New York University: Machine Learning in Economics

Lecture 2: Linear Methods for Regression

Dr. George Lentzas

Linear Model Selection

Motivation

- Is there any way to improve the simple linear regression model, perhaps by replacing the least squares fitting methodology by another fitting procedure? There are two ways we might want to "improve" the least squares model:
- Prediction Accuracy: Least squares will perform well when the true relationship between y and x is approximately linear (low bias) and $n >> p$, the number of observations is much larger than the number of variables (low variance). If the latter is not the case we can use *shrinkage methods* to reduce the variance with a small increase in bias.
- Feature Selection: Occasionally some of the variables in the multiple linear regression model will not be associated with the response variable and including such irrelevant variables would lead to unnecessary complexity. We can use *feature selection* or *variable selection* methods that will help us exclude such irrelevant variables.

Linear Model Selection

Subset Selection: Best Subset

- Subset selection refers to methods for choosing a particular subset of predictors among all the x variables.
- A naive approach for subset selection would be to fit a separate least squares (LS) regression for each of the 2^p possibilities.
- Then we use a criterion based on the estimated test MSE (cross-validated test error, for example) to choose the best model.
- Although this sounds appealing, it suffers from serious computational limitations and becomes infeasible for $p > 40$ or so.

Linar Model Selection

Subset Selection: Forward Stepwise

- Forward Stepwise Selection (FSS) starts with a trivial model which contains only an intercept and then adds predictors.
- At each step the variable that results in the greatest additional improvement is added (so it is a "greedy search" method so to speak). The criterion is usually cross validation or some other test MSE related metric such as AIC or BIC.
- This is computationally much more feasible than looking for the best model (we need to fit $1 + p\left(p + 1\right)/2$ models but is not guaranteed to find the best model out of all the 2^p models.

Linar Model Selection

Subset Selection: Backward Stepwise

- Backward Stepwise Selection (BSS) begins with a model containing all p predictors and the remove at each step the least useful predictions.
- As usual this is done using a criterion based on the estimated test MSE (cross-validated test error, for example) to choose the best model.
- Similar to FSS we only need to fit $1 + p(p+1)/2$ models and is not guaranteed to find the best model either.

Linar Model Selection

Subset Selection: Optimal Model

- All of the above methods require a way to determine when a model is better than another. R^2 is not a good way to compare models, as R^2 will always increase with the number of predictors.
- Instead we would like to choose a model that has a low test error. There are two approaches:
  1. Use the training error and make a formulaic adjustment to estimate the test error.
  2. Directly estimate the test error using the (cross) validation approach.
- Let us start with considering how we can make an adjustment to the training error to become a better estimate of the test error.

Linar Model Selection

Subset Selection: Optimal Model

- Recall that when we fit a linear regression model with d predictors using n observations using least squares we explicitly pick betas that minimize the training squared error (RSS = MSE / n).
- A first attempt to adjust RSS to estimate test MSE is the C_p measure:

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

  where $\hat{\sigma}^2$ is an estimate of the variance of the regression error.
- If $\hat{\sigma}^2$ is an unbiased predictor of the error variance then C_p will be an unbiased estimate of the test MSE.

Linar Model Selection

Subset Selection: Optimal Model

- Two alternatives are the AIC and BIC, which are actually defined for the larger class of models fit by maximum likelihood (and not just OLS).
- The AIC is given by

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

and the BIC by

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

- All of the above have theoretical underpinnings and are asymptotically equivalent so in practice what is important is consistent use of any of these.

Regularization

Shrinkage: Ridge Regression

- An alternative is to fit a model containing all p predictor variables using a technique that constrains (shrinks) some of the coefficient estimates to zero.
- A first such method is Ridge Regression where we pick $\hat{\beta}_R$ that minimize

$$RSS + \lambda \sum_{i=1}^{p} \beta_i^2.$$

- The parameter λ is called a tuning parameter and needs to be estimated too, usually done by cross-validation. It is very important to note that it does not apply to the intercept!
- Lets looks at an example using the "Credit" data from ISLR.

