E401/M518: Problem Set 2b

Predicting Wages in the US

Fall 2023

Due: October 20 2023

Please work on the following questions and hand in your solutions in groups of at most 2 students. You are asked to answer all questions, but I will only select 2 (sub)questions randomly to grade.

In this problem set we will predict wages in the US. You have access to US census data from the CPS in the year 2012 on the individual-level. The dependent variable is the logarithm of the hourly wage. All other variables in the data denote socioeconomic characteristics, e.g., marital status, education, and experience. The data can be found in the package `hdm` under the name `cps2012`. First, consider the following 16 predictors: `female`, `widowed`, `divorced`, `separated`, `nevermarried`, `hsd08`, `hsd911`, `hsg`, `cg`, `ad`, `mw`, `so`, `we`, `exp1`, `exp2`, `exp3`.

Question 1: Load, prepare, and summarize the data.

```
library(hdm)
# This command loads data contained in a R-package.
data(cps2012)
?cps2012
# Construct a regressor matrix for use in the different models.
x <- model.matrix( ~ -1 + female + widowed + divorced + separated + nevermarried +
hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3, data=cps2012)
dim(x)
```

```
## [1] 29217    16
```

```
# Extract the dependent variable.
y <- cps2012$lnw
```

Question 2: Apply Ridge Regression with CV

Apply ridge regression to the previous dataset using the default grid of values for λ. Plot the 10-fold CV MSE as a function of λ. Afterwards, select the optimal λ by cross-validation. How many variables are used in this Ridge regression fit? Why is the test MSE for a Ridge

regression often smaller than for OLS when λ is not zero? What is the optimal value of λ? Is the unrestricted OLS optimal here (in the sense that it results in the lowest test MSE)? Explain your answer.

Question 3: Apply Lasso Regression with CV

Apply Lasso regression to the previous dataset for the the default grid of values for λ. Plot the 10-fold CV MSE as a function of λ. Then, select the optimal λ by cross-validation. What is the optimal λ? How many variables are used in the optimal Lasso fit? What are their coefficients? Is there a big difference between Ridge regression and Lasso (in terms of test MSE) in this application? Which method of prediction would you choose and why? Is `gender` an important factor in the prediction model? Interpret the coefficient on `female`.

Question 4: Using a more flexible model

Now suppose you want to predict wages with a more flexible model that allows all marginal effects to depend on gender. You would like to analyze the effect of gender and interaction effects of other variables with gender on wage jointly, i.e., in one model. The dependent variable is still the logarithm of the hourly wage. The new design matrix is given below. Repeat Questions 2 and 3 with this more flexible model.

```
X <- model.matrix( ~ -1 +female+
                     female:(widowed+divorced+separated+nevermarried+
                             hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3) +
                     + (widowed + divorced + separated + nevermarried +
                          hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3)^2,
                   data=cps2012)
dim(X)
```

```
## [1] 29217   136
```

```
# Safety check: Exclude all constant variables.
X <- X[,which(apply(X, 2, var)!=0)]
dim(X)
```

```
## [1] 29217   116
```

```
index.gender <- grep("female", colnames(X))
```

Question 5: What is the most preferred prediction model of all?

Explain your choice. Do the effects of gender on wages depend on education, i.e., is the interaction between gender and education important for predicting wages? In particular, does Lasso select this variable?