ER190C: Data, Environment and Society

In class midterm exam

October 18, 2018

Instructions:

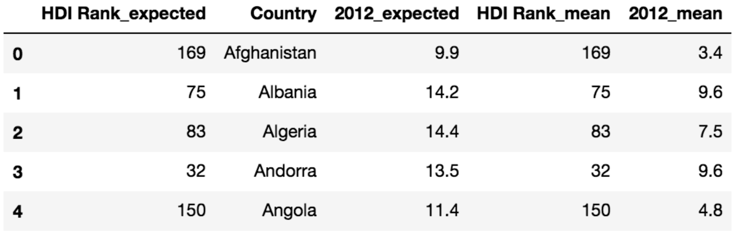
1. You have 80 minutes to take this test.
2. This booklet should have 7 problems and 9 pages (including this cover sheet). If it doesn't, you must let us know before you begin the exam.
3. There are 76 points total.
4. You may write on both sides of the paper in this booklet. Please be clear about where your work is continued.
5. You must show your work to receive full credit.
6. We will award partial credit.
7. **Write your name here**: ______________________________

Good luck!

Do not write below this line

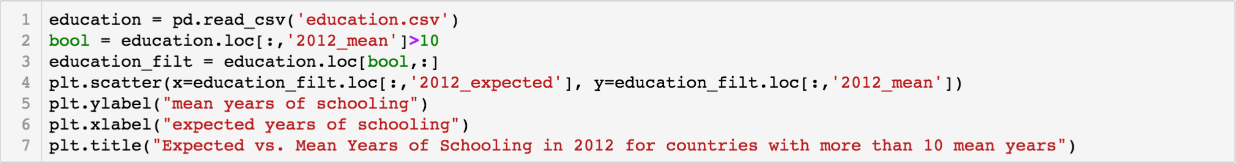
Problem	Max points	Points received
1	13	
2	14	
3	8	
4	14	
5	9	
6	12	
7	6	

1. Python and Pandas. Suppose you have a .csv file called “education.csv”. If you load it in Pandas and call `education.head()`, you get the following output:



	HDI Rank_expected	Country	2012_expected	HDI Rank_mean	2012_mean
0	169	Afghanistan	9.9	169	3.4
1	75	Albania	14.2	75	9.6
2	83	Algeria	14.4	83	7.5
3	32	Andorra	13.5	32	9.6
4	150	Angola	11.4	150	4.8

   (a) (7 points) Explain what each line of this code block does. (1 point per line)



```
education = pd.read_csv('education.csv')
bool = education.loc[:,'2012_mean']>10
education_filt = education.loc[bool,:]
plt.scatter(x=education_filt.loc[:,'2012_expected'], y=education_filt.loc[:,'2012_mean'])
plt.ylabel("mean years of schooling")
plt.xlabel("expected years of schooling")
plt.title("Expected vs. Mean Years of Schooling in 2012 for countries with more than 10 mean years")
```

   (b) (2 points) If you fired up a blank Jupyter notebook and pasted in the code block above, what additional code would you need to add to make it work?

   (c) (4 points) Describe how you would modify the code block above to generate a scatter plot of the *difference* between expected education (‘`2012_expected`’) and mean education (‘`2012_mean`’) versus ‘`HDI Rank_expected`’, for all countries with mean education levels below 6 years.

2. Energy and development reading.

   (a) (2 points) In Alstone *et al*, *Nature Climate Change* (2015), the authors argue in favor of what technology solution to electrify low-income people living off-grid?

   (b) (2 points) Name and describe two key variables that Alstone *et al* use to make their argument. Hint: they were very similar to data we used on HW4.

   (c) (2 points) In Lee *et al*, *American Economic Review* (2016), the authors argue in favor of what technology solution to electrify low income people living off the grid?

   (d) (2 points) Name and define one key data source that Lee *et al* use to make their argument.

(e) (2 points) Describe how some of the data listed above have finer granularity (i.e. higher resolution, or less aggregated) than other data listed above. (Note, this question is not asking about *scope,* i.e. the range of time or number people or events covered in the data set.) Be sure to explain in what way its granularity is finer (i.e. in what 'dimension' is the resolution higher?). Multiple answers are possible, but you need to describe just one. Justify your answer.

(f) (2 points) Describe how at least one of the data sets used in your answers above is broader in *scope* than others, i.e. encompassing a larger range of some variable (even if it is more aggregated). Note there is not a single correct answer, but you must justify your answer.

(g) (2 points) Suggest a way to change the scope and / or granularity of the data you described for (i) Alstone *et al* and (ii) Lee *et al* (one suggestion for each paper). The suggestion should improve the generalizability of the paper's conclusions, or provide further insight into the processes at play.