Regularization

Shrinkage: Ridge Regression

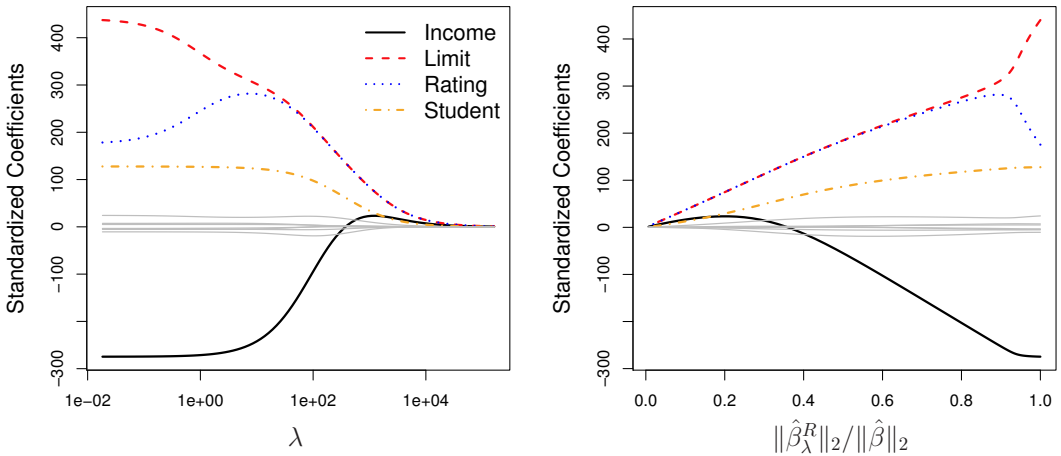

Figure: The standardized ridge regression coefficients as functions of λ and ridge vs ols coefficient ratios.

Regularization

Shrinkage: Ridge Regression

- Ridge Regression betas are unit dependent! That is they will change (not just scale as in with OLS) when multiplying a given predictor by a constant. Therefore, it is advisable to first standardize the predictors to make sure they have a variance of one:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{ij}\right)}}$$

- So why and when is this better than OLS? As the ridge coefficient increases, the flexibility of the model decreases, as a result increasing the bias but decreasing the variance of the estimated model.
- Therefore, in situations when the relationship between the predictors and the response variables is approximately linear but the least squares estimate has high variance we would expect Ridge Regression to do well. A common such scenario is when the number of variables p is in the same order of magnitude as the number of observations n.

Regularization

Shrinkage: Ridge Regression

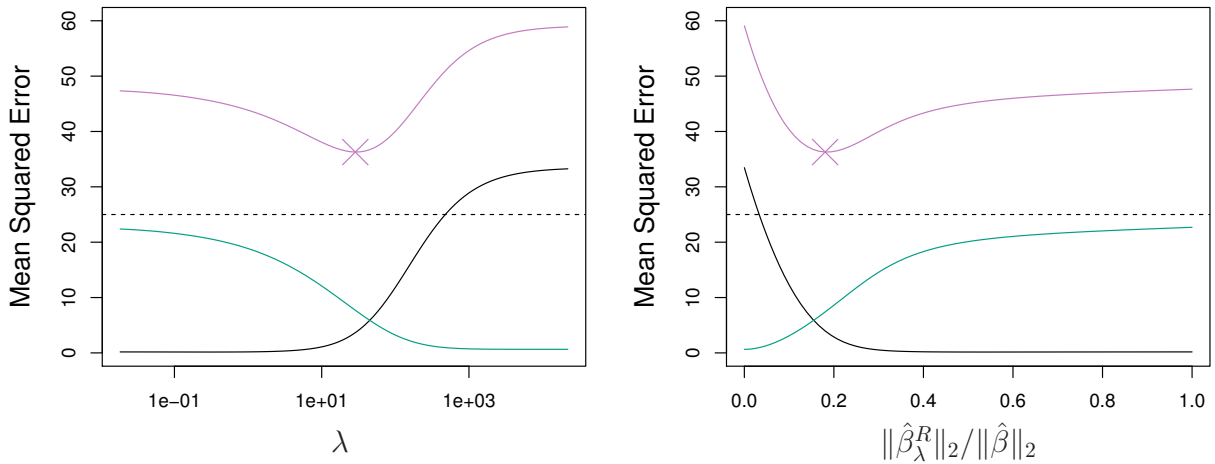

Figure: Squared bias(black), variance (green) and test MSE (purple) for Ridge Regression predictions on simulated data as a function of λ.

Regularization

Shrinkage: The Lasso

- Ridge Regression has the obvious drawback that it does not perform any kind of model simplification, i.e. it still includes all p predictors (no betas are set exactly equal to zero).
- The Lasso estimator is an alternative method that minimizes:

$$RSS + \lambda \sum_{i=1}^{p} |\beta_i| .$$

- This has the effect of forcing some of the parameters to be exactly zero and in this way allows the Lasso to perform an automatic variable selection as part of the estimation.

