ER190C: Data, Environment and Society

In class midterm exam

October 18, 2018

SOLUTIONS

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.

Grade distribution (mean YYY):

---

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)

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

i. Loads in csv to data frame
ii. Creates a series of TRUE/FALSE values; TRUE for countries with years of schooling greater than 10.
iii. Creates a new data frame with only those countries that have years of schooling greater than 10.
iv. Scatter plot for expected years of schooling versus mean years of schooling.
v. Labels vertical axis
vi. Labels horizontal axis
vii. Titles figure.

(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?

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
```

(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.

```
education = pd.read_csv('education.csv')
bool = education.loc[:,'2012_mean']<6
education_filt = education.loc[bool,:]
diff = education_filt.loc[:,'2012_expected'] - education_filt.loc[:,'2012_mean']
plt.scatter(x=education_filt.loc[:,'HDI Rank_expected'], y=diff)
plt.ylabel("difference between expected and mean")
plt.xlabel("HDI rank")
plt.title("Difference between expected and mean years of schooling")
```

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?

Decentralized energy systems such as solar lanterns and solar home systems.

   (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.

- HDI: Human development indicator. A weighted sum of per capita income, years of schooling, and life expectancy. Measured at the level of a country.
- Electrification rates: the fraction of the population with access to electricity. Measured at the level of a country.

   (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?

The grid: centralized generation extending out to customers via transmission and distribution.

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

Household survey responses: electrical appliances owned and desired, for Kenyans living on the grid, with solar home systems, and living off the grid.

(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.

Lee *et al* have higher resolution data across the population. They have data on individual people, whereas Alstone *et al*'s data have data on people aggregated at the country level.

(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.

Alstone *et al*'s data cover far more people – even though the data are more heavily aggregated. And Alstone *et al*'s data cover more years of time.

(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.

There are many possible answers here. For example, Alstone's paper would be more meaningful if the data were not aggregated to the country level – i.e. if the granularity was finer. This would provide more insight into how energy technology adoption supports (or doesn't) economic development, health and / or education. Lee *et al*'s paper would be more generalizable of the surveys extended beyond a handful of villages in Kenya.

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?

   From Novotny *et al:* "Land-use regression is an empirical-statistical technique that uses in situ concentration measurements and information about surrounding land-uses to estimate concentrations for nonmeasurement locations." It's used for prediction in spatial dimensions.

