Introdução a Inteligência Computacional

Cap. 6: Linear Model Selection and Regularisation

Prof. Cristiano Leite de Castro
Departamento de Engenharia Elétrica - DEE

PPGEE - Programa de Pós-Graduação em Engenharia Elétrica

Belo Horizonte - Abril, 2025

Sumário

Improving the Linear Model

- Despite its simplicity, the linear model has distinct advantages in terms of its interpretability, and often shows good predictive performance.
- Hence, we discuss in this lecture some ways in which the simple linear model can be improved, by replacing ordinary least squares fitting with some alternative fitting procedures.

Why consider alternatives to Least Squares? (1)

- Provided that the true relationship between the response and the predictors is approximately linear, the least squares estimator will have low bias.
- Also, if $n >> p$, it tends to have low variance.
- However,
  1. if n is not much larger than p, then the model tends to have high variance.
  2. if $p > n$, then there is no longer a unique least squares coefficient estimate.
- Feature Selection and Regularization: alternatives to least-squares.
  - often results in better model interpretability.
  - may result in a better prediction accuracy.

Why consider alternatives to Least Squares? (2)

- It is often the case that some of the variables used in a multiple regression model are not associated with the response.
- Including such irrelevant variables leads to unnecessary complexity in the resulting model.
- Feature Selection and Regularization: alternatives to least-squares.
  - often results in better model interpretability.
  - may result in a better prediction accuracy.

Two Classes of Methods

- Subset Selection:
  - identifies a subset of the p predictors that is truly related to the response.
  - Then, it fits a model using least squares on the reduced set of variables.
- Regularisation:
  - fits a model involving all p predictors, but the estimated coefficients are shrunken towards zero.
  - it can significantly reduce variance at the cost of a negligible increase in bias.
  - it can also perform variable selection.

Subset Selection Approaches

- Best Subset Selection.
- Forward Stepwise Selection.
- Backward Stepwise Selection.

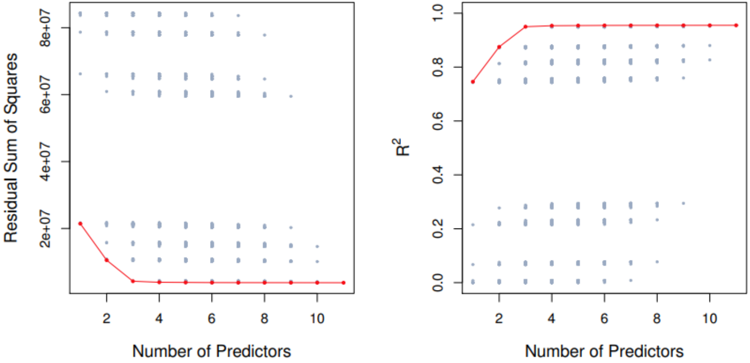
Subset Selection as a Search Problem

- The Subset Selection Problem can be modelled as a Search Problem, more specifically as a Constrained Satisfaction Problem (CSP):
  - Variables
  - Domains
  - Constraints
- One total assignment of the variables with their respective values is an candidate solution for the problem.
- Given a problem with p predictors, how many possible candidate solutions exist? (search space size)
- How to obtain the best solution among them?

Best Subset Selection (Brute-Force Search)

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Example: Best Subset Selection for Credit Data



Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p.
- Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- stepwise methods: explore a far more restricted set of models.

Forward Stepwise Selection (Best Improvement Heuristic)

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.

Forward Stepwise Selection (Best Improvement Heuristic)

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   2.1 Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   2.2 Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Best Subset x Forward Stepwise for the Credit Data

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

The first four selected models for best subset selection and forward stepwise selection on the `Credit` *data set. The first three models are identical but the fourth models differ.*

Backward Stepwise Selection

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.
- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   2.2 Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

