VARIABLE SELECTION

Model selection in linear regression

How to choose between competing linear regression models

- Model too small: “underfit” the data; poor predictions
- Model too big: “overfit” the data; poor predictions
- Model just right: good balance..

Model selection

How to extract the best set of features/model for my problem

- Subset selection (variable/feature selection)
- Shrinkage method (regularized model)
- Feature extraction (discussed later)

Variable selection

- What if there are p=1000 variables while we have only 50 samples?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{1000} x_{1000}?$$

- Find parsimonious model
  - More robust
  - Higher predictive accuracy
  - Parsimony (a.k.a. Occam's razor): the simpler, the better

- Methods
  - Exhaustive search
    - For p predictors, explore 2^p possibilities and choose the best
  - Partial search
    - Forward
    - Backward
    - Stepwise

Exhaustive Search (Best Subset Selection)

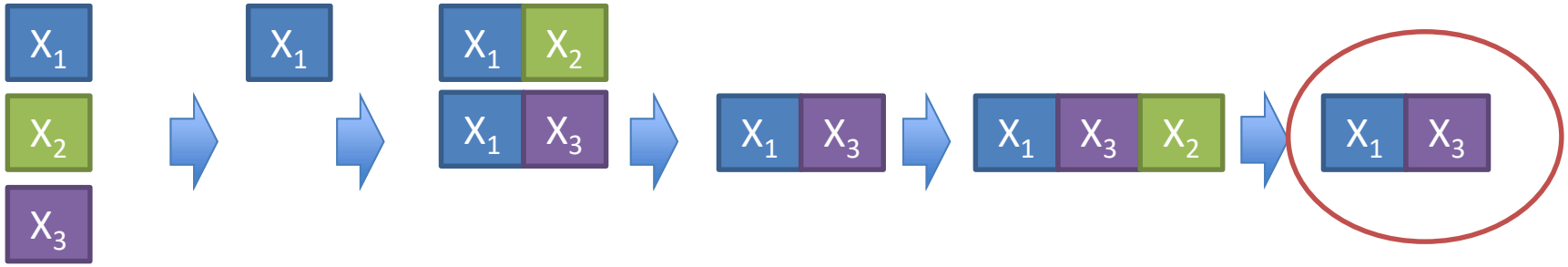
- All possible subsets of predictors assessed
  - For example, with three variables x_1, x_2, x_3
    - A total of seven (or eight) combinations are evaluated
    - Adjusted R^2 is used for performance criteria

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

  - For p predictors, explore 2^p possibilities and select the best combination of variables …

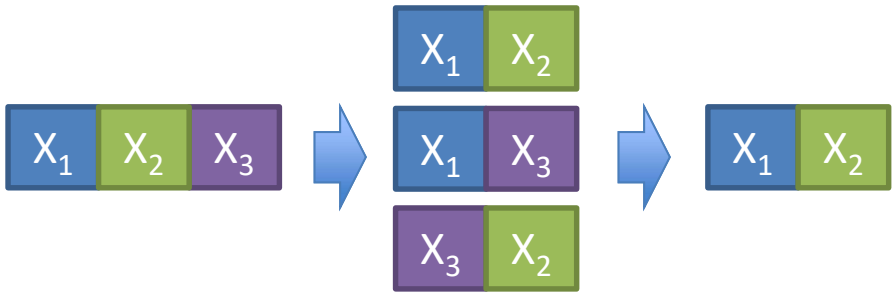
Forward selection

- Start with no predictors
- Add them one by one (add the one with the largest contribution)
- Stop when the addition is not statistically significant



Backward elimination

- Start with all predictors
- Successively eliminate least useful predictors one by one
- Stop when all remaining predictors have statistically significant contribution

