Assignment 06

Shrinkage Methods: Ridge Regression

Answer Key

This assignment is worth 20 points.

Exploratory Analysis

1. Create and report a correlation matrix of the outcome (`balance`) and the six predictors.

Table 1. *Correlation matrix for the credit data.*

Variable	1	2	3	4	5	6	7
1. Credit balance		.464	.862	.864	.086	.002	-.008
2. Income	.464		.792	.791	-.018	.175	-.028
3. Credit limit	.862	.792		.997	.010	.101	-.024
4. Credit rating	.864	.791	.997		.053	.103	-.030
5. Number of credit cards	.086	-.018	.010	.053		.043	-.051
6. Age	.002	.175	.101	.103	.043		.004
Tears of edcuation	-.008	-.028	-.024	-.030	-.051	.004	

2. Based on the correlation matrix, comment on whether there may be any potential collinearity problems. Explain.

The pairwise correlations between the income, credit limit, and credit rating variables are all quite high. This might indicate a potential collinearity issue.

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.

Table 2. *Coefficient-level output from the standardized regression model in which credit balance was regressed on a set of potential predictors.*

Coefficient	B	SE	t	p
income	-0.579	0.03	-19.79	< 0.001
limit	0.632	0.27	2.38	0.018
rating	0.694	0.27	2.60	0.01
cards	0.035	0.02	1.64	0.101
age	-0.033	0.02	-1.87	0.062
education	0.014	0.02	0.77	0.442

4. Compute and report the condition number for the $\mathbf{X}^\top\mathbf{X}$ matrix. Show your work.

```
# Create design matrix
X = z_credit[ , 2:7] %>%
  data.matrix()

# Compute eienvalues
e_values = eigen(t(X) %*% X)$values

# Compute condition number
max(e_values) / min(e_values)
```

```
## [1] 1262.212
```

5. Based on the condition number of the $\mathbf{X}^\top\mathbf{X}$ matrix, is there evidence of collinearity? Explain.

Yes, there is evidence of collinearity. The condition number is quite large indicating that there is likely collinearity collinearity among the predictors.

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.

```
##     income      limit     rating      cards        age  education
##   2.773276 228.848290 230.612596   1.433932   1.038541   1.008043
```

The sampling variances associated with credit limit and credit rating are extrodinarily inflated. The standard errors associated with those coefficients are over 30 times larger than they would be in a model where all the predictors are independent.

Finding an Optimal d Value

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.

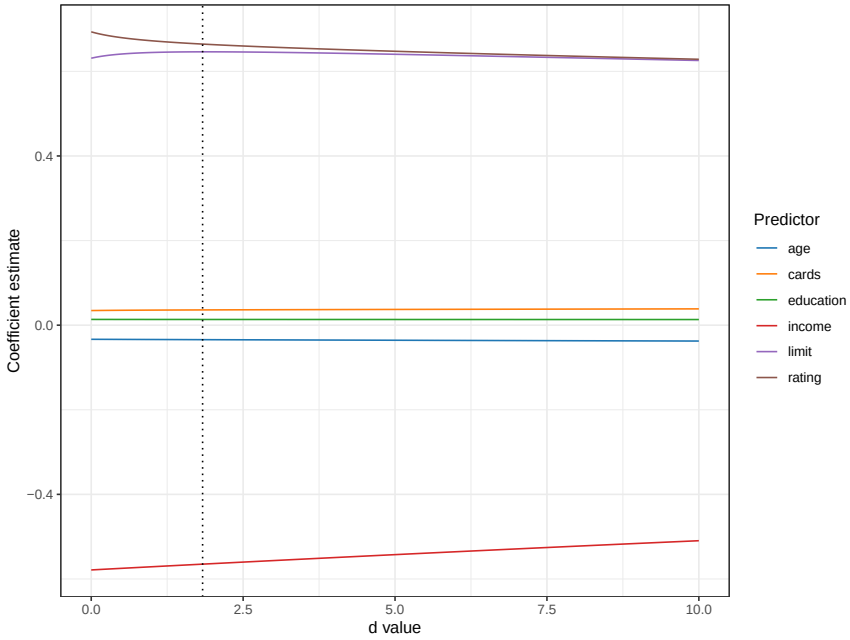
```
X = z_credit[ , 2:7] %>% data.matrix()
Y = z_credit[ , 1] %>% data.matrix()
d = seq(from = 0, to = 10, by = 0.001)


# FOR loop to cycle through the different values of d
aic = c()

for(i in 1:length(d)){

  b = solve(t(X) %*% X + d[i]*diag(6)) %*% t(X) %*% Y
  e = Y - (X %*% b)
  H = X %*% solve(t(X) %*% X + d[i]*diag(6)) %*% t(X)
  df = sum(diag(H))

  # Create and store the AIC value
  aic[i] = 400 * log(t(e) %*% e) + 2 * df
}

# Find d associated with smallest AIC
data.frame(d, aic) %>%
  filter(aic == min(aic))
```

```
##       d      aic
## 1 1.833 1564.124
```

The d value with the smallest AIC is $d = 1.833$.

