ECON3389 Machine Learning in Economics

Module 4
Feature Selection in Linear Models

Alberto Cappello

Department of Economics, Boston College

Fall 2024

Overview

Agenda:

- Subset selection methods.
- Ridge regression.
- Lasso regression.

Readings:

- ISLR Chapter 6, sections 6.1 and 6.2

Linear Model: Pros and Cons

- In this chapter, we are going to extend our understanding of the Linear Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

- Despite its simplicity, linear regression model estimated by OLS has two key advantages: interpretability and good predictive performance.
- However, the linear nature of regression function means that complex non-linear relationships cannot be easily modeled.
- One solution is to add more features to the model — interactions, powers, log-transformations and so on. This allows the model to retain its linear nature, yet approach non-linear models in terms of flexibility and predictive performance.
- The question then becomes — how does one select which features (regressors) to include? And why do we want to select a subset of the features?

Why Consider Alternatives to OLS?

- Model Interpretability
  - Including irrelevant variables in our model leads to unnecessary complexity in the resulting model. By removing these variables we can obtain a model that is more easily interpreted.
- Prediction Accuracy
  - Suppose n is the number of observations and p is the number of regressors
  - OLS estimates generally have low bias
  - When $n \gg p$, OLS estimates tend to also have low variance, and hence will perform well on test observations
  - When n is not much greater than p then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations
  - Finally, OLS is generally infeasible when $p > n$.

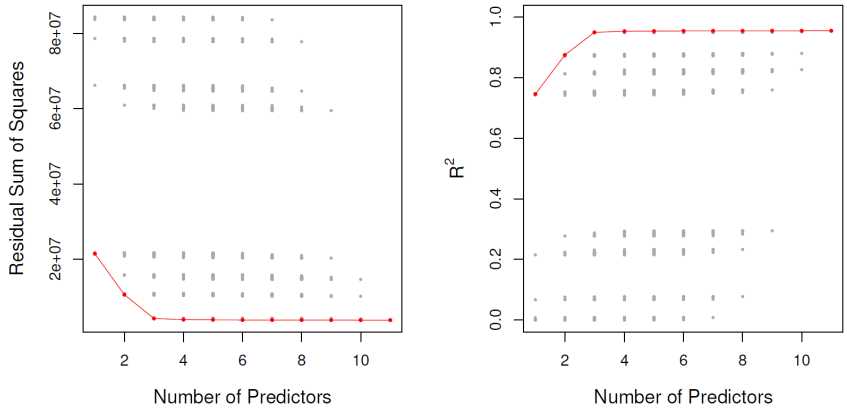
Feature Selection Methods

- *Subset selection*. Identify a subset of the p predictors that are believed to be related to the response Y. Then fit a model using least squares on the reduced set of variables.
  - Examples: best subset selection, forward stepwise selection, backward stepwise selection.
- *Shrinkage/regularization*. Fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage/regularization has the effect of reducing variance and can also perform variable selection.
  - Examples: ridge regression, lasso regression, elastic net regression.
- *Dimension reduction*. Project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.
  - Examples: principal component regression, partial least squares.

Best Subset Selection

- Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
- For $k = 1, 2, \ldots, p$:
  - Fit all $\binom{p}{k} = \frac{p!}{k!(p-k)!}$ models that contain all possible combinations of k predictors out of p.
  - For each value of k, pick the best out of these models as having the smallest value of the loss function (lowest RSS or highest R^2) on the training dataset and call that model $\mathcal{M}_k$.
- Select a single best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using either MSE from cross-validation, C_p (AIC), BIC or adjusted R^2 (more on these later).

Best Subset Selection



Stepwise Selection

- The total number of models to estimate in best subset selection algorithm is equal to 2^p and that number grows very quickly with p. There are 1024 models for $p = 10$, over a million for $p = 20$ and with $p = 40$ it becomes computationally infeasible even on fastest modern hardware.
- Because of mainly this reason, *stepwise* methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Forward Stepwise Selection

- *Forward stepwise selection* begins with a model containing no predictors, and then adds predictors to the model one-at-a-time, until all p predictors are in the model.
- At each step the variable that gives the greatest additional improvement to the fit is added to the model.
- Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
- For $k = 1, 2, \ldots, p-1$:
  - Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
  - Pick the best out of these models as having the smallest value of the loss function (lowest RSS or highest R^2) on the training dataset and call that model $\mathcal{M}_{k+1}$.
- Select a single best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using either MSE from cross-validation, C_p (AIC), BIC or adjusted R^2.

