Ekaterina Kazak

ECON61001: Econometric Methods

Semester 1

Programming Assignment # 6.

Please complete the following problems and submit a file named `PA6.R`.

Remember:

- Do not rename external data files or edit them in any way. In other words, don't modify `mroz_subset.csv`. Your code won't work properly on my version of that data set, if you do.
- Do not use global paths in your script. Instead, use `setwd()` interactively in the console, but do not forget to remove or comment out this part of the code before you submit. The directory structure of your machine is not the same as the one on Gradescope's virtual machines.
- Do not destroy or overwrite any variables in your program. I check them only after I have run your entire program from start to finish.
- Check to make sure you do not have any syntax errors. Code that doesn't run will get a very bad grade.
- Do not install or require any libraries unless specified in the task.
- Make sure to name your submission `PA6.R`

Tip: before submitting, it might help to clear all the objects from your workspace, and then `source` your file before you submit it. This will often uncover bugs. Before you start, download the file `mroz_subset.csv` from the Blackboard and save it in the same folder as the R script `PA6.R`. Make sure you set a correct working directory.

You are interested in estimating an earnings function for married women based on their educational and family background. The first column of the data frame contains the cross-sectional data on log of hourly wages, `lwage`, in US dollars for 428 women. The explanatory variables include:

Variable	**Definition**
`nwifeinc`	family income other than earnings of the woman
`exper`	years of labour market experience
`expersq`	squared `exper`
`city`	= 1 if lives in a city, = 0 else
`fatheduc`	father's years of schooling
`motheduc`	mother's years of schooling
`huseduc`	husbands's years of schooling
`educ`	years of schooling
`age`	womans's age in years
`kidslt6`	number of kids younger than 6 years old
`kidsge6`	number of kids 6-18 years old

1. Import the data stored in 'mroz_subset.csv' as a data frame with the name `df`.
2. Define a vector `Y` with dimensions (428×1), containing log of hourly wages, `lwage`, using the function `as.matrix`.
3. Define a variable `n0` containing the number of rows of the vector `Y`.
4. Define a matrix `X` with dimensions (428×12), containing all the explanatory variables from `df`, using the function `as.matrix`. Do not forget to include the intercept in the model by making sure the first column of the `X` matrix is a column of ones.
5. Calculate the OLS solution for the linear regression model $Y = X\beta + e$, where Y and X are defined in 2. and 4. correspondingly. Store the OLS solution as a 12×1 vector with the name `b0`.
6. Calculate how many elements of `b0` are larger than 0.01 in absolute value. Save this number as `nz0`.
7. Using the function `seq` define the sequence of parameters $\lambda = \{10, 11, 12, ..., 499, 500\}$ and save it under the name `lam`.

   Next, examine the Ridge regression estimator $\tilde{\beta}(\lambda) = (X'X + \lambda I_k)^{-1}X'y$. For each value of λ:

   (a) compute the Ridge estimator using `Y` and `X` defined in 2. and 4. correspondingly.

   (b) compute how many elements of the Ridge estimator are larger than 0.01 in absolute value. Save this number.

   (c) calculate the regression residuals $\hat{e} = Y - X\hat{\tilde{\beta}}(\lambda)$. Save the estimate of the error term variance $\sum \hat{e}^2/(n-k)$.
8. After the loop described above, using the function `c( )` define a variable `a1` which reports the number of non-zero elements of the ridge estimator for $\lambda = 30$ and $\lambda = 460$. Thus `a1` should consist of 2 values, combined with the function `c( )`.
9. Next, define a variable `b1` which reports the distance between the Ridge estimator $\hat{\tilde{\beta}}(\lambda)$ and OLS solution $\hat{\beta}_0$ for $\lambda = 30$ and $\lambda = 460$. The distance can be measure by the simple element-wise sum of the squared distance $\sum_{j=1}^{k}(\hat{\tilde{\beta}}_j(\lambda)) - \hat{\beta}_{0,j})^2$. Thus `b1` should consist of 2 values, combined with the function `c( )`.
10. Next, define a variable `c1` which reports the estimates of the error term variance for $\lambda = 30$ and $\lambda = 460$. Thus `c1` should consist of 2 values, combined with the function `c( )`.

    Repeat the above analysis of the Ridge estimator by restricting the dataset to $n = 100$ and $n = 15$, i.e. take only first n observations for estimation in X and Y correspondingly. Note, you have to re-estimate the model for every value of λ, but the OLS solution remains fixed as `b0` defined in 5.

11. For the ridge regression estimated on the first 100 observations, using the function `c( )` define a variable `a2` which reports the number of non-zero elements of the ridge estimator for $\lambda = 30$ and $\lambda = 460$. Thus `a2` should consist of 2 values, combined with the function `c( )`.

12. Next, define a variable `b2` which reports the distance between the Ridge estimator $\hat{\hat{\beta}}(\lambda)$ and OLS solution $\hat{\beta}_0$ for $\lambda = 30$ and $\lambda = 460$. The distance can be measure by the simple element-wise sum of the squared distance $\sum_{j=1}^{k}(\hat{\hat{\beta}}_j(\lambda)) - \hat{\beta}_{0,j})^2$. Thus `b2` should consist of 2 values, combined with the function `c( )`.

13. Next, define a variable `c2` which reports the estimates of the error term variance for $\lambda = 30$ and $\lambda = 460$. Thus `c2` should consist of 2 values, combined with the function `c( )`.

14. For the ridge regression estimated on the first 15 observations, using the function `c( )` define a variable `a3` which reports the number of non-zero elements of the ridge estimator for $\lambda = 30$ and $\lambda = 460$. Thus `a3` should consist of 2 values, combined with the function `c( )`.

15. Next, define a variable `b3` which reports the distance between the Ridge estimator $\hat{\hat{\beta}}(\lambda)$ and OLS solution $\hat{\beta}_0$ for $\lambda = 30$ and $\lambda = 460$. The distance can be measure by the simple element-wise sum of the squared distance $\sum_{j=1}^{k}(\hat{\hat{\beta}}_j(\lambda)) - \hat{\beta}_{0,j})^2$. Thus `b3` should consist of 2 values, combined with the function `c( )`.

16. Next, define a variable `c3` which reports the estimates of the error term variance for $\lambda = 30$ and $\lambda = 460$. Thus `c3` should consist of 2 values, combined with the function `c( )`.