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.

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

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

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.

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.

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.

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.

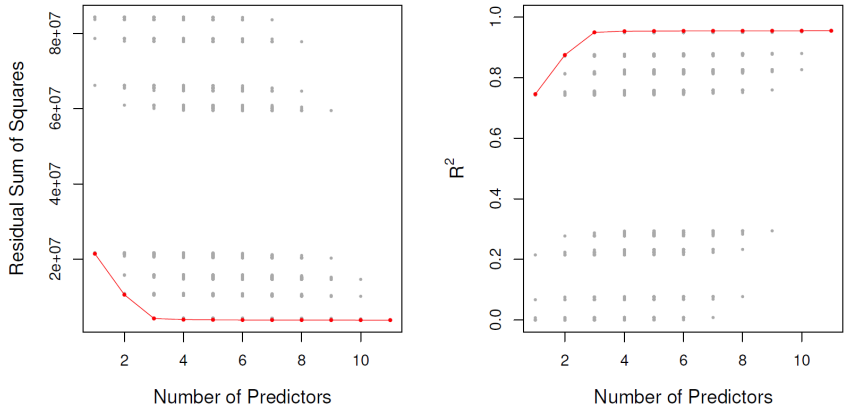
Best Subset Selection

- 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 $\mathcal{M}_k$.

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.

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.

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 $\mathcal{M}_{k+1}$.

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.

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 $\mathcal{M}_{k-1}$.

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.

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 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.

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.

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.

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.

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.

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.

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