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

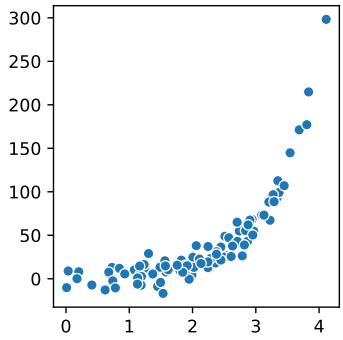
   She could use it to decide where to deploy additional monitors, or where to spend money on pollution reduction equipment.

   (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.

   The predictions are imperfect – carrying far more error than an actual pollution sensor would. Novotny *et al* don't consider non-traffic sources of NO2 in their set of predictors.

4. Visualization

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



You could log-transform the y-axis (vertical axis) data. This would "expand" the axis where values are small and "compress" it where values are large. However, note some values are negative, so you might need to shift the data upward before transforming it. (I did not take off points for not noticing this.) Another valid answer is a $(\cdot)^{\frac{1}{3}}$ transformation.

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

i. x-axis label
ii. y-axis label
iii. title
iv. a caption describing the data.

(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. .

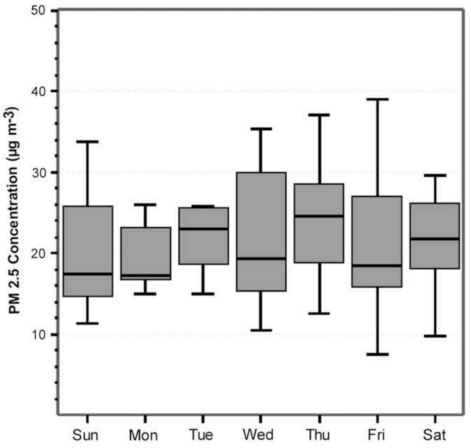
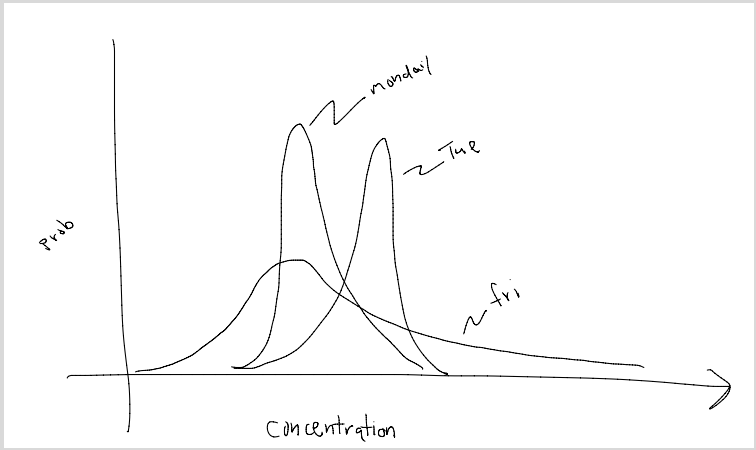


Fig. 2. Box-and-whisker plot of the daily PM 2.5 concentration for the Decatur, Georgia, station during the summers of 2003–2004 with an MT air mass present. The plots represent the interquartile range (shaded), the median (thick line), and outliers (the 10th and 90th percentiles as whiskers). The average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

Friday has the most extreme high values. It is not clear which days have the highest average because the box and whisker plots only show the median.

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.

$$\text{loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0)^2$$

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

$$\frac{\partial(\text{loss})}{\partial \beta_0} = -\frac{1}{n}\cdot 2\sum(y_i - \hat{\beta}_0) = 0 \quad \text{(at the optimal } \hat{\beta}_0\text{)}$$

$$\Rightarrow \quad \sum \hat{\beta}_0 = \sum y_i$$

$$n\hat{\beta}_0 = \sum y_i$$

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i = \bar{y} \quad \text{(the average)}$$

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

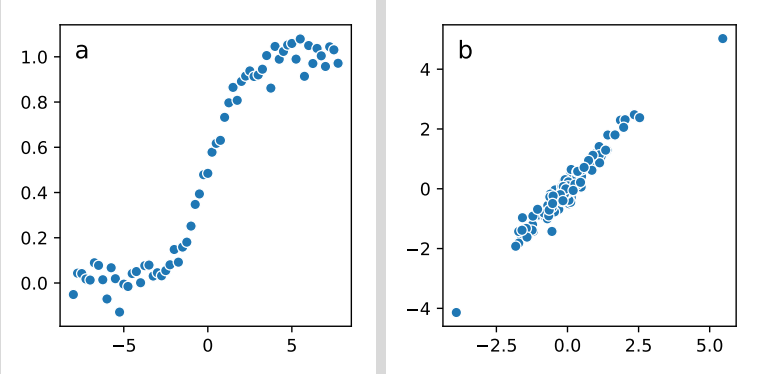
$$\frac{1}{n}\sum y_i = \frac{1}{3}(0 + 2 + 4) = 2$$

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.

- MSE is differentiable – you can solve for optimal parameters analytically, and basic gradient descent works
- MAE weights errors of all size equally, so it doesn't "chase" extreme values as much as MSE does.

(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.

- KNN is better for (a), because the data are clearly nonlinear. Linear regression would have significant bias (bias high for low x values and bias low for high x-values).
- Linear regression is better for (b). KNN would be biased high for low x values and biased low for high x values.

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.

   She could deploy sensors just in the places with the highest predicted concentrations, rather than everywhere.

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

   Resource allocation. Assuming she has limited budget, she is choosing where to allocate her spending efforts.