Backward Stepwise Selection

- *Backward stepwise selection* begins with a full model containing all p predictors, and then iteratively removes the least useful predictor one-at-a-time.
- Let $\mathcal{M}_p$ denote the full model, which contains all p predictors.
- For $k = p, p-1, \ldots, 1$:
  - Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$ for a total of $k-1$ predictors.
  - Pick the best out of these k models (lowest RSS or highest R^2) on the training dataset and call that model $\mathcal{M}_{k-1}$.
- Select a single best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using either MSE from cross-validation, C_p (AIC), BIC or adjusted R^2.

Backward vs Forward Selection

- Both backward and forward stepwise selection search only through a small subset of 2^p and thus can be applied in settings where p is too large for best subset selection.
- However, neither of them is guaranteed to yield the best model containing a subset of p predictors.

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income`	`rating`, `income`,
	`student`, `limit`	`student`, `limit`

- Backward selection requires that the sample size n is larger than the number of variables p (so that the full model with p predictors can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Choosing the Optimal Model

- The model containing all p predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error, which is typically negatively related to number of features used.
- We wish to choose a model with low test error, not a model with low training error. Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.
- We can directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in previous lectures.
- Alternatively, we can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.

C_p, AIC, BIC, and Adjusted R^2

- C_p, AIC, BIC, and Adjusted R^2 are different measures designed to introduce a correction to training error to help avoid overfitting issues.
- *Mallow's* C_p:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

where d is the total number of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error term ϵ.

- The *AIC* criterion is defined for a large class of models fit by maximum likelihood:

$$\text{AIC} = -2\log L + 2d$$

where L is the maximized value of the likelihood function for the estimated model.

- In case of linear model with normal errors C_p and AIC are equivalent.

C_p, AIC, BIC, and Adjusted R^2

- *BIC*

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

- Like with C_p and AIC, the lower value of BIC, the better.
- BIC replaces the term $2d\widehat{\sigma}^2$ used by C_p with a term $\log(n)d\widehat{\sigma}^2$. Since $\log(n) > 2$ for any $n > 7$, BIC places heavier penalty on models with many variables, and hence usually results in the selection of smaller models than C_p.

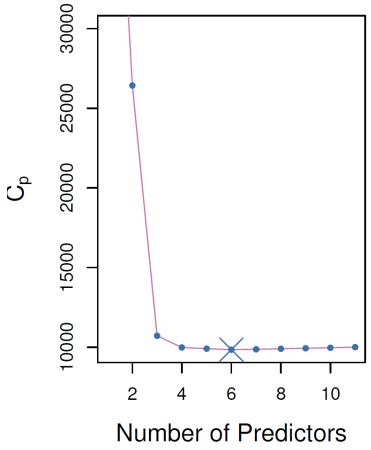
C_p, AIC, BIC, and Adjusted R^2

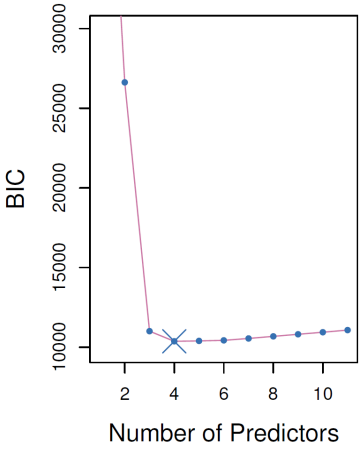
- For a least squares model with d variables the *adjusted* R^2 statistic is

