# 1 Data & Bias

(a) **(6 points)** A UCLA researcher wants to study how much sleep students get on average per night. To collect data, the researcher sends an online survey to a student health and wellness club and asks members to report their typical sleep duration. The researcher then concludes that UCLA students on average have good sleeping schedule. Does the researcher's data collection method exhibit any selection bias? Identify and explain each type of selection bias present in this study (Refer to Week 1 Lecture 2, slide 30 for a list).

Potential biases:

- Voluntary bias: Since the researcher sends an online survey, only students who are motivated to respond will participate. Those who are particularly concerned with health and wellness (and possibly have better sleep habits) are more likely to respond..
- Under-coverage bias: The researcher only surveys members of a student health and wellness club, who may not be representative of the general UCLA student population.
- Non-response bias: Some students may ignore the survey, particularly those who have poor sleep habits or are too busy to participate.
- Response bias: Students who are part of a health and wellness club might feel pressured to report healthier sleep habits than they actually have.

(b) **(6 points)** Since the early 2010s, banks and financial institutions have explored AI-driven systems to automate loan approvals. However, these AI systems were found to disproportionately deny loans to applicants from lower-income neighborhoods, leading to concerns about algorithmic bias and fairness. (1) Explain why the tool was discriminating against lower-income neighborhoods? (2) The developer decides to remove location-related data, such as ZIP codes, from the dataset. Would this effectively eliminate the bias? Why or why not?

(i) The training data for 'successful' loan approvals predominantly consisted of people from wealthy neighborhoods, as they were overrepresented in the bank's system before the implementation of this AI tool. As a result, the model learned to recognize features in applications that were more common among individuals from wealthy neighborhoods. (ii) Even if location-related data are removed, the AI can still infer an applicant's neighborhood based on other details in the application. This article describes it well, calling the issue "self-selection bias."

# 2 Linear Regression: goodness of fit & Interpretation

***Note:*** *The Python code for this Problem 2 can be found on the Bruinlearn.*

1- **(6 points)** US population was around 9 million in 1820, 40 million in 1870, 92 million in 1910, 151 million in 1950, and 281 million in 2000.

(a) The closed-form solution of linear regression with an MSE loss is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Use the formula to fit the above data. What will the population be like in 2020 under this model?

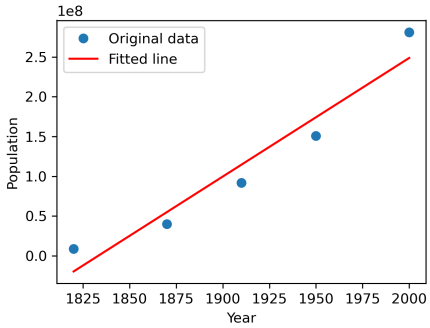
$\beta_0 = -2,732,678,350$ $\beta_1 = 1,490,722$

Prediction at the year of 2020 is: 278,579,381 or 278 million.

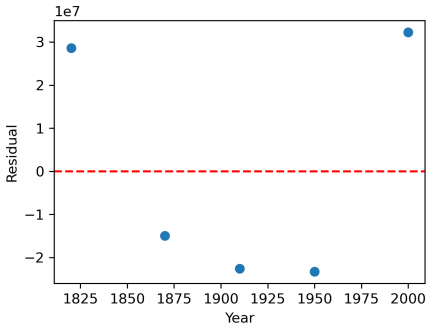
(b) What is $R^2$ for your model? Based on the value of $R^2$ can we say weather the estimated regression line fits the data well?

The value of $R^2$ is 0.9323. That is, only 6.77% of the variation in the U.S. population is left to explain after taking into account the year in a linear way. However, from the $R^2$ value alone, we cannot conclude if the regression line fits the data well or not.

(c) Plot the residuals versus year. Do you think this is a good model? Why



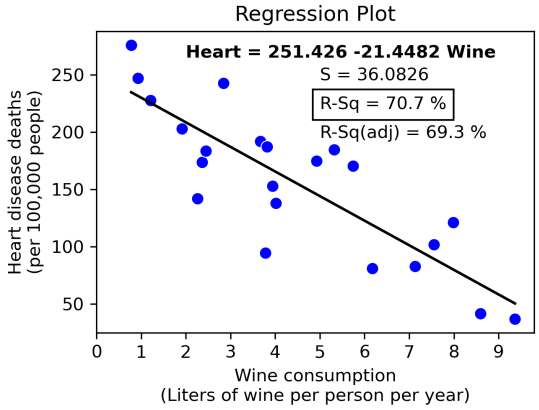

(a) Fitting results for Q2-1




(b) Residual plot for Q2-1(c)

This is not a good model. From looking at the plot of residuals versus year, we can see that the residuals are not randomly distributed about 0, which indicates that a linear model is not expressive enough for this data. A linear model would be inappropriate to use - even though we have a high $R^2$ value, it doesn't mean that a linear model is a good fit for the data.

2- **(4 points)** The following plot shows how the number of deaths due to heart disease varies with wine consumption, in different countries. Is there a strong correlation between heart disease and wine consumption? Can we conclude that drinking more wine will reduce the risk of heart disease? Explain your reasoning.



