)          x1          x2
   3.021159   21.121729  -19.089170
```

- Although there are only two covariates, the strong correlation between X_1 and X_2 causes a great deal of trouble for maximum likelihood

Ridge regression for the simple example

- The problem here is that the likelihood surface is very flat along $\beta_1 + \beta_2 = 2$, leading to tremendous uncertainty
- When we introduce the added assumption that small coefficients are more likely than large ones by using a ridge penalty, however, this uncertainty is resolved:

```
> ridge(y ~ x1 + x2, lambda = 0.1)
(Intercept)          x1          x2
  3.0327231   0.9575176   0.9421784
```

Ridge regression always has unique solutions

- The maximum likelihood estimator is not always unique: If $\mathbf{X}$ is not full rank, $\mathbf{X}^\top\mathbf{X}$ is not invertible and an infinite number of β values maximize the likelihood
- This problem does not occur with ridge regression
- **Theorem:** For any design matrix $\mathbf{X}$, the quantity $n^{-1}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is always invertible provided that $\lambda > 0$; thus, there is always a unique solution $\widehat{\beta}$.

Is ridge better than maximum likelihood?

- In our simple example from earlier, the ridge regression estimate was much closer to the truth than the MLE
- An obvious question is whether ridge regression estimates are systematically closer to the truth than MLEs are, or whether that example was a fluke
- To address this question, let us first derive the bias and variance of ridge regression

Bias and variance

- The variance of the ridge regression estimate is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \frac{\sigma^2}{n}\mathbf{W}(\tfrac{1}{n}\mathbf{X}^\top\mathbf{X})\mathbf{W},$$

  where $\mathbf{W} = (\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$
- Meanwhile, the bias is

$$\mathrm{Bias}(\widehat{\boldsymbol{\beta}}) = -\lambda\mathbf{W}\boldsymbol{\beta}$$

- Both bias and variance contribute to overall accuracy, as measured by mean squared error:

$$\begin{aligned}\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) &= \mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \\ &= \sum_j \mathrm{Var}(\widehat{\beta}_j) + \sum_j \mathrm{Bias}(\widehat{\beta}_j)^2\end{aligned}$$

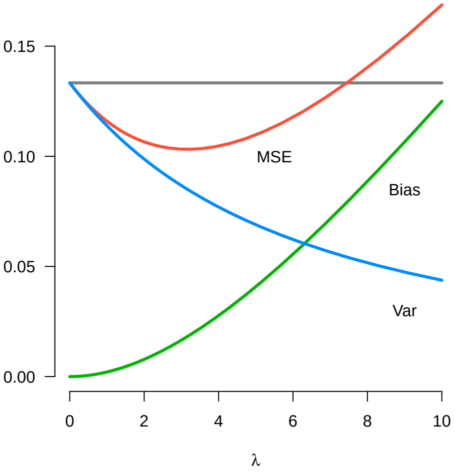
Existence theorem

- So, is ridge regression better than maximum likelihood (OLS)?
- **Theorem:** There always exists a value λ such that

$$\mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}_\lambda\right) < \mathrm{MSE}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}\right)$$

- This is a rather surprising result with somewhat radical implications: despite the typically impressive theoretical properties of maximum likelihood and linear regression, we can *always* obtain a better estimator by shrinking the MLE towards zero

Sketch of proof



Bayesian justification for the penalty

- From a Bayesian perspective, one can think of the penalty as arising from a formal prior distribution on the parameters
- Let $p(\mathbf{y}|\boldsymbol{\beta})$ denote the distribution of $\mathbf{y}$ given $\boldsymbol{\beta}$ and $p(\boldsymbol{\beta})$ the prior for $\boldsymbol{\beta}$; then the posterior density is

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}),$$

or

$$\log p(\boldsymbol{\beta}|\mathbf{y}) = \log p(\mathbf{y}|\boldsymbol{\beta}) + \log p(\boldsymbol{\beta}) + \text{constant}$$

on the log scale; this is exactly the generic form of a penalized likelihood

Ridge regression from a Bayesian perspective

- By optimizing the objective function, we are finding the mode of the posterior distribution of $\boldsymbol{\beta}$; this is known as the *maximum a posteriori*, or MAP, estimate
- Specifically, suppose that we assume the prior

$$\beta_j \overset{\perp\!\!\!\perp}{\sim} N(0, \tau^2);$$

  the resulting log-posterior is exactly the ridge regression objective function (up to a constant)
- Furthermore,
  - The ridge regression estimator $\widehat{\boldsymbol{\beta}}$ is the posterior mean (in addition to being the posterior mode)
  - The regularization parameter λ is the ratio of the prior precision $(1/\tau^2)$ to the information (n/σ^2)

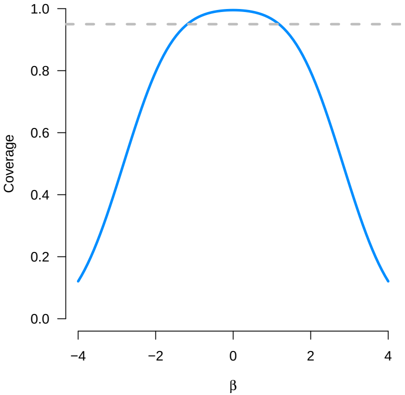
Similarities and differences

- Thus, we arrive at the same estimator $\widehat{\beta}$ whether we view it as a modified maximum likelihood estimator or a Bayes estimator
- In other inferential respects, however, the similarity between Bayesian and Frequentist breaks down
- Two aspects, in particular, are worthy of mention

Properties of intervals

- First is the inferential goal of constructing intervals for β and what properties such intervals should have
- Frequentist confidence intervals are required to maintain a certain level of coverage for any fixed value of β
- Bayesian posterior intervals, on the other hand, may have much higher coverage at some values of β than others

Properties of intervals (cont'd)



- Bayes coverage for a 95% posterior interval at $\beta_j \approx 0$ is $> 99\%$, but only $\approx 20\%$ for $\beta_j \approx 3.5$
- The interval nevertheless maintains 95% coverage across a collection of β_j values, integrated with respect to the prior

Point properties at 0

- The other aspect in which a clear divide emerges between Bayes and Frequentist perspectives is with regard to the specific value $\beta = 0$
- From a Bayesian perspective, the posterior probability that $\beta = 0$ is 0 because its posterior distribution is continuous
- From a Frequentist perspective, however, the notion of testing whether $\beta = 0$ is still meaningful and indeed, often of interest in an analysis

Final remarks

- The penalized regression literature generally adopts the perspective of maximum likelihood theory, although the appearance of a penalty in the likelihood somewhat blurs the lines between Bayes and Frequentist ideas
- The majority of research into penalized regression methods has focused on point estimation and its properties, so these inferential differences between Bayesian and Frequentist perspectives are relatively unexplored
- Nevertheless, developing inferential methods for penalized regression is an active area of current research, and we will come back to some of these issues when we discuss inference for high-dimensional models