Assignment 06

Shrinkage Methods

This goal of this assignment is to give you experience using shrinkage methods for alleviating interpretability problems that arise because of collinearity. Turn in a printed version of your responses to each of the questions on this assignment.

In questions that ask you to "use matrix algebra" to solve the problem, you can either show your syntax and output from carrying out the matrix operations, or you can use Equation Editor to input the matrices involved in your calculations.

In addition, please adhere to the following guidelines for further formatting your assignment:

- All graphics should be set to an appropriate aspect ratio and sized so that they do not take up more room than necessary. They should also have an appropriate caption.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document.
- Syntax or computer output should not be included in your assignment unless it is specifically asked for.

This assignment is worth 18 points.

Data Set

The data set you will use to answer the questions in this assignment contains simulated credit card data for 400 individuals. Specifically, the variables in the data set are:

- `balance`: Customer's average credit card balance (in dollars)
- `income`: Customer's reported income (in $10,000 dollars)
- `limit`: Credit limit issued to customer
- `rating`: Customer's credit rating; higher values indicate a better credit rating
- `cards`: Number of credit cards the customer has
- `age`: Customer's age
- `education`: Number of years of education

The goal of the analysis you are going to undertake in this assignment is to build a model that predicts customers' credit card balance. All of the variables included in the dataset have been previously shown to predict credit card balance.

These data were provided by: James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An introduction to statistical learning with applications in R*. New York: Springer-Verlag.

Exploratory Analysis

1. Create and report a correlation matrix of the outcome (`balance`) and the six predictors.
2. Based on the correlation matrix, comment on whether there may be any potential collinearity problems. Explain.
3. Fit the OLS model that regresses customers' credit card balance on the six predictors. (Don't forget to standardize any numeric variables prior to fitting the model.) Report the coefficient-level output, including the estimated coefficients, standard errors, t-valus, and p-values.

4. Compute and report the variance inflation factors.
5. Based on the VIF values, is there evidence of collinearity? Explain.
6. Compute the condition indices for the correlation matrix of the predictors. Based on these values, is there evidence of collinearity? Explain.
7. Which estimates from the coefficient-level output you reported in Question 3 are likely affected by the collinearity?

Shrinkage Method 1: Ridge Regression

In this section, you will carry out a ridge regression analysis to obtain better sampling variances for the coefficients.

8. Carry out a cross-validation to select the λ value to use in the ridge regression. Prior to carrying out this analysis, set the seed for the random number generation to 100. Based on the cross-validation results, report the λ value to be used in the ridge regression.
9. Fit the ridge regression model to the standardized credit data using the λ value you identified in Question #9. Report the equation for the fitted model based on the ridge regression.
10. Based on the coefficient estimates from the ridge regression model, which of the predictors seem important in explaining variation in customers' credit card balances? Explain.

Shrinkage Method 2: LASSO Regression

In this section, you will carry out a LASSO regression analysis to obtain better sampling variances for the coefficients. Recall that ridge regression minimized a penalized SSE based on the following:

$$\text{SSE}_{\text{Penalized}} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

LASSO regression also minimizes a penalized sum of squared error, but uses a slightly different penalty:

$$\text{SSE}_{\text{Penalized}} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid$$

To fit the LASSO model, we use the `cv.glmnet()` and `glmnet()` functions identically to how we used them to fit the ridge regression, with the exception that in both we specify the argument `alpha=1` to force a different penalty term. An advantage of the LASSO ovr ridge regression is that the LASSO will shrink coefficients all the way to zero in some cases, and in this way will also perform variable selection for you. (If this happens, the coefficient will be omitted from the `tidy()` output.) This is convenient since we do not have to rely on statistical inference to select predictors. (Remember that the SEs and hence the t- and p-values are not meaningful when shrinkage methods are used.)

11. Carry out a cross-validation to select the λ value to use in the LASSO regression. Prior to carrying out this analysis, set the seed for the random number generation to 100. Based on the cross-validation results, report the λ value to be used in the LASSO regression.
12. Fit the LASSO regression model to the standardized credit data using the λ value you identified in Question #11. Report the equation for the fitted model based on the LASSO regression.
13. Based on the coefficient estimates from the LASSO model, which of the predictors seem important in explaining variation in customers' credit card balances? Explain.

Standard Errors

14. Although they are not meaningful in practice, as an exercise, I still want you to compare the SEs from each of the three models (OLS, ridge regression, and LASSO regression). Create and report a table that allows a comparison of the standard error estimates for the coefficients estimated in each of the three models.

15. Compare and contrast the standard errors from the three models.

16. Based on your response to Question #7, explain why the comparisons of the standard errors from the OLS model to those from the shrinkage methods you just made in Question #15 make sense.