Assignment 06

Shrinkage Methods: Ridge Regression

This goal of this assignment is to give you experience using ridge regression 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 20 points.

Data Set

The goal of the analysis you are going to undertake in this assignment is to build a model that predicts customers' credit card balance. To do this you will use the data provided in the file *credit.csv* (see the data codebook). All of the variables included in the dataset have been previously shown to predict credit card balance.

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-values, and p-values.
4. Compute and report the condition number for the $\mathbf{X}^\top\mathbf{X}$ matrix. Show your work.
5. Based on the condition number of the $\mathbf{X}^\top\mathbf{X}$ matrix, is there evidence of collinearity? Explain.
6. Compute and report the VIF values for the standardized regression. Based on the VIF values, which estimates from the coefficient-level output you reported in Question 3 are likely affected by the collinearity? Explain.

Finding an Optimal d Value

In this section, you will find a d value to use in a ridge regression analysis to alleviate estimation problems associated with the near perfect collinearity.

7. Use the AIC to help you select the d value to use in the ridge regression. What is the value of d you will use in the ridge regression? Show your work.
8. Create a ridge trace plot that includes the values of d you examined in Question 7. Also include a guideline indicating the value of d chosen by the AIC.
9. Use the ridge trace plot you created to indicate the direction of bias for each of the coefficients. Explain.

Fitting the Ridge Regression Model

In this section, you will fit a ridge regression using your identified d value.

10. Use matrix algebra to compute the ridge regression coefficient estimates using the d value you identified in Question #7. Show your work.
11. Fit the ridge regression model to the standardized credit data using the d value you identified in Question #7 and the `glmnet()` function. Show your syntax (not the output) and report the fitted equation based on the ridge regression.

Coefficient-Level Summaries

12. Although they are not meaningful in practice, as an exercise, I still want you to compare the SEs from the standardized OLS and ridge regression models. Create and report a table that allows a comparison of the standard error estimates for the coefficients estimated in each of the two models.
13. Which coefficients saw the biggest reduction in their SEs? How could you predict this? Explain. (Hint: Revisit your response to Question #6.)
14. Create a coefficient-level regression table that reports the estimates, SEs, t-values, p-values, and confidence intervals for each of the predictors from the ridge regression model.
15. Compute and report the amount of bias in each of the coefficients. Show your work.
16. Compute the VIF values for each of the coefficients from the ridge regression model. The VIF value for the ith coefficient is computed as:

$$\text{VIF}(\hat{\beta}_i) = \text{R}_{i,i} \times \frac{\det(\text{R}_{-i,-i})}{\det(\text{R})}$$

where R is the standardized (correlation) matrix of the sampling variances and covariances of the ridge coefficients, $\text{R}_{i,i}$ is the element in the ith row and ith column of R, and $\text{R}_{-i,-i}$ is the matrix composed of all rows and columns of R except the ith. (Hint: You computed the unstandardized matrix of the sampling variances and covariances of the ridge coefficients in Question #12. To turn a variance–covariance matrix into a correlation matrix use the `cov2cor()` function.)

17. Based on the VIF values, have we eliminated the collinearity problems? Explain.

Model-Level Summaries

18. Compute and report the model-level R^2 for the ridge regression model. (Hint: Remember that the model-level R^2 is the squared correlation between the observed and predicted values of the outcome.) Show your work. How does this compare to the R^2 from the OLS model?

19. Compute and report the F-value associated with the R^2 value you computed in Question #18.

20. Compute and report the p-value associated with the test of whether $\rho^2 = 0$.