Problem Set #2

MACS 30200, Dr. Evans and Dr. Soltoff
Due Monday, Apr. 30 at 11:30am

1. **2D kernel density estimator (5 points).** The data BQmat_orig.txt is a 78×7 matrix of percentages representing the values of a two-dimensional histogram of the percent of the U.S. population that receives all the bequests (inheritances) by a recipient's age (ages 18 to 95, rows) and by a recipient's lifetime income group (7 categories, columns). The seven lifetime income groups are percentiles. Let $prcntl_j$ be the percent of the population in lifetime income group j. The lifetime income groups in the $J = 7$ columns of the BQmat_orig.txt data are the following.

$$\boldsymbol{prcntl} = \big[0.25, 0.25, 0.20, 0.10, 0.10, 0.09, 0.01\big], \quad \text{such that} \quad \sum_{j=1}^{7} prcntl_j = 1$$

You can read this file into memory using the `numpy.loadtxt` function.

```
bq_data = np.loadtxt('BQmat_orig.txt', delimiter=',')
```

So the $[11, 5]$-th element of the `bq_data` matrix represents the percent of total bequests (inheritances) received by age-28 and lifetime income group $j = 5$ (80th to 90th percentile of lifetime income).

   (a) Read in the bequests data as a 78×7 `NumPy` array. Plot the 2D empirical histogram of these data as a 3D surface plot with age and income group on the x-axis and y-axis and the histogram density on the z-axis using a 3D surface plot tool (not a 3D bar histogram tool). Make sure that the axes are labeled correctly. And make sure that your 3D histogram is presented from a perspective that allows a viewer to see that data (don't let the data be hidden by a poor angle of the plot.)

   (b) Fit a bivariate kernel density estimator to the data. Use a Gaussian kernel. Choose a bandwidth parameter λ that you think is best. Justify your choice of that parameter. Plot the surface of your chosen kernel density estimator. Make sure that the axes are labeled correctly. And make sure that your 3D histogram is presented from a perspective that allows a viewer to see that data. What is the estimated density for bequest recipients who are age 61 in the 6th lifetime income category ($j = 6$, 90th to 99th percentile).

2. **Interaction terms (5 points).** `biden.csv` contains a selection of variables from the 2008 American National Election Studies. Estimate the following linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{1}$$

where Y is the Joe Biden feeling thermometer, X_1 is age, and X_2 is education. Report the parameters and standard errors.

(a) Evaluate the marginal effect of age on Joe Biden thermometer rating, conditional on education. Consider the magnitude and direction of the marginal effect, as well as its statistical significance.

(b) Evaluate the marginal effect of education on Joe Biden thermometer rating, conditional on age. Consider the magnitude and direction of the marginal effect, as well as its statistical significance.

For both elements, provide both a written answer and supporting graphical evidence.

3. **Parallel computing versus serial computing a bootstrapped cross validation (bonus).** In Exercise 3 of Problem Set 4 of the MACS 30100 Perspectives on Computational 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`).

$$Pr(mpg\_high = 1 | \boldsymbol{X\beta}) = \frac{e^{\boldsymbol{X\beta}}}{1 + e^{\boldsymbol{X\beta}}}$$

$$\text{where} \quad \boldsymbol{X\beta} = \beta_0 + \beta_1 cyl_i + \beta_2 dspl_i + \beta_3 hpwr_i + \beta_4 wgt_i + \beta_5 accl_i + \beta_6 yr_i + \beta_7 orgn_i$$

(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. 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?