Universidad Carlos III de Madrid, Department of Economics
ECONOMETRICS I Fall 2020
Assingment 3, Due December 3th

1. **Predicting Wages in USA**. The problem is prediction of wages in the US. To that end, you collect US census data from the CPS in the year 2012. The dependent variable is the logarithm of the wage. All other variables denote some other socio-economic characteristics, e.g. marital status, education, and experience. The data can be found in the R package "hdm" (`install.packages("hdm")`) under the name cps2012. First, consider the 16 predictors female + widowed + divorced + separated + nevermarried + hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3.

(a) Load, prepare and summarize data.

(b) Apply Ridge-Regression with cross-validation (CV):

   (i) Apply ridge regression to the previous dataset for the the default grid of values of lambda. Plot the 10-fold CV MSE as a function of lambda.

   (ii) Then, select the optimal lambda (λ) by cross-validation. How many variables are used in the Ridge fit?

   (iii) Why is the test MSE for Ridge often smaller than for OLS when lambda is not zero?

   (iv) What is the optimal value of lambda? Is unrestricted OLS optimal here, in a test MSE sense?

(c) Apply LASSO with cross-validation (CV):

   (i) Apply Lasso regression to the previous dataset for the the default grid of values of lambda. Plot the 10-fold CV MSE as a function of lambda.

   (ii) Then, select the optimal lambda (λ) by cross-validation. What is the optimal lambda?

   (iii) How many variables are used in the optimal Lasso fit? What are their coefficients? Is there a big difference here between Ridge and Lasso (in terms of test MSE)?

   (iv) Which method of prediction would you choose and why? Is gender an important factor in the prediction model? Interpret the coefficient of female.

(d) Repeat (b) and (c) for a more flexible specification: You would like to analyse the effect of gender and interaction effects of other variables with gender on wage jointly. The dependent variable is still the logarithm of the wage.

(e) Based on the variable selection provided by Lasso, study the Gender Pay Gap (GPG) with this data set, including an Oaxaca-Blinder decomposition for the GPG.

2. Shao's Chapter 2: 105; Chapter 3: 101, 103(a,b,c,d), 107; Chapter 4: 96(even cases), 144.