Econ 493 B1 - Winter 2023

Homework 1

Assignment Information

This assignment is due on Friday January 27 at 11:59 am.

Submit the assignment on eClass. Late assignments will receive **NO MARKS**.

Answers to computing exercises must include R commands and output files when applicable. All answers must be transcribed to your written answers which must be separate from the R printout.

Total marks = 50 (5 questions).

Exercise 1

Let $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$ where y_i is the number of hot dogs sold at an amusement park on a given day, x_i is the number of admission tickets sold that day, z_i is the daily maximum temperature, and u_i is a random error.

Answer the following questions.

a. State whether **each** of y_i, x_i, z_i, β_0, β_1, and β_2 is a coefficient or a variable.

b. Determine the units of β_0, β_1, and β_2 and describe the interpretation of each.

c. What are your expectations for the signs of β_0, β_1, and β_2 (negative, zero, positive or unsure)? Justify your answer.

d. Is it reasonable to allow for a non-zero intercept? Explain.

Exercise 2

You have a sample size $n = 1$ with data $y_1 = 2$ and $x_1 = 1$. You are interested in the value of β in the regression $y = \beta x + u$. Note there is no intercept.

a. Plot the sum of squared residuals $(y_1 - bx_1)^2$ as a function of b.

b. Show that the least squares estimate of β is $\hat{\beta}_{OLS} = 2$.

c. Using $\lambda_{ridge} = 1$, plot the ridge penalty term $\lambda_{ridge} b^2$ as a function b.

d. Using $\lambda_{ridge} = 1$, plot the ridge penalized sum of squared residuals $(y_1 - bx_1)^2 + \lambda_{ridge} b^2$.

e. Find the value of $\hat{\beta}_{ridge}$.

f. Using $\lambda_{ridge} = 0.5$, repeat (c) and (d). Find the value of $\hat{\beta}_{ridge}$.

g. Using $\lambda_{ridge} = 4$, repeat (c) and (d). Find the value of $\hat{\beta}_{ridge}$.

h. Use the graphs that you produced in (a)-(d) for the various values of λ_{ridge} to explain why a larger value of λ_{ridge} results in more shrinkage of the OLS estimate.

Exercise 3

You have a sample size $n = 1$ with data $y_1 = 2$ and $x_1 = 1$. You are interested in the value of β in the regression $y = \beta x + u$. Note there is no intercept.

a. Plot the sum of squared residuals $(y_1 - bx_1)^2$ as a function of b.

b. Show that the least squares estimate of β is $\hat{\beta}_{OLS} = 2$.

c. Using $\lambda_{lasso} = 1$, plot the lasso penalty term $\lambda_{lasso}|b|$ as a function b.

d. Using $\lambda_{lasso} = 1$, plot the lasso penalized sum of squared residuals $(y_1 - bx_1)^2 + \lambda_{lasso}|b|$.

e. Find the value of $\hat{\beta}_{lasso}$.

f. Using $\lambda_{lasso} = 0.5$, repeat (c) and (d). Find the value of $\hat{\beta}_{lasso}$.

g. Using $\lambda_{lasso} = 4$, repeat (c) and (d). Find the value of $\hat{\beta}_{lasso}$.

h. Use the graphs that you produced in (a)-(d) for the various values of λ_{lasso} to explain why a larger value of λ_{lasso} results in more shrinkage of the OLS estimate.

Exercise 4 (R)

We will now perform model selection using information criteria and k-fold cross-validation on a simulated data set.

Generate a simulated data set as follows:

```r
set.seed(1234)
n.obs <- 100
x1 <- rnorm(n.obs)
y <- x1 - 2*x1^2 + rnorm(n.obs)
```

a. Write out the model used to generate the data in equation form.

b. Create a scatterplot of x against y. Comment on what you find.

c. Compute the BIC and AIC for the following four models estimated using least squares. Which model would you prefer?

d. $y = \beta_0 + \beta_1 x + \varepsilon$

ii. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

iii. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

iv. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \varepsilon$

v. Compute the k-fold cross-validation errors that result from fitting the four models. Use $k = 5$. Which model would you prefer? Is this what you expected? Explain your answer.

e. Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on information criteria (c) and the cross-validation results (d)?

Exercise 5 (R)

The data set `males1987.csv` contains 545 observations of young working males in the United States with some professional and personal characteristics for the year 1987. Variable names and descriptions are provided in `males1987.txt`. We want to explain log-wages.

a. Create a matrix $\boldsymbol{X}$ (545×9) with the 7 explanatory variables described above plus experience and schooling squared. Scale the matrix $\boldsymbol{X}$ such that all variables have the same variance. Create a vector $\boldsymbol{y}$ (545×1) with log wage.

b. Plot the ridge regression standardized coefficients for different values of λ. Plot the lasso standardized coefficients for different values of λ. Comment on your results.

c. Estimate the parameters by OLS using the full sample. Which variables are statistically significant at the 10% level.

d. Perform best subset selection using the 9 explanatory variables and the full sample. Based on the BIC, which variables are included in the best model? Based on the AIC, which variables are included in the best model? Compare your results with OLS.

e. Estimate the lasso coefficients using the full sample. Are any of the coefficients forced to be exactly 0? Compare your results with OLS and best subset selection.

f. Re-estimate all the models (parts c, d, and e) using a training set with (about) half of the observations. Obtain the test sample errors for all the models. Comment on your results.