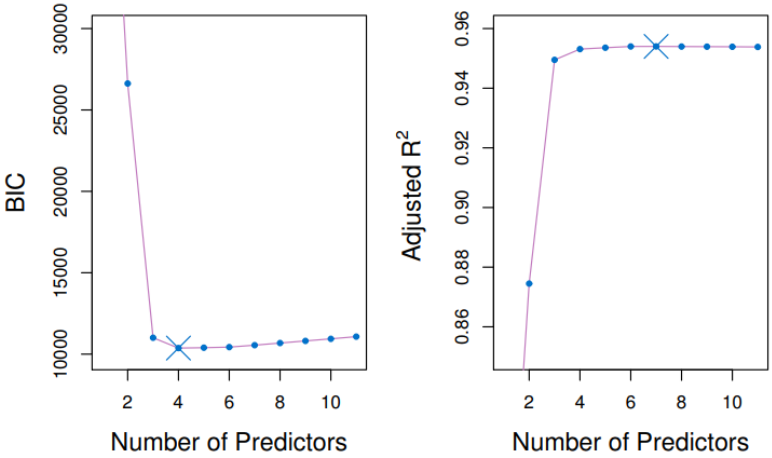
Choosing the Optimal Model

- The model containing all of the predictors will always have the smallest RSS and the largest R2, since these quantities are related to the training error.
- We wish to choose a model with low test error, not a model with low training error.
- Therefore, RSS and R2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

Choosing the Optimal Model

- We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.
- Model Selection Measures: BIC (Bayesian Information Criterion), Adjusted R^2.
- We can directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in previous lectures.

Example: BIC x Adjusted R^2



Regularization Methods for Linear Models

- Rigde Regression.
- The LASSO.
- Elastic Net.

Ridge Regression

- Ridge Regression estimates the coefficients $\vec{\beta} = [\beta_0, \ldots, \beta_p]$ of the multiple linear regression through the following equation

$$J = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j (x_j) \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

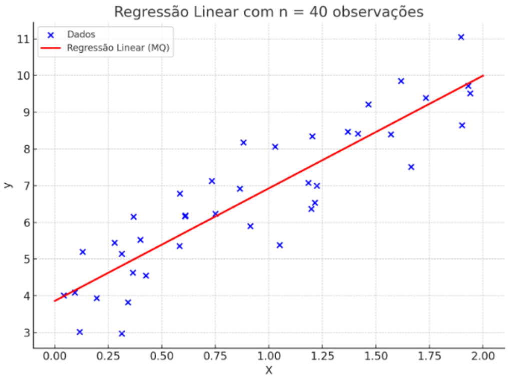
  where $\lambda \geq 0$ is a tuning (regularisation) parameter.
- In a more simplified view,

$$J = SSE + \lambda ||\vec{\beta}||_2$$

  where *SSE* is the sum of the squared errors (or residuals).

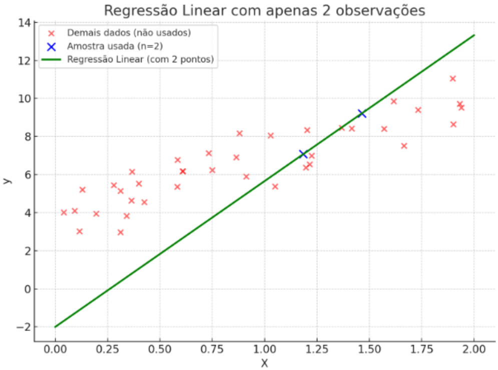
The Intuition Behind Ridge Regression

- Suppose that you estimate a Linear Regression model (*Least Squares*) with $n = 40$ observations.



The Intuition Behind Ridge Regression

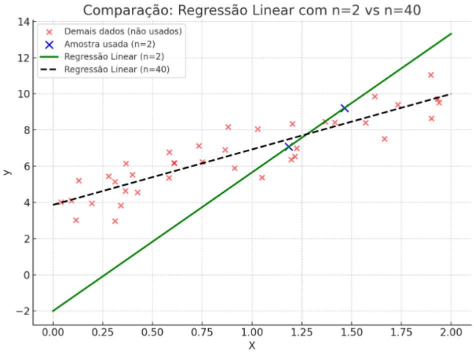
- Now, suppose that you selected $n = 2$ observations at random from this dataset and estimated another Least Squares model.



The Intuition Behind Ridge Regression

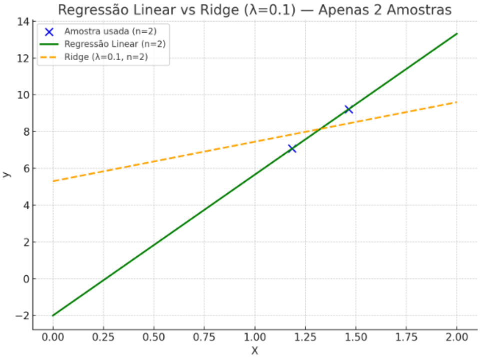
Comparing the two Lines estimated from Ordinary Least Squares:

- the green line has a $MSE_{train} = 0$ and a MSE_{test} higher than the dotted black line. Why?



