Statistics 203: Introduction to Regression and Analysis of Variance

Penalized models

Jonathan Taylor

Today's class

- Bias-Variance tradeoff.
- Penalized regression.
- Cross-validation.

Bias-variance tradeoff

- Arguably, the goal of a regression analysis is to "build" a model that predicts well.
- We saw in model selection that C_p and AIC were trying to estimate the MSE of each model which included some bias.
- One way to measure this performance is in the population mean squared-error of the model

$$MSE_{pop}(\mathcal{M}) = \mathbb{E}\left((Y_{new} - (\widehat{\beta}_0 + \sum_{j=1}^{p-1} \widehat{\beta}_j X_{new,j}))^2 \right)$$

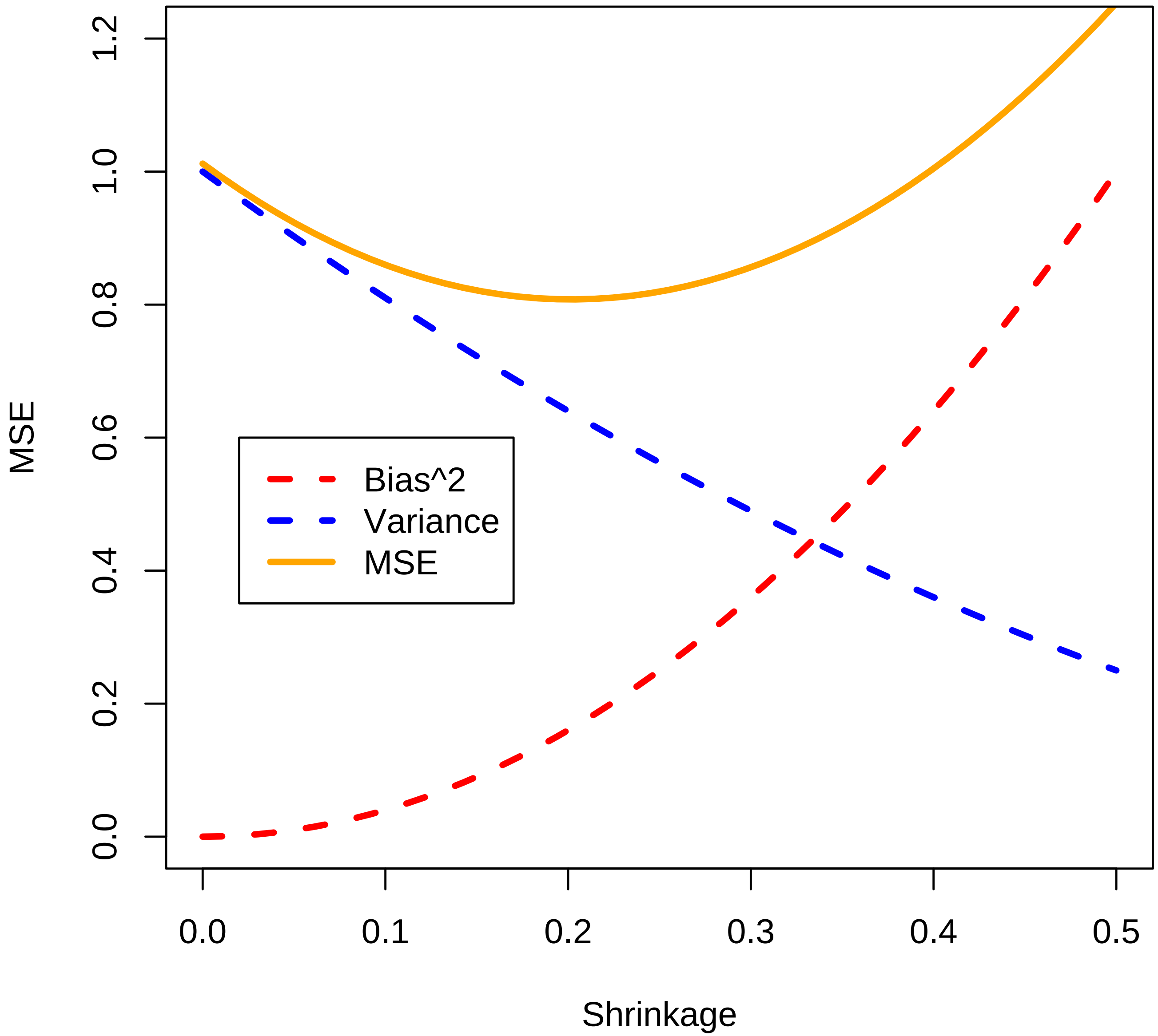
$$= \text{Var}(Y_{new} - (\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_{new,j})) +$$

$$\text{Bias}(\widehat{\beta})^2.$$

Bias-variance tradeoff

- In choosing a model automatically, even if the "full" model is correct (unbiased) our resulting model may be biased – a fact we have ignored so far.
- Inference (F, χ^2 tests, etc) is not quite exact for biased models.
- Sometimes, it is possible to find a model with lower MSE than an unbiased model! This is called the "bias-variance tradeoff."
- It is "generic" in statistics: almost always introducing some bias yields a decrease in MSE, followed by an later increase.