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.

Figure 1. *Ridge trace plot of the coefficients for differetn d values. The vertical line is place at 1.833, the d value that has the lowest AIC.*



9. Use the ridge trace plot you created to indicate the direction of bias for each of the coefficients. Explain.

The number of credit cards, age, and education predictors show little difference in bias (≈ 0). The income predictor is biased upwards, while the credit limt and credit rating predictors are biased downward.

Fitting the Ridge Regression Model

10. Use matrix algebra to compute the ridge regression coefficient estimates using the d value you identified in Question #7. Show your work.

```
X = z_credit[ , 2:7] %>% data.matrix()
Y = z_credit[ , 1] %>% data.matrix()
b = solve(t(X) %*% X + 1.849 * diag(6)) %*% t(X) %*% Y
b
```

```
##                 [,1]
## income    -0.56551627
## limit      0.64698247
## rating     0.66514320
## cards      0.03608720
## age       -0.03433985
## education  0.01346353
```

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.

```
# Fit ridge regression
ridge.3 = glmnet(
  x = X,
  y = Y,
  alpha = 0,
  lambda = 1.833 / 400,
  intercept = FALSE,
  standardize = FALSE
  )

# Show coefficients
tidy(ridge.3)
```

```
## # A tibble: 6 x 5
##   term       step estimate  lambda dev.ratio
##   <chr>     <dbl>    <dbl>   <dbl>     <dbl>
## 1 income        1  -0.566  0.00458     0.878
## 2 limit         1   0.670  0.00458     0.878
## 3 rating        1   0.643  0.00458     0.878
## 4 cards         1   0.0370 0.00458     0.878
## 5 age           1  -0.0343 0.00458     0.878
## 6 education     1   0.0134 0.00458     0.878
```

The fitted equation is:

$$\begin{aligned}\hat{\text{Balance}}_i = &-0.566(\text{Income}_i) + 0.670(\text{Credit Limit}_i) + 0.643(\text{Credit Rating}_i) + 0.037(\text{Cards}_i) \\ &- 0.034(\text{Age}_i) + 0.013(\text{Education}_i)\end{aligned}$$

where all variables are standardized.

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.

Table 3. *Standard errors from the standardized regression model (OLS) and the ridge regression model with d=1.849.*

	OLS	Ridge
income	0.029	0.029
limit	0.266	0.087
rating	0.267	0.087
cards	0.021	0.018
age	0.018	0.018
education	0.018	0.018

13. Which coefficients saw the biggest reduction in their SEs? How could you predict this? Explain. (Hint: Revisit your response to Question #6.)

The ridge regression model shrunk the SEs for the credit limit and credit rating coefficients the most. This was predictable as the collinearity diagnostics pointed toward these coefficients as being problematic.

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.

```
b_ols = solve(t(X) %*% X) %*% t(X) %*% Y # OLS estimates
dI = 1.833 * diag(6) # Compute dI

# Estimate bias in ridge regression coefficients
bias = -1.833 * solve(t(X) %*% X + dI) %*% b_ols
round(bias, 3)
```

```
##             [,1]
## income     0.014
## limit      0.015
## rating    -0.029
## cards      0.002
## age       -0.001
## education  0.000
```

16. Compute the VIF values for each of the coefficients from the ridge regression model.

```
R = cov2cor(var_b)

VIF = c()

for (i in 1:6) {
    VIF[i] <- R[i, i] * det(R[-i, -i]) / det(R)
}
```

Table 4. *VIF values for the ridge regression coefficients (d=1.833).*

Predictor	VIF
income	2.71
limit	24.65
rating	24.82
cards	1.05
age	1.04
education	1.00

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

Not really. The VIF values for the credit limit and credit rating predictors are still quite high.

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?

```
# Compute R2 for the OLS
glance(lm.1)$r.squared
```

```
## [1] 0.8782453
```

```
# Compute R2 for ridge regression
b = tidy(ridge.3)$estimate
yhat = X %*% b
R2 = cor(Y, yhat)^2
R2
```

```
##           [,1]
## [1,] 0.8781954
```

The model-level R^2 is 0.878. This is virtually identical to the R^2 for the OLS model.

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

```
F = (R2) / (1 - R2) * df_residual / df_model
F
```

```
##          [,1]
## [1,] 536.6715
```

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

```
p = 1 - pf(F, df1 = df_model, df2 = df_residual)
p
```

```
##      [,1]
## [1,]    0
```