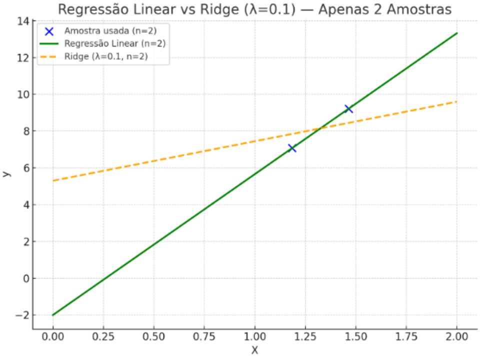
The Intuition Behind Ridge Regression

- Now, using the same $n = 2$ observations, let's estimate a Ridge Regression Line with $\lambda = 0.1$ and compare it with the corresponding Least Squares Line. What happened?



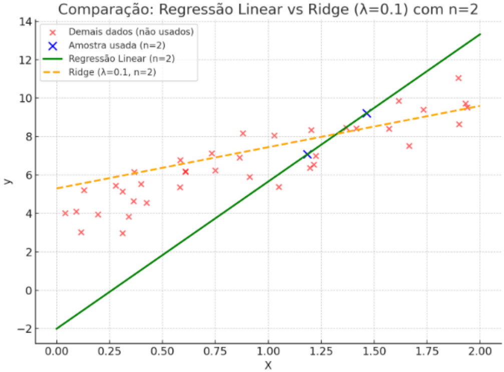
The Intuition Behind Ridge Regression

- the dashed yellow line doesn't fit the training data as well;
- ridge regression introduces a small amount of bias into how this line is fit.



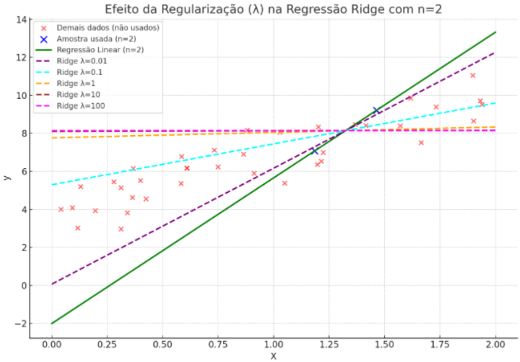
The Intuition Behind Ridge Regression

- Ridge Regression introduces a small amount of bias to reduce variance significantly.
- This also leads to a reduction in test error.



The Effect of λ in Ridge Regression

$$J = SSE + \lambda||\vec{\beta}||_2$$

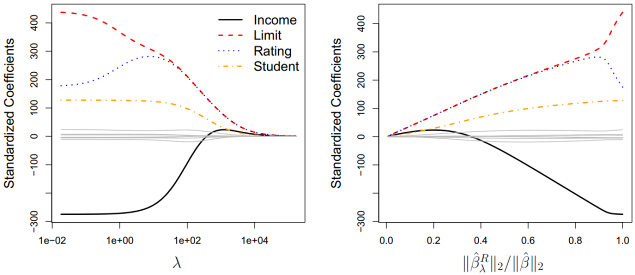


The Effect of λ in Ridge Regression

- The larger we make λ, the slope gets asymptotically close to 0.
- This means that the model predictions $\left(\hat{f}(x_i)\right)$ become less and less sensitive to the predictive variable X.
- How do we choose the value of λ?
  - try a range of different values for $\lambda = \{0, 0.1, \ldots, 10^4\}$.
  - use k-fold cross-validation to determine which one results in the lowest validation error.

Credit Data Set

- Ridge regression coefficients as a function of λ and $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$.



Why does Ridge Regression Improve over Least Squares?

- The bias-variance trade-off.
- simulated data with $n = 50$ observations and $p = 45$ predictors.

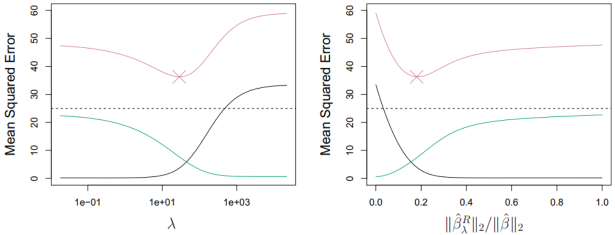


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

The LASSO

- Ridge Regression has one obvious disadvantage:
  - It includes ALL p predictors in the final model.
- The LASSO is an alternative to Ridge Regression that overcomes this disadvantage.