Bias-Variance Tradeoff



Shrinkage & Penalties

- Shrinkage can be thought of as "constrained" minimization.
- Minimize

$$\sum_{i=1}^n (Y_i - \mu)^2 \quad \text{subject to } \mu^2 \leq C$$

- Lagrange: equivalent to minimizing

$$\sum_{i=1}^n (Y_i - \mu)^2 + \lambda_C \mu^2$$

- Differentiating:

$$-2\sum_{i=1}^n (Y_i - \widehat{\mu}_C) + 2\lambda_C \widehat{\mu}_C = 0$$

- Finally

$$\widehat{\mu}_C = \frac{\sum_{i=1}^n Y_i}{n + \lambda_C} = K_C \overline{Y}, \qquad K_C < 1.$$

Shrinkage & Penalties

- The precise form of λ_C is unimportant: as $C \to 0$,

$$\widehat{\mu}_C \to \overline{Y}.$$

- As $C \to \infty$

$$\widehat{\mu}_C \to 0.$$

Penalties & Priors

- Minimizing
$$\sum_{i=1}^n (Y_i - \mu)^2 + \lambda \mu^2$$
is similar to computing "MLE" of μ if the likelihood was proportional to
$$\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (Y_i - \mu)^2 + \lambda \mu^2\right)\right).$$
- This is not a likelihood function, *but* it is a posterior density for μ if μ has a $N(0, \sigma^2/\lambda)$ prior.
- Hence, penalized estimation with this penalty is equivalent to using the MAP (Maximum A Posteriori) estimator of μ with a Gaussian prior.

Biased regression: penalties

- Not all biased models are better – we need a way to find "good" biased models.
- Generalized one sample problem: penalize large values of β. This should lead to "multivariate" shrinkage of the vector β.
- Heuristically, "large β" is interpreted as "complex model". Goal is really to penalize "complex" models, i.e. Occam's razor.
- Equivalent Bayesian interpretation.
- If truth really is complex, this may not work! (But, it will then be hard to build a good model anyways ... (statistical lore))

Ridge regression

- Assume that columns $(X_j)_{1 \leq j \leq p-1}$ have zero mean, and length 1 (to distribute the penalty equally – not strictly necessary) and Y has zero mean, i.e. no intercept in the model.
- This is called the *standardized model*.
- Minimize

$$SSE_\lambda(\beta) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2.$$

- Corresponds (through Lagrange multiplier) to a quadratic constraint on β's. LASSO, another penalized regression uses $\sum_{j=1}^{p-1} |\beta_j|$.
- Normal equations

$$\frac{\partial}{\partial \beta_l} SSE_\lambda(\beta) = -2\langle Y - X\beta, X_l \rangle + 2\lambda\beta_l$$

Solving the normal equations

- $$-2\langle Y - X\widehat{\beta}_\lambda, X_l\rangle + 2\lambda\widehat{\beta}_{l,\lambda} = 0, \qquad 1 \leq l \leq p-1$$
- In matrix form
$$-Y^tX + \widehat{\beta}^t_\lambda(X^tX + \lambda I) = 0$$
- Or
$$\widehat{\beta}_\lambda = (X^tX + \lambda I)^{-1}X^tY.$$
- This is identical to the previous $\widehat{\mu}_C$ in matrix form.
- Essentially equivalent to putting a $N(0, CI)$ prior on the standardized coefficients.

LASSO regression

- Another “popular” penalized regression technique.
- Use the standardized model.
- Minimize

$$SSE_\lambda(\beta) = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|.$$

- Corresponds (through Lagrange multiplier) to an ℓ^1 constraint on β's. In theory, it works well when many β_j's are 0 and gives “sparse” solutions unlike ridge.
- Corresponds to a Laplace prior on standardized coefficients.

Choosing λ: cross-validation

- If we knew MSE as a function of λ then we would simply choose the λ that minimizes MSE.
- To do this, we need to estimate MSE.
- A popular method is "cross-validation." Breaks the data up into smaller groups and uses part of the data to predict the rest.
- We saw this in diagnostics: i.e. Cook's distance measured the fit with and without each point in the data set.

Generalized Cross Validation

- A computational shortcut for n-fold cross-validation (also known as leave-one out cross-validation). Later, we will talk about K-fold cross-validation.
- Let

$$S_\lambda = (X^t X + \lambda I)^{-1} X^t$$

be the matrix in ridge regression.

- Then

$$GCV(\lambda) = \frac{\|Y - S_\lambda Y\|^2}{n - \text{Tr}(S_\lambda)}.$$

Effective degrees of freedom

- If $\lambda = 0$ then S_λ is a projection matrix onto a $p-1$ dimensional space so

$$n - \mathrm{Tr}(S_0) = n - p + 1$$

is basically the degrees of freedom in the error, ignoring the fact we have forgotten the intercept term.

- For any linear "smoother"

$$\widehat{Y} = SY$$

the quantity

$$n - \mathrm{Tr}(S)$$

can therefore be thought of as the *effective* degrees of freedom.