Lecture 4: Regression Contd.

Rajdeep Banerjee

Some common metrics for regression

- Residual standard error (RSE):

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Estimates the standard deviation of the error. Has same unit as y and therefore hard to compare.

- R^2:

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

Proportion of the variance explained.

- Other metrics: Cp, AIC, BIC, adjusted R^2, refer to Hastie-Tibshirani ISL pp: 232-235.

Obtaining best model: Subset selection

- Best feature selection:
  - Take all possible combinations to select the target number of features and build model.
  - Select the best
- Forward feature selection:
  - Start with lowest number of features and iteratively increase number of features, each time selecting the best.
  - Select the one with best metric.
- Backward feature selection:
  - Start with all features and iteratively reduce the number of features, each time selecting the best.
  - Select the one with best metric.

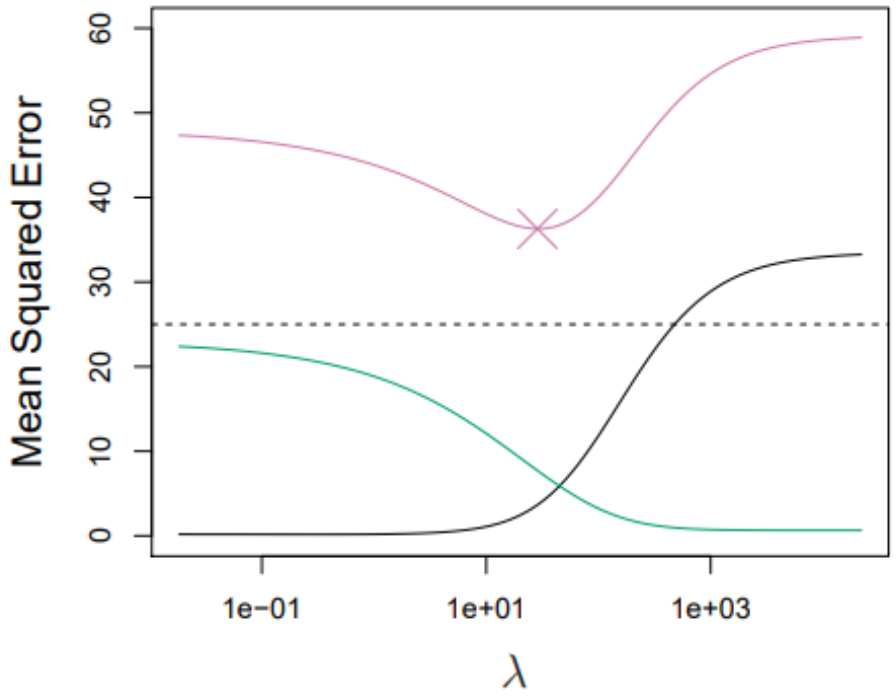
Shrinkage methods: Ridge

- The goal is to reduce variance → reduce overfit.
- A method that constrains or regularizes the coefficients or shrinks them toward zero.
- **Ridge regression**:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

- $\lambda \geq 0$, is a tuning parameter, determines the penalty.

  $\lambda \uparrow \Rightarrow \beta \rightarrow 0$.
- Increasing λ reduces flexibility of the coefficients to t to any data, reducing variance.



Shrinkage methods: Lasso

- Disadvantage of ridge: Does not reduce number of features.
- Lasso: Changes the penalty term $l2 \rightarrow l1$ norm.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

- Advantage: Coefficients can be made exactly 0
- Enables variable selection.

Ridge vs. Lasso

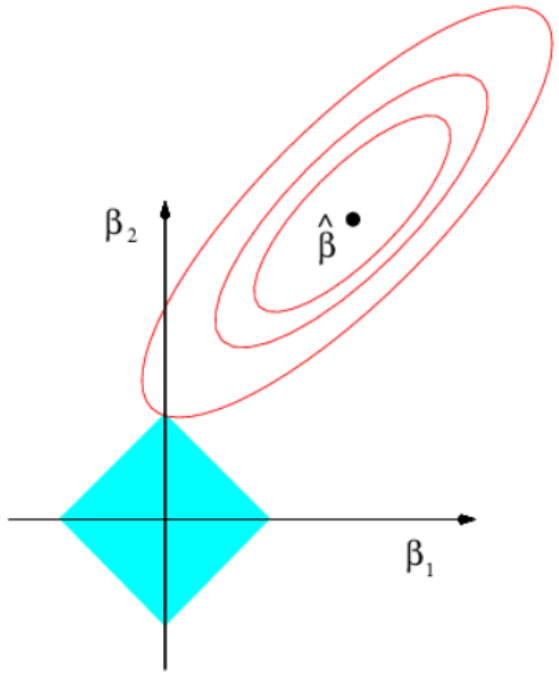
Another way to write the optimization problem:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

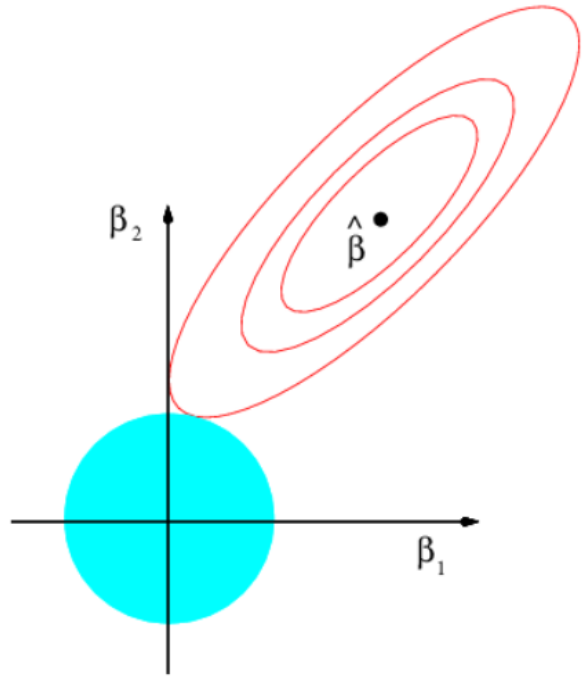
$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \qquad \textbf{Lasso}$$

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s, \qquad \textbf{Ridge}$$



Lasso



Ridge

There is a one to one relation between λ and s.