RECSM Summer School:

Machine Learning for Social Sciences

Session 1.4:
Ridge Regression

Reto Wüest

Department of Political Science and International Relations
University of Geneva

Shrinkage Methods

Shrinkage Methods

- Shrinkage methods shrink the coefficient estimates of a regression model towards 0.
- This leads to a decrease in variance at the cost of an increase in bias.
- If the decrease in variance dominates the increase in bias, this leads to a decrease in the test error.
- The two best-known methods for shrinking regression coefficients are ridge regression and the lasso.

Shrinkage Methods

Ridge Regression

Ridge Regression

- When we fit a model by least squares, the coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 . \qquad (1.4.1)$$

- In ridge regression, the coefficient estimates are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{shrinkage penalty}}, \quad (1.4.2)$$

where $\lambda \geq 0$ is a tuning parameter.

Ridge Regression

- Tuning parameter λ controls the relative impact of the two terms on the coefficient estimates:
  - If $\lambda = 0$, then the ridge regression estimates are identical to the least squares estimates.
  - As $\lambda \to \infty$, the ridge regression estimates will approach 0.
- Note that the shrinkage penalty is applied to $\beta_1, \ldots, \beta_p$, but not to the intercept β_0, which is a measure of the mean value of the response variable when $x_{i1} = x_{i2} = \ldots = x_{ip} = 0$.

Ridge Regression

- Least squares estimates are scale equivariant: multiplying predictor X_j by constant c leads to a scaling of the least squares estimate by factor $1/c$ (i.e., $\hat{\beta}_j X_j$ remains the same).
- Ridge regression estimates can change substantially when multiplying a predictor by a constant, due to the sum of squared coefficients term in the objective function.
- Therefore, the predictors should be standardized as follows before applying ridge regression

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}, \qquad (1.4.3)$$

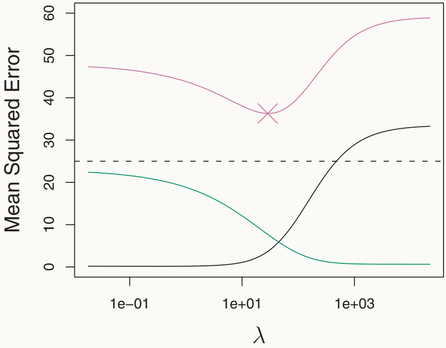
so that they are all on the same scale.

Shrinkage Methods

Why Does Ridge Regression Improve Over Least Squares?

Why Does Ridge Regression Improve Over Least Squares?

- As λ increases, the flexibility of ridge regression decreases, leading to increased bias but decreased variance.
- Simulated data containing $n = 50$ observations and $p = 45$ predictors (test MSE is a function of variance and squared bias):



(Squared bias (black), variance (green), and test MSE (purple) for the ridge regression predictions on a simulated data set. Source: James et al. 2013, 218.)

Why Does Ridge Regression Improve Over Least Squares?

- When the relationship between the response and the predictors is close to linear, the least squares estimates have low bias but may have high variance.
- In particular, when the number of predictors p is almost as large as the number of observations n (as in the above simulated data), the least squares estimates are extremely variable.
- Hence, ridge regression works best in situations where the least squares estimates have high variance.