Regularization

Shrinkage: The Lasso

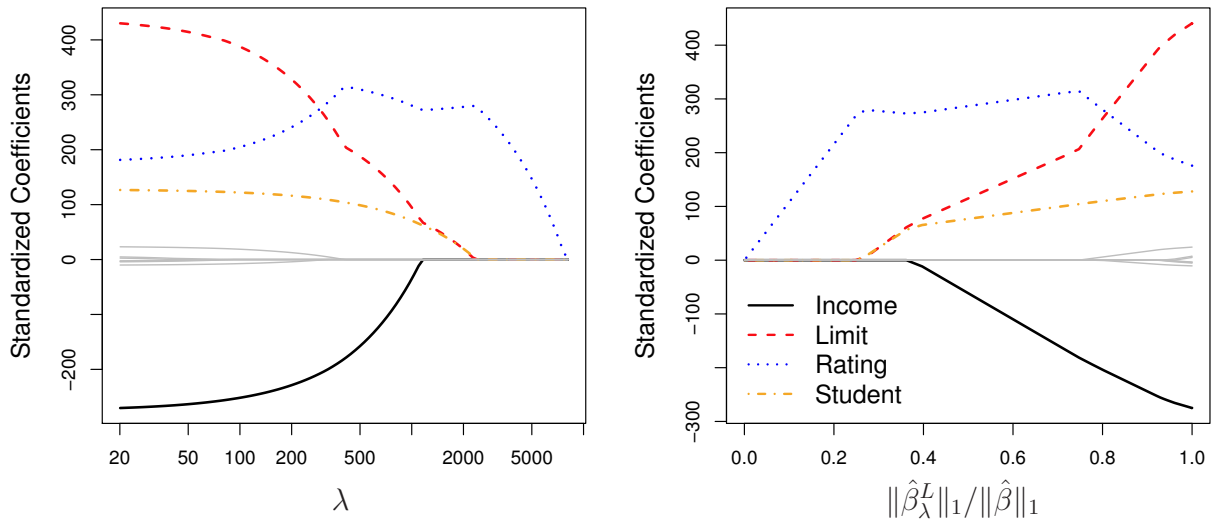

Figure: The standardized Lasso coefficients on the "Credit" data set from ISLR as functions of the tuning parameter and beta estimate ratios.

Regularization

Shrinkage

- When is the Lasso preferable (in terms of prediction accuracy) to Ridge Regression and vice versa?
- As you might expect it depends on the situation; we expect the Lasso to do better when a small subset of predictors have large coefficients while the remaining predictors have coefficients that are zero. On the contrary, Ridge Regression should do better when the response variable is a function of all the predictors, all with coefficients of comparable size.
- This can be seen in the next slide where the variance, squared bias and test MSE of the Lasso and Ridge Regression are plotted.

Regularization

Shrinkage: The Lasso

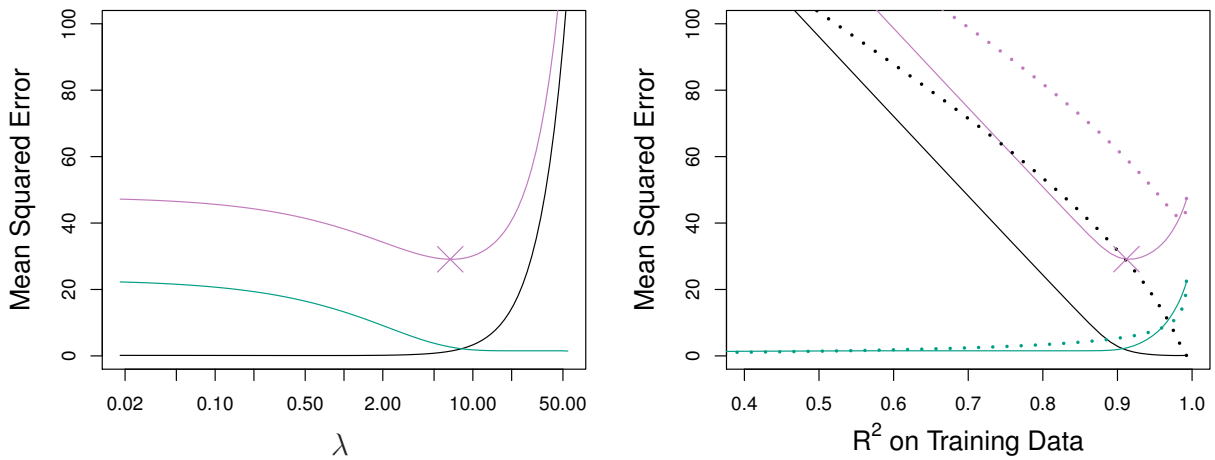

Figure: Squared bias (black), variance (green) and test MSE (purple) for the Lasso on the left. Comparison between Lasso (solid) and Ridge Regression (dotted) on the right.

