Penalized regression: Introduction

Patrick Breheny

August 30

Maximum likelihood

Much of 20th-century statistics dealt with maximum likelihood estimation:

- One begins by specifying a distribution $f(\mathbf{x}|\theta)$ for the data
- This distribution in turn induces a log-likelihood function $L(\theta|\mathbf{x})$, which specifies how likely the observed data is for various values of the unknown parameters
- Estimates can be found by finding the most likely values of the unknown parameters
- Hypothesis tests and confidence intervals can be carried out by evaluating how the likelihood changes in response to moving away from these most likely values

Problems with maximum likelihood in regression

Maximum likelihood estimation has many wonderful properties, but is often unsatisfactory in regression problems for two reasons:

- Large variability: when p is large with respect to n, or when columns of $\mathbf{X}$ are highly correlated, the variance of $\hat{\beta}$ is large
- Lack of interpretability: if p is large, we often desire a smaller set of predictors in order to gain insight into the most relevant relationships between y and $\mathbf{X}$

Subset selection

- Probably the most common solution to this problem is choose a subset of important explanatory variables based on ad hoc trial and error using p-values to guide whether a variable should be in the model or not
- However, this approach is impossible to replicate and therefore to study the statistical properties of
- A more systematic approach is to exhaustively search all possible subsets and then use some sort of criterion such as AIC or BIC to choose an optimal subset

Problem #1: Computational infeasibility

- One problem with this approach is that the number of "all possible subsets" grows exponentially with p
- With today's computers, searching all possible subsets is computationally infeasible for p much larger than 40 or 50
- For this reason, modifications such as forward selection, backward selection, and forward/backward hybrids have been proposed

Problem #2: Instability

- Another problem is the fact that subset selection is discontinuous, in the sense that an infinitesimally small change in the data can result in completely different estimates
- As a result, subset selection is often unstable and highly variable, especially in high dimensions
- As we will see, penalized regression allows us to accomplish the same goals as subset selection, but in a more stable, continuous, and computationally efficient fashion

Subset selection pitfalls

- It should be pointed out that it is common practice to report inferential results from ordinary least squares models following subset selection as if the model had been prespecified from the beginning
- This is unfortunate, as the resulting inferential procedures violate every principle of statistical estimation and hypothesis testing:
  - Test statistics no longer follow t/F distributions
  - Standard errors are biased low, and confidence intervals falsely narrow
  - p-values are falsely small
  - Regression coefficients are biased away from zero

Penalized maximum likelihood

A different way of dealing with this problem is to introduce a penalty: instead of maximizing $\ell(\theta|\mathbf{x}) = \log\{L(\theta|\mathbf{x}\}$, we maximize the function

$$M(\theta) = \ell(\theta|\mathbf{x}) - \lambda P(\theta)$$

where

- P is a function that penalizes what one would consider less realistic values of the unknown parameters
- λ controls the tradeoff between the two parts
- The function M is called the *objective function*

Penalized maximum likelihood (cont'd)

In regression, we usually think about minimizing a loss function (usually squared error loss) rather than maximizing likelihood; an equivalent formulation is that we estimate θ by minimizing

$$M(\theta) = L(\theta|\mathbf{x}) + \lambda P(\theta)$$

where L here is a loss function (usually a quantity that is proportional to a negative log likelihood, such as the residual sum of squares)

Penalty functions

- What exactly do we mean by "less realistic" values?
- In the first section of this course, we mean that coefficient values around zero are more believable than those far away from zero, and P is therefore a function which penalizes coefficients as they get further away from zero
- The two penalties that we will cover in depth in this section of the course are ridge regression and the lasso:

$$\text{Ridge:} \quad P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$$

$$\text{Lasso:} \quad P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$$

Penalty functions (cont'd)

- Later in the course, we will use the idea of penalization when fitting curves to data, to favor simple, smoother curves over wiggly, erratic ones; here, P is a measure of the "wiggliness" of the curve
- There are also plenty of other varieties of penalization in existence that we will not address in this course
- The all operate on the same basic principle, however: all other things being equal, we would tend to favor the simpler explanation over the more complicated one

The regularization parameter

- As mentioned earlier, the parameter λ controls the tradeoff between the penalty and the fit (loss/likelihood):
  - When λ is too small, we tend to overfit the data and end up models that have high variance
  - When λ is too large, we tend to underfit the data and end up with models that are too simplistic and thus potentially biased
- The parameter λ is called the *regularization* parameter; changing the regularization parameter allows us to directly balance the bias-variance tradeoff
- Obviously, selection of λ is a very important practical aspect of fitting these models

Bayesian interpretation

- From a Bayesian perspective, one can think of the penalty as arising from a prior distribution on the parameters
- The objective function is then (proportional to) the log of the posterior distribution of θ given the data
- By optimizing this objective function, we are finding the mode of the posterior distribution of θ

Constrained regression

- Yet another way to think about penalized regression is that they imply a constraint on the values of θ
- Suppose we were trying to maximize the likelihood subject to the constraint that $P(\theta) \leq t$
- A standard approach to solving such problem is to introduce a *Lagrange multiplier*; when we do so, we arrive at the same objective function as earlier

Penalization and the intercept term

- Earlier, we mentioned that penalized regression revolves around an assumption that coefficient values around zero are more believable than those far away from zero
- Some care is needed, however, in the application of this idea
- First of all, it does not make sense to apply this idea to the intercept, unless you happened to have some reason to think that the mean of y should be zero
- Hence, the intercept is not included in the penalty; if it were, the estimates would not be invariant to changes of location

Standardization

- A separate consideration is how to make "far from zero" mean the same thing for all the variables
- For example, suppose x_1 varied from 0 to 1, while x_2 varied from 0 to 1 million; clearly, a one-unit change in x does not mean the same for both of these variables
- Thus, the explanatory variables are usually *standardized* prior to model fitting to have mean zero and standard deviation 1; *i.e.*,

$$\bar{x}_j = 0$$
$$\mathbf{x}_j^T \mathbf{x}_j = n$$

for all j

Standardization (cont'd)

This can be accomplished without any loss of generality:

- Any location shifts for $\mathbf{X}$ are absorbed into the intercept
- Scale changes can be reversed after the model has been fit:

$$\begin{aligned} x_{ij}\beta_j &= \frac{x_{ij}}{a} a\beta_j \\ &= \tilde{x}_{ij}\tilde{\beta}_j; \end{aligned}$$

i.e., if we had to divide $\mathbf{x}_j$ by a to standardize it, we simply divide the transformed solution $\tilde{\beta}_j$ by a to obtain β_j on the original scale

Further benefits

Centering and scaling the explanatory variables has added benefits:

- The explanatory variables are now orthogonal to the intercept term, meaning that in the standardized covariate space, $\hat{\beta}_0 = \bar{y}$ regardless of what goes on in the rest of the model
- In other words, if we center y by subtracting off its mean, we don't even need to estimate β_0
- Also, standardization simplifies the solutions; recall from BST 760 that for simple linear regression

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

- If we center and scale x and center y, however, then we get the much simpler expression $\hat{\beta} = \mathbf{x}^T\mathbf{y}/n$

Standardization summary

To summarize, centering the response and centering and scaling each explanatory variable to have mean 0 and standard deviation 1 has the following benefits:

- Estimates of $\boldsymbol{\beta}$ are location-scale invariant
- Computational savings (we only need to estimate p parameters, not $p+1$)
- Simplicity
- No loss of generality, as we can transform back to the original scale once we finish fitting the model