Least Squares Regression Workshop CS 480

Write your answers below. When you are finished, turn this page in, but also share your Python code with me. If you are using Google Colab, share your code with: `lins.brian@gmail.com`

High Bridge half-marathon

In this exercise, you will create a model to predict how long it will take a runner to finish a half-marathon in minutes based on two variables: age and gender. The data comes from the results of the 2018 Farmville High Bridge half-marathon. Your model will have the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

where b_0, b_1, and b_2 are constants and x_1 is the age of a runner, and x_2 is gender (0 for male, 1 for female). To get the data, we will use the `read_excel` function from the Pandas library. The code below will store the data from each column of the Excel spreadsheet in a Python list.

```
import numpy as np
import pandas as pd

df = pd.read_excel("https://people.hsc.edu/faculty-staff/blins/classes/spring23/math121/halfmarathon.xlsx")
genders = list(df.Gender)
ages = list(df.Age)
minutes = list(df.Minutes)
```

1. Use the data to create two numpy arrays. First, construct a matrix X with 3 columns, one where every entry is 1.0, one the contains the runners' ages and one with either 1.0 for female, and 0.0 for male in each entry. Second, construct the numpy array y which contains the runners' race times in minutes.

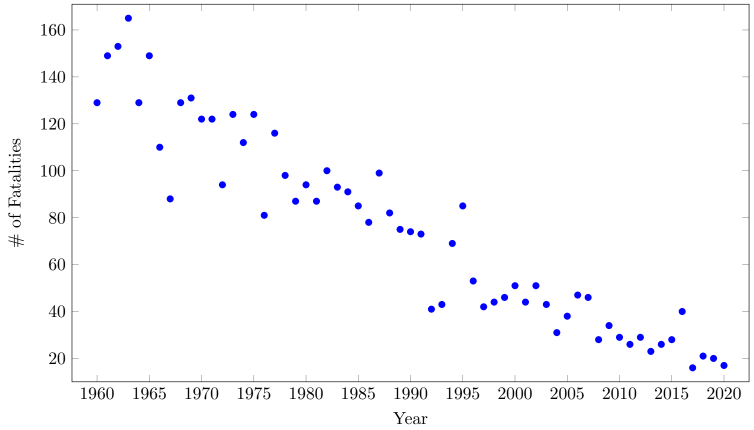
2. Numpy has a function to make it easy to find the least squares solution β that makes $\hat{y} = X\beta$ as close as possible to y (it minimizes the sum of square errors $\|\hat{y} - y\|^2$). The function is `np.linalg.lstsq`. To use it, enter: `np.linalg.lstsq(X,y)[0]`

   Note: You need to add the `[0]` since the function returns multiple values including the solution β, the sum of squared error $\|\hat{y} - y\|^2$, and two other linear algebra things that we don't actually care about.

3. Use your model to predict the race times for a typical 50 year old man and for a typical 30 year old woman.

4. How much difference does gender make? How much do runners slow down as they get older? Explain what the model says.

Lightning fatalities



The striking downward trend in lightning fatalities over the last 60 years almost looks like it follows a straight line. But it can't continue to follow a linear trend forever. An exponential decay model

$$\hat{y} = Ce^{\alpha x}$$

to predict fatalities $\hat{y}$ based on year x would probably be more accurate. We can use least squares regression to find the optimal constants C and α if we take the natural log of both sides and then look for the constants that minimize the sum of squared errors in the resulting formula:

$$\log \hat{y} = \log C + \alpha x.$$

5. Find the optimal constants C and α by using the `np.linalg.lstsq` function to find the least squares solution to

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \log C \\ \alpha \end{pmatrix} \approx \begin{pmatrix} \log y_1 \\ \log y_2 \\ \vdots \\ \log y_n \end{pmatrix}.$$

   You can get the data for the problem with the commands:

```
df2 = pd.read_excel("http://people.hsc.edu/faculty-staff/blins/StatsExamples/Lightning.xlsx")
years = np.array(df2.year)
deaths = np.array(df2.deaths)
logDeaths = np.log(deaths) # notice that functions work elementwise on np.arrays.
```

6. How many lightning fatalities does the model predict will happen this year?