Regularization

Shrinkage

- A key element in all the above is the choice of the optimal tuning parameter λ. This is done typically by cross validation over a grid of values. The tuned model is then re-fit with all observations.

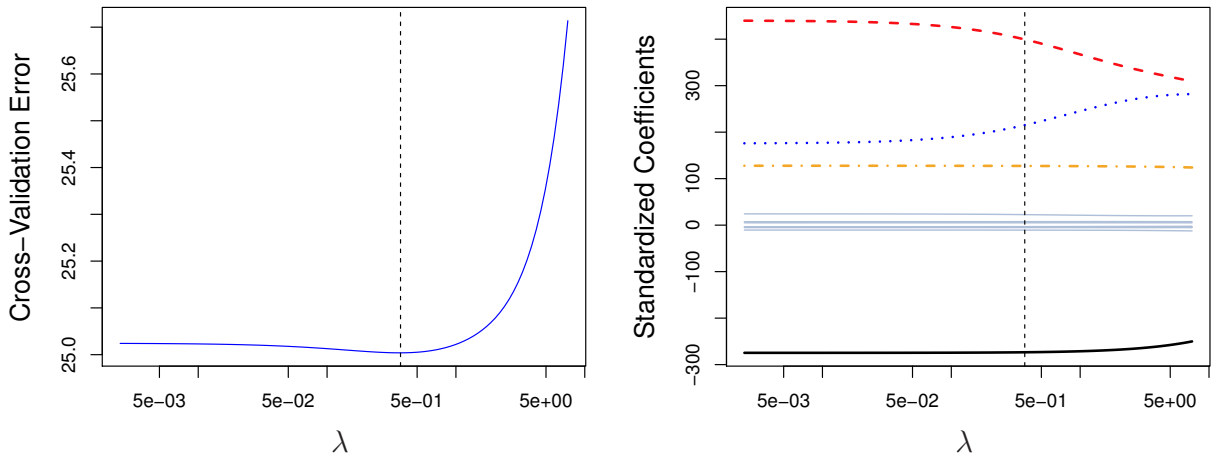

Figure: Cross validation estimation of the tuning parametr. On the left we have the CV Errors from Ridge Regression estimation on the "Credit" data set and on the right the coefficient estimates. The vertical line indicates the CV tuning parameter estimate.

Regularization

High Dimensions

- Classical econometric and statistical techniques were developed with datasets for which $n >> P$ in mind.
- Recently however, because of ever increasing data-collection, we are facing an increasing number of situations where the number of predictors is in the same order of magnitude as the number of observations. Data sets containing more predictors/features than observations are known as "high dimensional". For such data sets the classical least squares regression does no work.
- The crux of the problem is that in high dimensional data sets OLS will severely overfit the data. Unfortunately, none of the Subset Selection methods can address this either because estimating the error variance also become problematic.