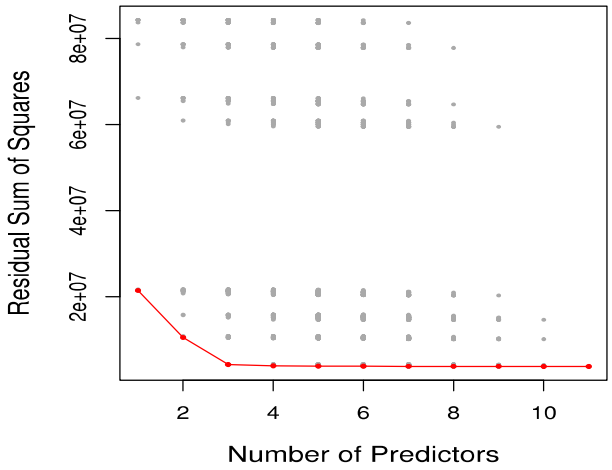


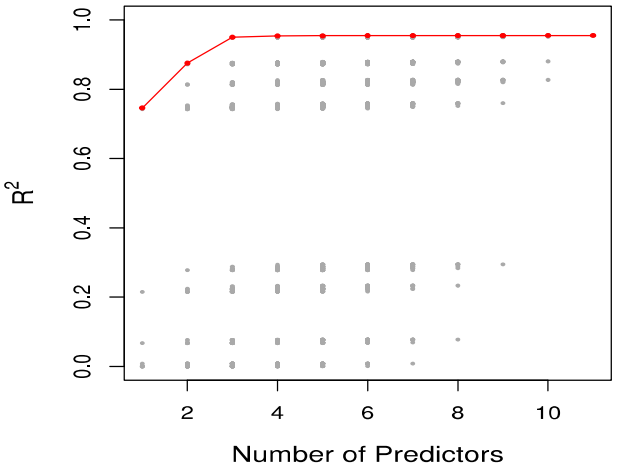
Stepwise selection

- Like backward elimination (or forward selection)
- except at each step, also consider adding back significant predictors (or dropping non-significant predictors)

R^2 vs. Subset Size

- The RSS/R^2 will always decline/increase as the number of variables increase so they are not very useful
- The red line tracks the best model for a given number of predictors, according to RSS and R^2





Other Measures of Comparison

- To compare different models, we can use other approaches:
  - Adjusted R^2
  - AIC (Akaike information criterion)
  - BIC (Bayesian information criterion)
  - C_p (equivalent to AIC for linear regression)
- These methods add penalty to RSS for the number of variables (i.e. complexity) in the model
- None are perfect

Choosing the optimal model

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

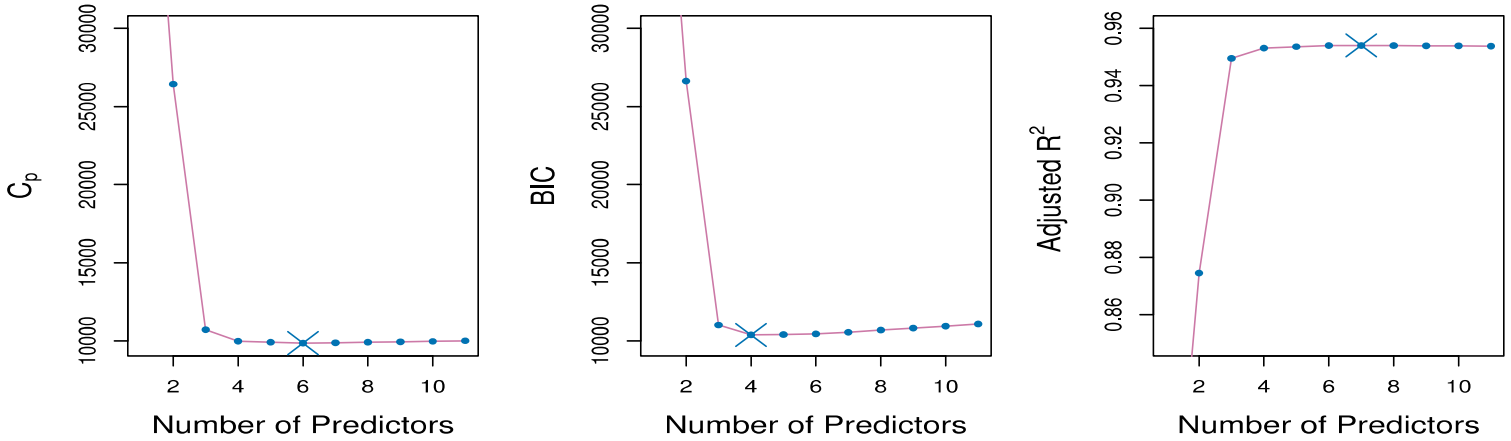
$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right)$$