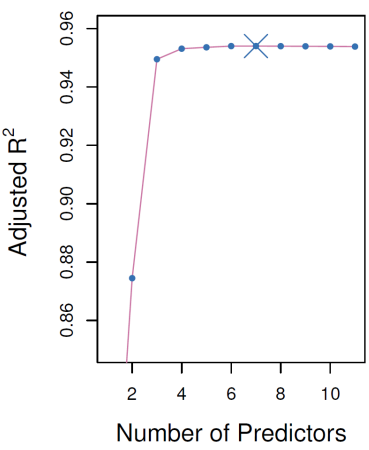
$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

- Unlike C_p, AIC and BIC, a larger value of R^2_{adj} indicates a model with a smaller test error.
- Maximizing R^2_{adj} is equivalent to minimizing $\frac{RSS}{(n-d-1)}$. While RSS always decreases as the number of variables in the model increases, R^2_{adj} may increase or decrease due to the presence of d in the denominator.
- In other words, unlike the standard R^2, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.

Credit Data Set Example







Shrinkage Methods

- The subset selection methods fit a linear model that contains only a subset of the predictors. This is equivalent to setting the coefficients on excluded predictors to zero prior to running the estimation algorithm.
- As an alternative, one can fit a model containing all p predictors using a technique that *regularizes* the coefficient estimates, or equivalently, that *shrinks* the coefficient estimates towards zero as part of the estimation algorithm.
- It may not be immediately obvious why such a constraint should improve the fit or if the algorithm will work in the first place, but it turns out that shrinking the coefficient estimates can significantly reduce their variance at a cost of a minor increase in bias.
- Two most common shrinkage/regularization methods are *ridge regression* and *lasso regression*.

Ridge Regression

- Standard least squares regression fits the model by picking values of $\beta_0, \beta_1, \ldots, \beta_p$ that minimize

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

- Ridge regression instead picks coefficient values $\widehat{\beta}^R$ that minimize

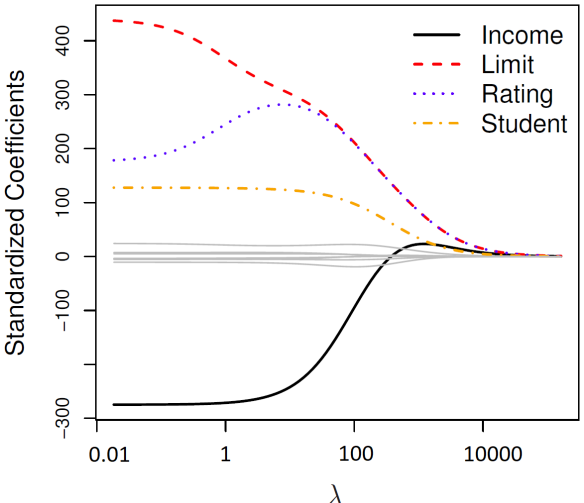
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Parameter λ is the *tuning parameter*, to be determined separately.

Ridge Regression

- The nature of ridge regression is similar to that of OLS: seek coefficient values that make the model fit the data well (by making RSS small).
- However, now we can no longer set values of coefficients to arbitrary values, even if that significantly decreases RSS. This is because the second term $\lambda \sum_{j=1}^{p} \beta_j^2$, called a *shrinkage penalty*, will increase our loss function if values of $\beta_0, \beta_1, \ldots, \beta_p$ are far away from zero.
- Because loss function now has two terms to balance out, the extra second term has the effect of *shrinking* the estimates of β_j towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates. Selecting a good value for λ is critical, and is done via cross-validation.

Credit Card Data Example



- As can be seen from the picture, it is always possible to set λ to a value that will shrink all coefficients arbitrary close to zero.
- As such, we need to perform cross-validation testing to see which value of λ achieves minimal total value of ridge loss function.
- The process is usually done via a grid search algorithm (more on that later)