Regularization

High Dimensions

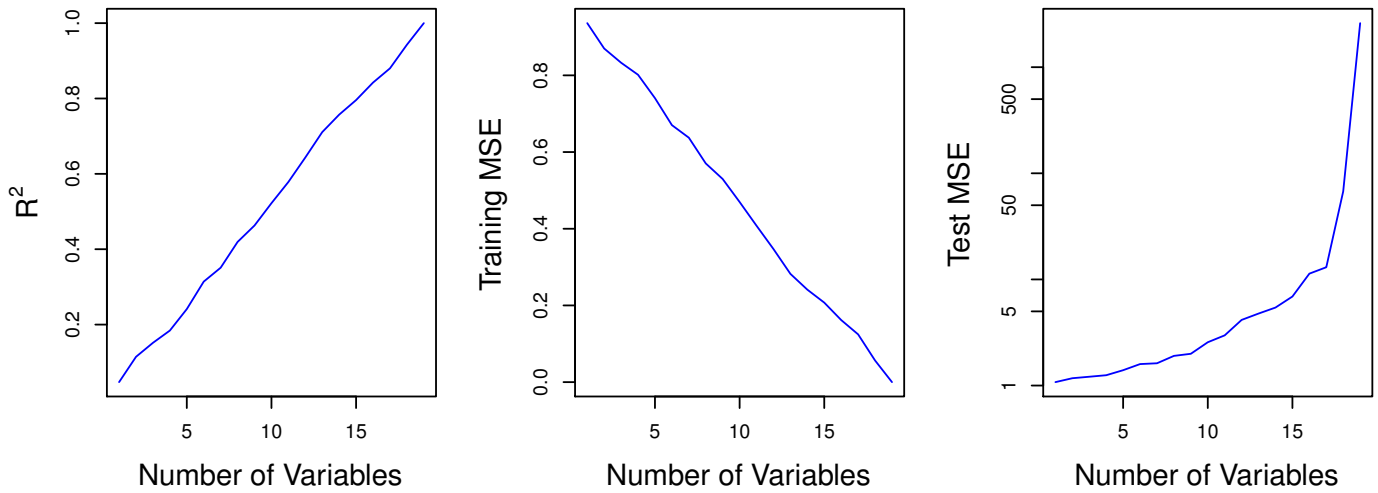

Figure: R^2, Training and Test MSE for a simulated data of $n = 20$ as a function of the number of regressors included in the model.

Regularization

High Dimensions

- Fortunately, Stepwise Selection, Ridge Regression and Lasso are all good ways to deal with regression in a high dimensional data set.
- However, it is key to understand that the test error rate tends to increase as the dimensionality increases, unless the additional predictors are truly associated with the response variable. This is known as the *curse of dimensionality*. We can see this clearly illustrated in the next graph where adding noise features leads to a deterioration in the fitted model, captured by and overall increase in the test error.

Regularization

High Dimensions

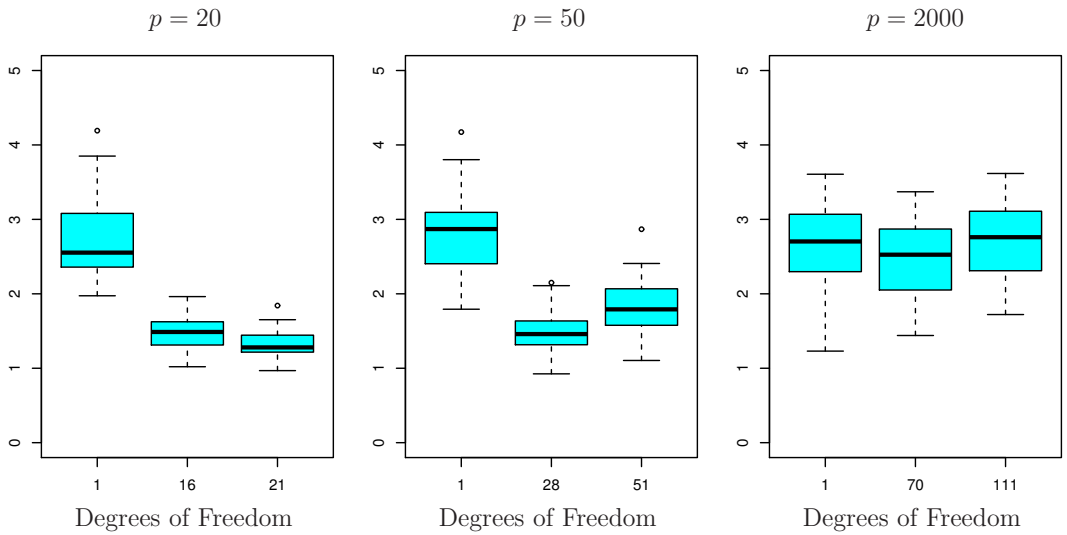

Figure: The Lasso for a sample of $n = 100$ and three different number of predictors in the model. Of these predictors 20 are associated with the true model. Degrees of freedom are the number of non-zero estimated coefficients.

Acknowledgements

- This presentation is based on and follow closely the excellent books (i) "Introduction to Statistical Learning with applications in R" by G. James, D. Witten, T. Hastie and R. Tibshirani and (ii) "The Elements of Statistical Learning" by T. Hastie, R. Tibshirani and J. Friedman.
- Also, some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani as well as from from "The Elements of Statistical Learning" (Springer, 2016) with permission from the authors: T. Hastie, R. Tibshirani and J. Friedman.