- *d*: the number of predictors in the model
- σ: the variance of the error ε associated with each sample
- TSS: total sum of squares $\sum(y_i - \bar{y})^2$

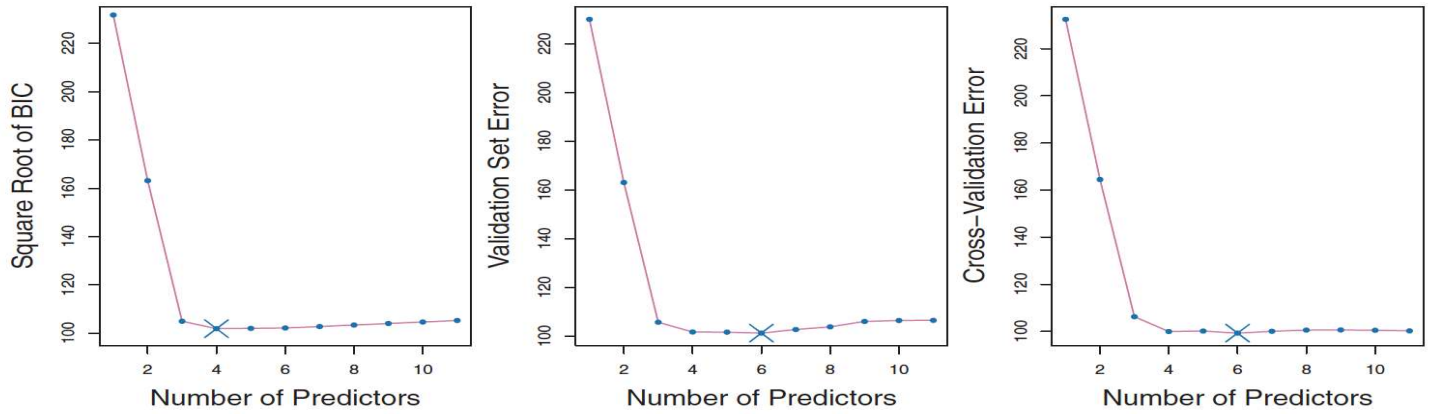
C_p, BIC, and Adjusted R^2

- A small value of C_p and BIC indicates a low error, and thus a better model
- A large value for the Adjusted R^2 indicates a better model



Cross-validation for model selection

- Nowadays with fast computers, the computations for cross-validations are hardly an issue



- One-standard-error rule: select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve

REGULARIZED LINEAR MODEL

Regularization

- Ordinary least square fit

$$\beta^* = \text{argmin}\, L(X, Y; \beta) = \text{argmin} \sum (y(x_n, \beta) - y_n)^2$$

- Shrinkage

$$\beta^* = \text{argmin}\, L(X, Y; \beta) + \Omega(\beta)$$

  - Ridge $\Omega(\beta) = \lambda \sum_i |\beta_i|^2$
  - Lasso $\Omega(\beta) = \lambda \sum_i |\beta_i|$
  - Elastic net $\Omega(\beta) = \lambda_2 \sum_i |\beta_i|^2 + \lambda_1 \sum_i |\beta_i|$