Feature Scaling and Standardization

- Standard least squares coefficient estimates are *scale invariant*: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimate $\widehat{\beta}_j$ by a factor of $1/c$. In other words, regardless of how the j-th predictor is scaled, $\widehat{\beta}_j X_j$ will always remain the same.
- In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression loss function.
- Unlike the first term, which contains $\widehat{\beta}_j X_j$ parts, the shrinkage penalty contains values of only $\widehat{\beta}_j^2$, thus making is scale-dependent.
- Therefore, it is best to apply ridge regression after *standardizing* the predictors:

$$\widetilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2}}$$

Why Does Ridge Regression Improve Over LS: Bias-variance trade-off

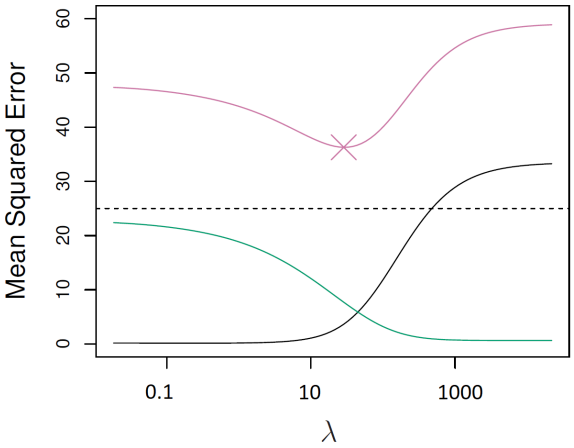
- Suppose our test data `Te` consists of a single data point (x_0, y_0). Then

$$MSE = \mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] + \text{Var}(\epsilon_0)$$

$$= \underbrace{\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}\left[\hat{f}(x_0)\right]\right)^2\right]}_{\text{Var}\left(\hat{f}(x_0)\right)} + \underbrace{\mathbb{E}\left[\left(f(x_0) - \mathbb{E}\left[\hat{f}(x_0)\right]\right)^2\right]}_{\mathbb{E}\left[\text{Bias}^2\left(\hat{f}(x_0)\right)\right]} + \text{Var}(\epsilon_0)$$

- Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set
- Bias refers to the error that is introduced by approximating a real-life problem by a much simpler model
- Typically as the flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on MSE amounts to a bias-variance trade-off.

Why Does Ridge Regression Improve Over LS?



- **Squared bias**, **variance** and **MSE**.
- Because OLS is free to choose any coefficient values, it tends to pick the ones that provide best fit, meaning less bias and more variance.
- Ridge regression, on the other hand, is penalized for choosing coefficients with high second moments, thus leading to less variance, slightly more bias, but lower MSE overall.

Lasso regression

- Unlike subset selection, which generally selects models that involve just a subset of all variables, ridge regression will include all p predictors in the final model. This makes ridge regression completely infeasible when $p > n$, as if often the case, for example, with Internet-related data.
- The LASSO (Least Absolute Shrinkage and Selection Operator) is an alternative that overcomes this disadvantage. It achieves that by using a different type of shrinkage penalty:

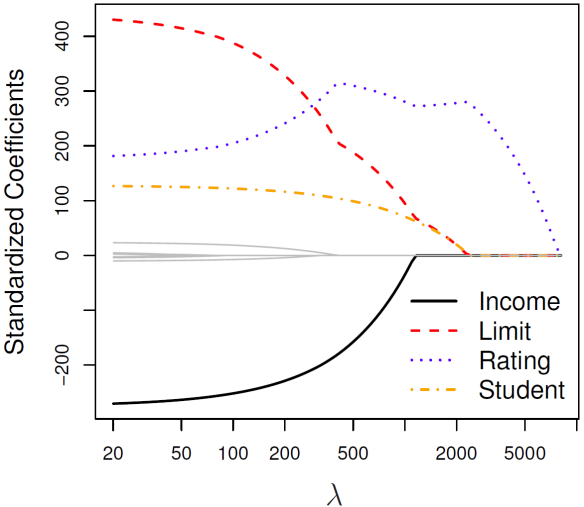
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

- In statistical lingo, this type of penalty is known as ℓ_1-penalty, because it uses ℓ_1 norm of coefficient vector β given by $||\beta||_1 = \sum|\beta_j|$. Ridge regression, on the other hand, uses ℓ_2 norm as a penalty, given by $||\beta||_2 = \sum\beta_j^2$

Lasso Variable Selection

- As with ridge regression, lasso shrinks all coefficient estimates towards zero.
- However, unlike ℓ_2 penalty in ridge regression, lasso penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, lasso performs variable selection, starting with the full set of p variables. We say that lasso yields *sparse* models — that is, models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of λ for lasso is critical; cross-validation is again the method of choice.

