SF2950 Regression analysis

Exercise session 5 - Ch. 9: Multicollinearity (Ridge regression, principal component regression (PCR)), Ch. 10: Variable selection and model building

In class:

1. Montgomery et al., 9.24. Show that the ridge estimator is the solution to the problem
$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ \text{subject to} \quad & \boldsymbol{\beta}^T \boldsymbol{\beta} \leq d. \end{aligned}$$

2. Montgomery et al., 10.13. Suppose that the full model is
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$
$i = 1, \ldots, n$, where x_{i1} and x_{i2} have been coded so that (the sum of squares) $S_{11} = S_{22} = 1$. We will also consider fitting a subset model say
$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i.$$
   (a) Let $\hat{\beta}^*$ be the least squares estimate of β_1 in the full model. Show that $Var(\hat{\beta}_1^*) = \sigma^2/(1 - r_{12}^2)$, where r_{12} is the correlation between x_1 and x_2.
   (b) Let $\hat{\beta}$ be the least squares estimate of β_1 in the subset model. Show that $Var(\hat{\beta}_1) = \sigma^2$. Is β_1 estimated more precisely from the subset model or from the full model?
   (c) Show that $E[\hat{\beta}_1] = \beta_1 + r_{12}\beta_2$. Under what circumstances is $\hat{\beta}_1$ an unbiased estimate of β?
   (d) Find the mean squared error (MSE) for the subset estimator $\hat{\beta}_1$. Compare MSE($\hat{\beta}_1$) with $Var(\hat{\beta}_1^*)$. Under what circumstances is $\hat{\beta}_1$ a preferable estimator, with respect to MSE?

   You may find it helpful to reread Section 10.1.2.

3. James et al., Ch. 6.6.9. In this exercise, we will predict the number of applications received using the other variables in the College data set (`library(ISLR)`).
   a) Split the data set into a training set and a test set.
   b) Fit a linear model using least squares on the training set, and report the test error obtained.

c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

d) Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

e) Fit a PCR model on the training set, with M chosen by cross- validation. Report the test error obtained, along with the value of M selected by cross-validation.

g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

4. James et al., Ch. 6.8.8. In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

   a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

   b) Generate a response vector Y of length n = 100 according to the model

   $$Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

   where $\beta_0, \beta_1, \beta_2$ and β_3 are constants of your choice.

   c) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set containing both X and Y.

   d) Repeat (c), using forward stepwise selection and also using back- wards stepwise selection. How does your answer compare to the results in (c)?

   e) Now fit a lasso model to the simulated data, again using $X, X^2, \ldots, X^{10}$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

Recommended exercises:

Book	Theory	Implementation
Rawlings et al.:	7.1, 7.2, 7.3, 7.4	
James et al.:	Lab 2 Ch. 6.6.1	
Montgomery et al.		9.2, 9.4, 9.5, 9.19, 9.20, 9.21, 10.1, 10.5 10.14

References

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461471387.

D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9780470542811.

J.O. Rawlings, S.G. Pantula, and D.A. Dickey. *Applied Regression Analysis: A Research Tool*. Springer Texts in Statistics. Springer New York, 2001. ISBN 9780387984544.