Ridge Regression

Ordinary Least Squares (OLS) estimates $\beta's$ by minimizing

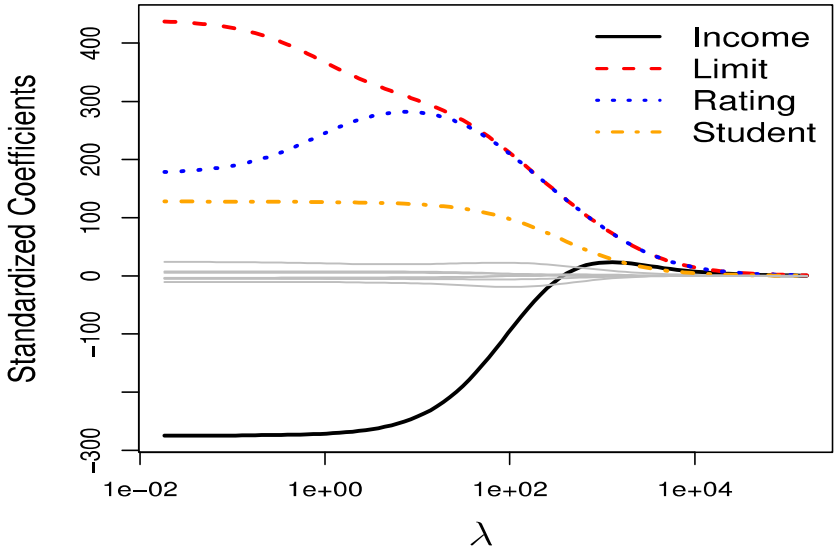
$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Ridge Regression uses a slightly different equation

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

Credit Data: Ridge Regression

- As λ increases, the standardized coefficients shrinks towards zero.



Why can shrinking towards zero be a good thing to do?

- The L2 penalty has the effect of "shrinking" large values of towards zero.
- In particular when n and p are of similar size or when n < p, then the OLS estimates will be extremely variable.
- The penalty term makes the ridge regression estimates biased but can also substantially reduce variance
- Notice that when $\lambda = 0$, we get the OLS!

The LASSO

- One significant problem of Ridge regression is that the penalty term will never force any of the coefficients to be exactly zero. Thus, the final model will include all variables, which makes it harder to interpret
- A more modern alternative is the LASSO
- The LASSO works in a similar way to Ridge Regression, except it uses a different penalty term

LASSO's Penalty Term

- Ridge Regression minimizes

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

- The LASSO estimates the $\beta's$ by minimizing the

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

What's the Big Deal?

- This seems like a very similar idea but there is a big difference
- Using this penalty, it could be proven mathematically that some coefficients end up being set to exactly zero
- With LASSO, we can produce a model that has high predictive power and it is simple to interpret

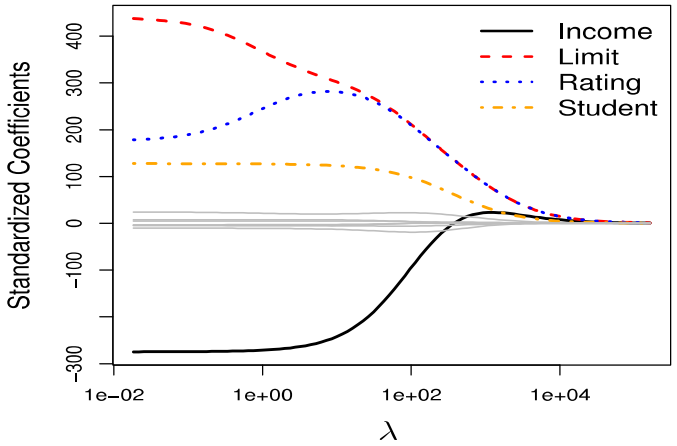
Ridge vs. LASSO

- In Ridge regression, the penalty term will never force any of the coefficients to be exactly zero. Thus, the final model will include all variables, which makes it harder to interpret