Credit Card Data Example



- Similar to ridge regression, setting λ to sufficiently high value will shrink all coefficients to zero.
- Unlike ridge regression, lasso coefficients will get shrunk exactly to zero in a single jump, without smooth continuous decline.
- Additionally, while ridge regression shrinks all coefficients close to zero around the same values of λ, lasso sets some coefficients to zero much earlier than others.

Lasso vs Ridge

- Why is it that in lasso regression we get some of the coefficients shrunk exactly to zero, but not in ridge regression?
- One can show that lasso and ridge regression coefficient estimates solve the following problems:

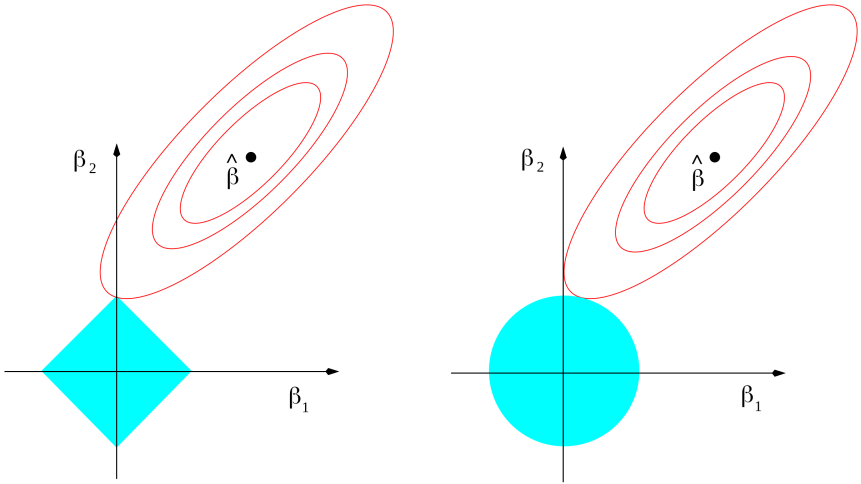
$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

- These two problems have a useful geometric representation that shows exactly why lasso induces sparsity among coefficients.

Lasso vs Ridge



Left: lasso Right: Ridge

Selecting the Tuning Parameter λ

- Both with ridge and with lasso we need to select the value for the tuning parameter λ or equivalently, the value of the constraint s in a way that will not lead to overfitting or other mistakes. Cross-validation provides a simple way to tackle this problem.
- We choose a grid of λ values and fit a separate model for every value from that grid using K-fold cross-validation.
- We then compute the cross-validation error for each value of λ and select the one for which that error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Pros and Cons of Lasso

- In terms of overall fit neither ridge regression nor lasso will universally dominate the other.
- In general, one might expect lasso to perform better when the response is a function of only a relatively small number of predictors. However, that is never known *a priori* with real-life data.
- Ridge regression can perform better if $p < n$ and there is no a priori reason for some of the variables to not be included in the model.
- Lasso can perform variable selection and model estimation with $p > n$, but has a known issue of ignoring groups of correlated variables (e.g. performance metrics of NBA players) and almost randomly selecting only one variable out of the group.

Lasso and Economics

- Despite its know flaws, over the past decade lasso has become very popular with both academic researchers and applied economists.
- From the theoretical perspective, multiple extensions and variations of lasso has been suggested, and today advanced versions of it can deal both with correlated regressors (group lasso, elastic net) and biased estimates (adaptive lasso, post-lasso).
- The main driving force behind is the ability to tackle datasets that previously were completely unusable due to number of variables p being close or even larger than sample size n.
- Additionally, lasso allows economists to utilize *sparse* structural models, e.g. consumer preferences across hundreds of product attributes with most of them having zero importance.