3. Land use regression.

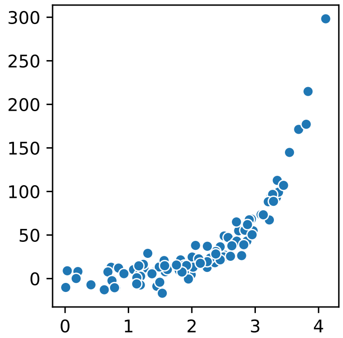
   (a) (4 points) Define "land use regression" (LUR), such as what we learned about in Novotny *et al*. Is it used for prediction or inference?

   (b) (2 points) Describe how LUR might be used by a policy-maker. There are many possible answers, you just need to give one.

   (c) (2 points) Describe why LUR predictions should be interpreted with caution. Your answer could be based on the details of the Novotny *et al* approach, or LUR in general.

4. Visualization

   (a) (2 points) Suggest a data transformation that would make this plot easier to interpret for small y-values. Explain your answer.



   (b) (2 points) List two things that would improve the above plot's provision of context.

   (c) (6 points) Using the following box and whisker plot (Lacke *et al*, "Aerosols and associated precipitation patterns in Atlanta", *Atmospheric Environment* (2009)), do your best to draw probability distributions for Monday, Tuesday and Friday concentrations. Put them on the same plot (with good labeling so we can tell them apart). You might want to sketch a few on the back of this sheet before drawing the final one here. You don't need to put numeric values on the plot. .

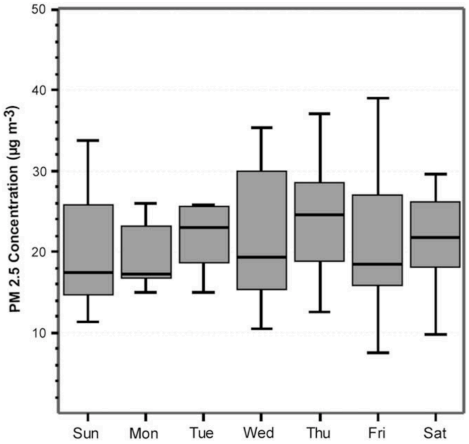


Fig. 2. Box-and-whisker plot of the daily PM 2.5 concentration for the Decatur, Georgia, station during the summers of 2003–2004 with an MT air mass present. The plots represent the interquartile range (shaded), the median (thick line), and outliers (the 10th and 90th percentiles as whiskers). The average PM 2.5 concentration per day is indicated along the *y*-axis.

   (d) (4 points) Both *average* and *extreme* values of PM2.5 matter for human health. Based on the box and whisker figure above, on which days should you avoid physical exertion? Justify your answers.

5. For this problem we'll work with a simple model (the "constant model"):

$$y_i = \beta_0 + \epsilon_i$$

We're going to estimate that model using:

$$\hat{y}_i = \hat{\beta}_0$$

(a) (3 points) Write down a mean squared error loss function for this model.

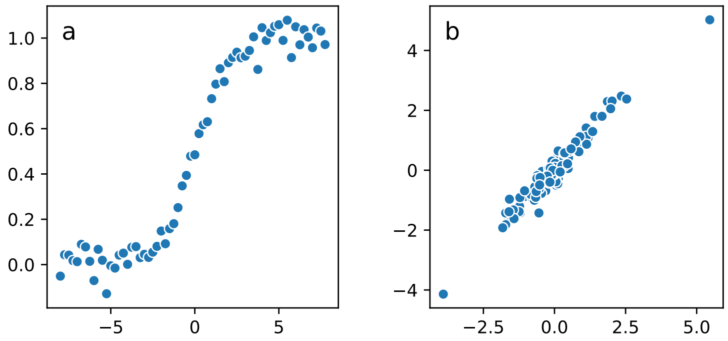
(b) (4 points) Use the loss function to derive an expression (a mathematical formula) for the optimal parameter for this model.

(c) (2 points) Using a data set $(x, y) = \{(0,0), (1,2), (2,4)\}$, compute the optimal β_0.

6. Modeling

(a) (4 points) Give one reason you might prefer to use a mean squared error loss function over a mean absolute error loss function. Give another reason you might prefer MAE over MSE. Justify your answers.

(b) (4 points) Look at the data in plots (a) and (b) below. Draw and label *approximations* of the fit lines that would result from running k-nearest neighbors (using a moderate-sized K, e.g. 10-20) and linear regression on each data set. Each plot should have one KNN line and one linear regression line.



(c) (4 points) Explain which model (KNN or linear regression) is better for each plot. Defend your answer in terms of model bias.

7. Eliza P. Annaswamy is a bureaucrat at a U.S. federal agency overseeing policies to reduce air pollution. She was just handed a model built to predict ozone concentrations in parts of the country where she doesn't have air quality sensors in place.

   (a) (3 points) Suggest how her model might help to reduce how much her agency spends to improve the overall health of the U.S. population.

   (b) (3 points) Is your suggestion an example of resource allocation or policy evaluation? Justify your answer.