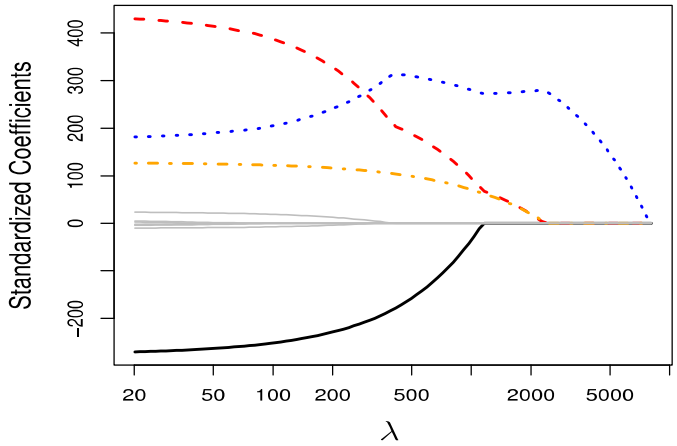
- The LASSO works in a similar way, except it uses a different penalty term. It could be proven proven mathematically that some coefficients end up being set to exactly zero.

Hyperparameter λ

- As λ increases, the standardized coefficients shrinks towards zero.



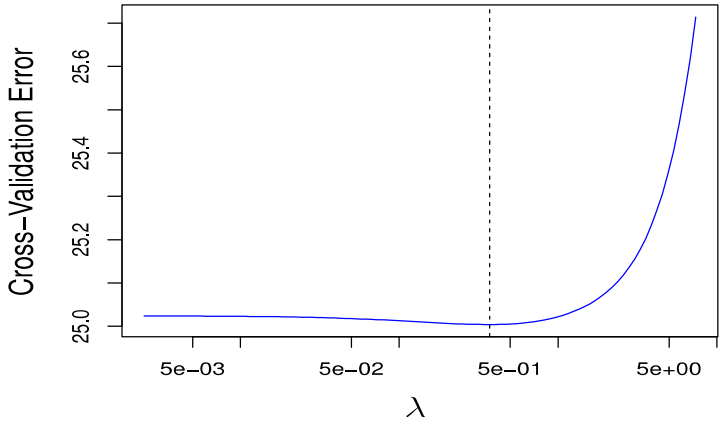
Ridge Regression

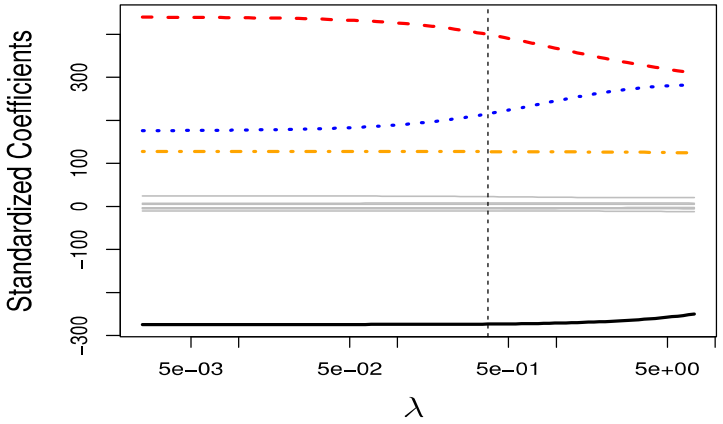


Lasso

Selecting the Tuning Parameter λ

- We need to decide on a value for λ
- Select a grid of potential values, use cross validation to estimate the error rate on test data (for each value of λ) and select the value that gives the least error rate





Summary

- Linear regression models are very popular, not only for explanatory modeling, but also for prediction
- A good predictive model has high prediction accuracy
- Predictive models are built using a training set and evaluated on a separate validation set
- Removing redundant predictors is key to achieving predictive accuracy and robustness