Problem Set #2

MACS 30250, Dr. Evans
Due Monday, May. 11 at 1:30pm

1. **Parallel computing versus serial computing a bootstrapped cross validation (10 points).** In Exercise 3 of Problem Set 6 of the MACS 30150 (Winter 2020) Perspectives on Computational Economic Modeling class, you estimated a multivariable logistic model and evaluated its fit using the validation set approach (one training set and one test set). For this exercise, you will use the same Auto.csv file. This dataset includes 397 observations on miles per gallon (`mpg`), number of cylinders (`cylinders`), engine displacement (`displacement`), horsepower (`horsepower`), vehicle weight (`weight`), acceleration (`acceleration`), vehicle year (`year`), vehicle origin (`origin`), and vehicle name (`name`). We will study the factors that make miles per gallon high or low. Create a binary variable `mpg_high` that equals 1 if `mpg_high`$\geq$ median(`mpg_high`) and equals 0 if `mpg_high`$<$ median(`mpg_high`). Create two indicator variables for vehicle origin 1 (`orgn1`) and vehicle origin 2 (`orgn2`). [This leaves the third origin category out as the excluded indicator variable.]

$$Pr(mpg\_high = 1|\boldsymbol{X\beta}) = \frac{e^{\boldsymbol{X\beta}}}{1 + e^{\boldsymbol{X\beta}}}$$
$$\text{where} \quad \boldsymbol{X\beta} = \beta_0 + \beta_1 cyl_i + \beta_2 dspl_i + \beta_3 hpwr_i + \beta_4 wgt_i + \beta_5 accl_i + \beta_6 yr_i + \beta_7 orgn1_i + \beta_8 orgn2_i$$

   (a) Using serial computation, perform an estimation of the logistic model on 100 bootstrapped training sets (with replacement) on random draws of training sets of 65% of the data. Use `sklearn.linear_model.LogisticRegression()` function and make sure that the `n_jobs` option is set to `None` or `1`. This will guarantee that it runs in serial. Compute the error rate for each of the 100 test sets. Calculate the average error rate. Make sure to set the seed on each of the 100 random draws so that these draws can be replicated in part (b). What is your error rate? How long did this computation take?

   (b) Now write a function that takes as arguments the bootstrap number (1 through 100 or 0 through 99), random seed, and the the data, and estimates the logistic model on 65% of the data and calculates an error rate on the remaining 35%. Use `Dask` to parallelize these bootstraps. What is your error rate from this parallelized list of error rates? It should be the same as part (a). How long did this computation take?