1 Data & Bias

(a) **(6 points)** Your friend working at UCLA housing has been given the task of determining how students feel about the new dorms. Your friend wants to accomplish this by scraping Twitter and news article comment sections for tweets containing keywords and hashtags related to the new dorm and then running them through a model that does sentiment analysis. This model contains an algorithm that says whether the text exhibits positive, neutral, or negative sentiment. What kinds of selection bias might your friend's data collection method exhibit? (Refer to Week 1 Lecture 2, slide 39 for a list).

(b) **(6 points)** In 2018, news reports came out about how Amazon tried to use an AI tool to assist in the hiring process, but this tool was scrapped due to it displaying bias against hiring women. (i) Explain why the tool was discriminating against women? (ii) The prediction of the AI program was based on the CV of the candidates, including their gender, the name of their university, previous positions, extra-curricular activities, etc. Amazon modified the AI program to make it not consider the gender of the candidate, but this was not enough to eliminate the bias. Explain why dropping the gender could not eliminate the bias.

2 KNN regression

Consider the following data points (x, y) = (0, 1), (1, 0), (2, 5), (3, 2), (4, 5).

(a) **(5 points)** Draw the prediction for all x from [0, 4] using 1-NN, 3-NN, 5-NN regression based on those data points.

(b) **(3 points)** Given a test data set (0.5, 1), (1.5, 3), (2.5, 4), (3.5, 3). Which value of K (in the range 1 to 6) is the best and why?

(c) **(3 points)** What is R^2 when K = 1 on the training data? Is this a good or bad model? Why?

3 Linear Regression

Assume that you are fitting a linear regression of the form $Y = \beta_0 + \beta_1 X$ to a data set of n points: $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, using a MSE loss: $\mathcal{L}(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_i])^2$. The closed form solution for β_0, β_1 can be derived analytically, i.e. by taking the derivative of the loss w.r.t $\beta_1, \beta 0$, and set it equal to 0.

(a) **(5 points)** Derive the derivative of the above loss function and find the closed-form solution for β_0, β_1.

(b) **(6 points)** Assume that you want to fit a line $Y = \beta_0$ to your training dataset. Using the argument in part (a), show that the optimal value for β_0 is equal to the mean of the y_i values if we use MSE, and is equal to the median of the y_i values if we use MAE.

4 Linear Regression: goodness of fit & Interpretation

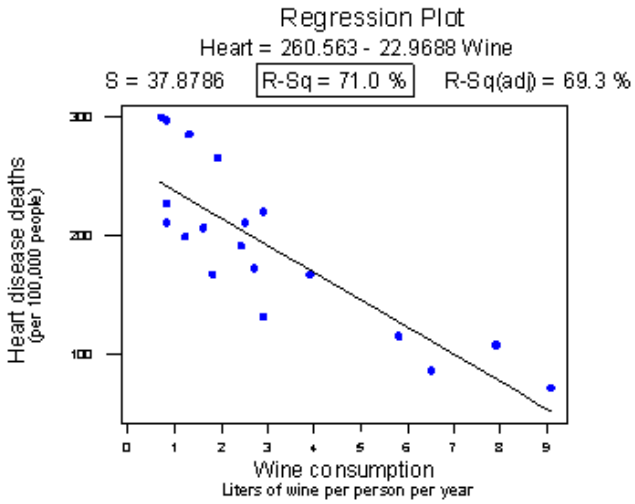
1- **(6 points)** US population was around 5 million in 1800, 23 million in 1850, 76 million in 1900, 161 million in 1950, and 291 million in 2000.

(a) use the analytic solution you derived in question 3(a) to fit a linear regression to the above data. What are β_0, β_1?

(b) What is R^2 for your model? Based on the value of R^2 can we say weather the estimated regression line fits the data well?

(c) Plot the residuals versus year. What do you conclude?

2- **(4 points)** The following plot shows how the number of deaths due to hearth disease varies with wine consumption, in different countries. Based on the values of R^2 and correlation reported on the top of the plot, can we conclude that drinking more wine reduces the risk of heart disease?



3- **(6 points)** *[You can use Python]* The Agriculture data contains data from 14 countries. The first column shows the amount of diseased crops (diseased) in million tons, that needs to be destroyed. the second column shows the amount of pesticide in tons (pesticide), and the third column shows the percentage of sunny days in a year (sun),

(a) Report β_0, β_1 for two *linear* classifiers that model: (i) diseased based on pesticide, and (ii) pesticide based on sun.

(b) Report R^2 for the above classifiers and explain the relationships between diseased, pesticide, and sun. Explain why it seems that the disease rate increases as more pesticide is used?

4-**(15 points)** *[You can use Python].* The Experiment dataset containing a thousand (x, y) data points, from a scientific experiment.

(a) Fit a linear model to the data and compute β_0, β_1.

(b) Compute and interpret the R^2 value. Does it show a strong linear relation between x and y?

(c) Conduct the test $H_0 : \beta_1 = 0$ (reject the null hypothesis if the p-value for β_1 is less than 0.05). What is your conclusion?

(d) Calculate a 95% confidence interval for β_1, using $\beta_1 \pm 2 \times SE(\beta_1)$, and interpret your interval. Suppose that if $\beta_1 \geq 1$, then we consider it to be meaningfully different from 0, in our research. Does the 95% confidence interval suggests that β_1 is meaningfully different from 0?

(e) Summarize the contradiction you've observed in parts (c) and (d). What is causing the contradiction, and what would you recommend we should always do while analyzing data?

5- **(10 points)** *[You can use Python]* The Volcano dataset contains 21 consecutive volcanic eruptions. Use a linear model to predict the time until the next eruption (next), given the duration of the last eruption (duration).

(a) Compute and interpret the R^2 value.

(b) If the duration of the last eruption was 3 minutes, obtain a 95% prediction interval for the time until the next eruption occurs, and interpret your prediction interval.

(c) Suppose you can only wait 60 minutes for the next eruption to occur. Can you make a decision based on the above prediction interval?

5 Interpretation of Coefficients in Linear Regression

(a) **[15 points]** Suppose that we want to model the effect of sunlight on the growth of three different types of plants. Suppose we are expecting a linear growth-response over a given range of sunlight, and hence we can model the outcome Y (amount of growth) as a linear function of the sunlight X_1 and the plant type X_2. How do you model the effect of sunlight on the growth of different plant types? How do you interpret each coefficients in your model?

(b) **[5 points]** How do you test the null hypothesis for each independent variable X_1 and X_2 to indicate if they have a significant correlation with the dependent variable Y?

6 Model Evaluation

(5 points) You have a dataset where the only y values are 0 or 1 (a binary classification problem). Out of the 100 data points in this dataset, there is a minority group of 5 data points with $y = 0$ and the rest of data points have $y = 1$. What is the issue if you randomly sample 80% of the data points as your training data? What evaluation strategy you should use to address this?