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

where $\lambda \geq 0$.
- In a more simplified view,

$$J = SSE + \lambda|\vec{\beta}|_1$$

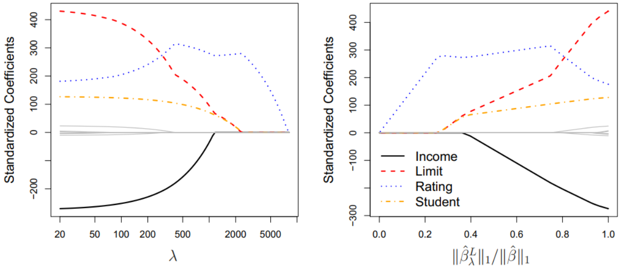
where $|\vec{\beta}|_1$ is the l_1 norm of the coefficient vector $\vec{\beta}$.

The LASSO

- The LASSO shrinks the coefficient estimates towards zero.
- However, the l_1 penalty has the effect of forcing some of the coefficients to be exactly equal to zero, when the tuning parameter is sufficiently large.
- Thus, much like subset selection, the LASSO performs feature selection.
- The LASSO yields sparse models.
- As is Ridge Regression, selecting a good value for λ is critical.

Credit Data Set

- LASSO coefficients as a function of λ and $|\hat{\beta}^R_\lambda|_1/|\hat{\beta}|_1$.



The Variable Selection Property of the LASSO

- Why is the LASSO results in coefficients estimates that are exactly equal to zero? And why Ridge Regression does not do that?
- The optimization problem solved by the LASSO

$$\min_{\vec{\beta}} \; SSE + \lambda \sum_{j=1}^{p} |\beta_j|$$

can also be written as

$$\min_{\vec{\beta}} \; SSE$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq \alpha$$

The Variable Selection Property of the LASSO

- Ridge Regression optimization problem

$$\min_{\vec{\beta}} \; SSE$$
$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leq \alpha$$

- The LASSO optimization problem

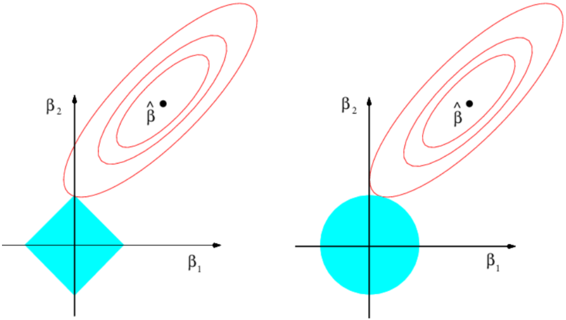
$$\min_{\vec{\beta}} \; SSE$$
$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq \alpha$$

Geometric Interpretation of the Optimization Problems

- The LASSO and Ridge Regression coefficient estimates (SSE contour lines) are given by the first point at which an ellipse intersects the constraint region (in blue).

$$\min_{\vec{\beta}} \; SSE$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq \alpha$$

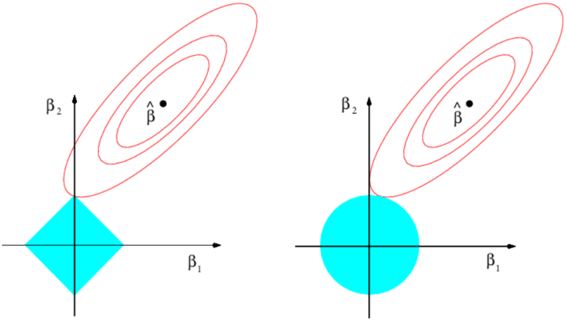


Geometric Interpretation of the Optimization Problems

- The LASSO constraint has corners at each of the axes. So, the ellipse will often intersects the constraint region at an axis, leading some coefficients to zero.

$$\min_{\vec{\beta}} \; SSE$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq \alpha$$



The LASSO x Rige Regression

- Ridge Regression can only shrink the coefficient estimates close to zero, while the LASSO can shrink these estimates all the way to zero.
- LASSO can exclude irrelevant variables, so it can be a little bit better than Ridge Regression at reducing the variance in models that contain a lot of irrelevant variables (predictors).
- In contrast, Ridge Regression tend to perform a little bit better when most variables are relevant.

Selection the parameter λ for Ridge Regression and the LASSO

- *K* - Fold Cross-Validation
  - Choose a grid of λ values and compute cross-validation error rate for each value of λ.
  - Then, select the tuning parameter λ for which the cross-validation error rate is smallest.
  - Finally, the model is re-fit using all the available observations and the selected value of the tuning parameter.

Referências bibliográficas

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Python*. 2023.
- Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference and prediction."The Mathematical Intelligence, 2001
- Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.
- Bishop, C. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc, 2006.