📖

Statistics Unit 3

Topic: Ridge Regression

- Date: @March 18, 2024
- Lane: *Optimization, Statistics*

What it covers

- What is linear regression?
- An optimization problem: Minimizing mean squared error
- Variance, errors & diagnostics
- Revisting Bias- variance tradeoffs.

Requires

Random variables, expectation, variance, projection operators.

Required by

Testing, Estimation

Density estimation

What is ridge regression?

We revisit the basic problem in science - recover a function (an unobservable) of a set of *explanatory* ("independent") variables, ***X,*** for an *outcome* ("dependent") variable, ***Y.*** Then the function can then be applied to a new point, ***x*** for either an "interpolation" or an "extrapolation."

Ridge regression improves on general regression methods, by improving accuracy and reducing the variance of coefficient estimates. It is one variety of "penalized" regression - a modification of the objective function that the regression optimizes. It introduces a bias in the regression coefficients, "shrinking" them towards zero. Surprisingly, even though the coefficients are no longer "correct", this tends to improve the prediction accuracy, especially for small data samples

Shrinkage: Adding a penalty to mean squared error

This is accomplished by adding a penalty term to the error term that is minimized, that trades off the minimization of the *mean squared error* with a term for the magnitude of the coefficients. Thus the objective function becomes:

$$\min_{\beta} ||\hat{y} - y||^2 + \lambda ||\beta||^2$$

Note that the first term is the same as in linear regression, where $\hat{y}$ is a function of the betas. The lambda is a free parameter whose optimal value is determined by experimentation. This typically is done by running the regression over various values of lambda and comparing the accuracy across values.

The ridge regression solution

Rewriting the objective function as a function of lambda in matrix form gives:

$$\mathrm{RSS}(\lambda) = (y - \boldsymbol{X}\beta)^T(y - \boldsymbol{X}\beta) + \lambda\beta^T\beta$$

This new objective function can be solved analytically for the betas

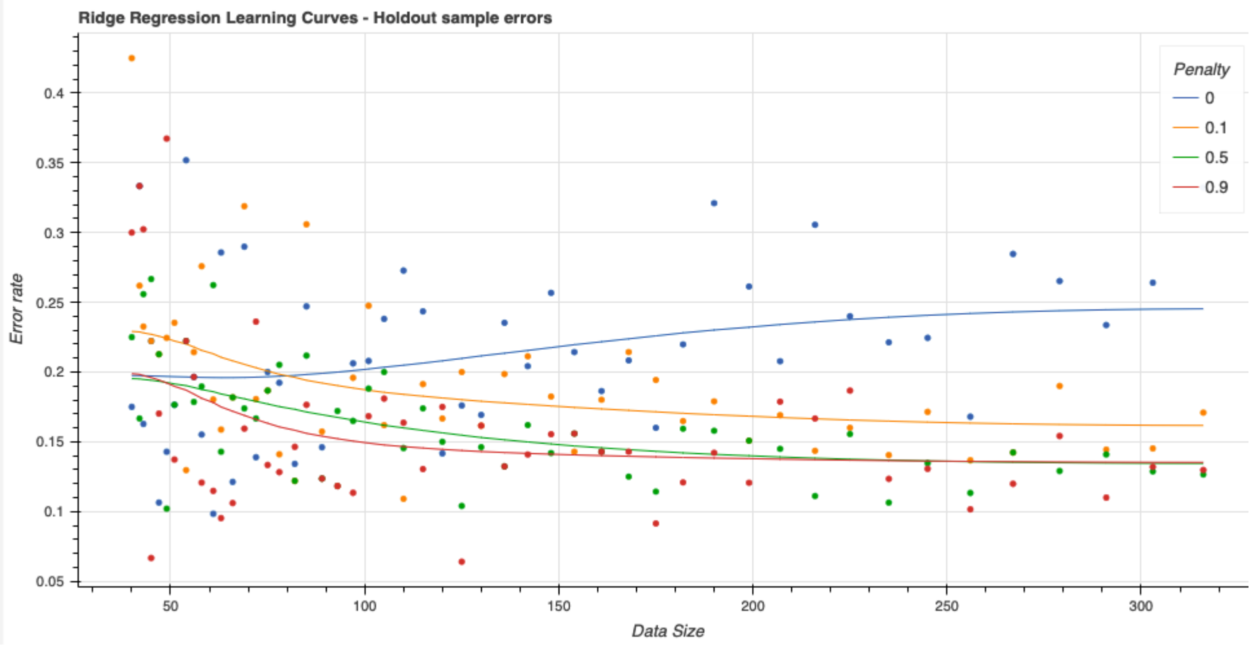
$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T y$$

As before the prediction becomes $\hat{y} = \boldsymbol{X}\hat{\beta}$.

$$\sigma^2(\beta) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2(\mathrm{RSS})$$

Learning curves - an example of measuring accuracy

Learning curves plot the accuracy of the regression as the size of the sample - the number of rows in the data matrix - increases. In general we expect accuracy to increase as the amount of data used increases. However this is not always the case, as the learning curves from this experiment reveals. It shows four learning curves, with successfully greater lambda penalties. As might be expected, both increasing the penalty value and providing more data, on average, tends to increase accuracy towards a limit. But surprisingly this is not true for linear regression — the zero penalty curve, shown in blue. It is not always the case the linear regression fails like this, but this example serves as a cautionary tale.



Shrinkage - as reduction in variance

The term “shrinkage” is used both to describe the bias “downward”, towards zero that the penalty term causes in the beta coefficients, but also the reduction in variance of the coefficients. So even though the coefficients loose their properties as the “true” coefficients in terms of their interpretation as sensitivity measures to the outcome, the are in a sense “better” especially when it comes to making predictions.

Shrinkage can also be used to select which variables to keep and which to remove from the regression. It is possible that as more penalty is applied, some coefficients “shrink” more than others, and their importance (or lack of importance) in the regression becomes evident.

When is Bias good: “Bias - variance” tradeoff.

Recall this expansion from the previous unit:

$$\sigma^2(x) = \sigma^2(\epsilon) + \mathrm{Bias}^2(\hat{f}(x)) + \sigma^2(\hat{f}(x))$$

Ridge regression is a clear example of how the “bias - variance “ tradeoff applies in regression. It has improved variance at the cost of adding some bias. In the limit of “infinite” data variance would eventually decrease and the tradeoff would no longer apply — one could have the best of both worlds. But in practice penalization methods are widely appllied, not only for linear regression, but for all predictive methods in machine learning.

Readings

Specific chapters on regularization and ridge regression:

- Boyd Ch. 15.4 Regularized data fitting.
- ISLP_python, ch 6: Linear models, subset selection and regularization.