From the figure, we can see that $R^2 = 0.707$ and $R = -\sqrt{R^2} = -0.841$. Thus, we can conclude that drinking more wine is strongly correlated with reduced heart disease, but we cannot necessarily say that drinking more wine *reduces* the risk of heart disease. Correlation is not equal to causation. Indeed, there may be other differences in the behavior of the people in the various countries that really explain the differences in the heart disease death rates, such as diet, exercise level, stress level, social support structure, and so on.

3- **(6 points)** *[You can use Python]* The Business Data contains data from 14 companies. The first column shows the investment per year (Investment), the second column shows the business growth per year (Business Growth), and the third column shows the profits per year (Profit).

(a) Report $\beta_0, \beta_1$ for two *linear* classifiers that model: (i) Business Growth based on Investment, and (ii) Investment based on Profit.

Business Growth vs. Investment: $\beta_0 = 1.10$, $\beta_1 = 0.62$

Investment vs. Profit: $\beta_0 = 40.0$, $\beta_1 = 7.62$

(b) Report $R^2$ for the above classifiers and explain the relationships between investment, business growth, and profit. Analyze the potential reason behind this.

Business Growth vs. Investment: $R^2 = 0.589$. This shows a moderate positive association between the variables, possibly suggesting that as investment increases, so too does business growth.

Investment vs. Profit: $R^2 = 0.605$. This shows a moderate positive association between the variables, possibly suggesting that as profit increases, investment increases.

If the profits are high, the investors are more likely to increase their investments, and a high investment tend to be associated with higher values of business growth, as the investment helps the company to expand. Reasonable answers would suffice.

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 $\beta_0, \beta_1$.

$\beta_0 = 5.5411$, $\beta_1 = 0.0975$

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

$R^2 = 0.201$, indicating that 20.1% of the variation in $x$ is explained by $y$. This shows a weak linear relationship between the 2 variables.

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

Running the dataset through statsmodels.OLS and printing out the summary, we get that $\beta_1$ has a standard error of 0.006 and a t value of 15.836. The value $P(t > 15.836)$ for a t-distribution with $n - 2 = 1000 - 2 = 998$ degrees of freedom is equal to 0, which is less than the p-value threshold of 0.05, so we reject the null hypothesis. We conclude that there is a significant relationship between $x$ and $y$.

(d) Calculate a 95% confidence interval for $\beta_1$, 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 suggest that $\beta_1$ is meaningfully different from 0?

95% CI $= [0.0975 - 2 \times 0.006, 0.0975 + 2 \times 0.006] = [0.0855, 0.1095]$. From this, we see that the 95% CI does not suggest that $\beta_1$ 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?

In (c), we concluded that there is a meaningful relationship between the response variable and the explanatory variable, but in (d) we observed no such meaningful relationship. The contradiction is caused by different evaluation criteria. In (c), we set a p-value of 0.05 to determine significance, while in (d), we picked a value to determine the cutoff for if $x$ is significant. Thus, when analyzing data, we should always establish clear evaluation criteria first.

5- **(10 points)** *[You can use Python]* The Earthquake dataset contains 21 consecutive earthquake events. Use a linear model to predict the time until the next aftershock (next), given the duration time of current aftershock (duration).

(a) Is the linear model a good model? Analyze your result using $R^2$.

Based on $R^2$ alone, the linear model is a good fit here. $R^2 = 0.748$, indicating that 74.8% of the variation in the time until the next eruption is explained by eruption duration.

Students may make comments about judging from $R^2$ itself is not sufficient to judge if a model is good or not, which is a fair point. As long as the $R^2$ calculation is correct and there is some explanation on its meaning, the student can get the full score.

(b) If the duration of the last eruption was 5 minutes, obtain a 95% prediction interval for the time until the next eruption occurs, and interpret your prediction interval.

Using get_prediction(input).summary_frame(alpha=0.05), where input is 5, we get a 95% confidence interval of [68.10739, 95.811722] minutes. This means that we are 95% confident that the true time until the next eruption will be between 68.11 and 95.81 minutes. Refer to this link for more details about the implementation and this link for the principle.

Alternatively, it is also fine to use $\mu \pm 2\sigma_\epsilon$ to approximate the 95% prediction interval, where $\mu$ is the average prediction when input is 5 minutes, and $\sigma_\epsilon = \sqrt{\frac{n \cdot MSE}{n-2}}$ represents the standard deviation of the residuals (error term) in the regression model. In this way, the approximate 95% prediction interval is [69.81171, 94.10740].

(c) If you need to leave in 50 minutes, can you determine if you can see the eruption based on the data? Explain your reasoning.

Based on the result in (b), we are at least 95% confident that we will not see the eruption within 50 minutes.

## 3 Interpretation of Coefficients in Linear Regression

Suppose that we want to model the sales of fish at a market. We will consider two datasets individually. Both datasets share the same structure, containing columns for *sales*, *weight*, and *fish type*.

- **Dataset A** includes the following fish types: *Tuna*, *Swordfish*, and *Blobfish*.
- **Dataset B** exclusively has data for different grades of Salmon: *Canned-Salmon*, *Commercial-Grade Salmon*, and *Sashimi-Grade Salmon*.

Across both datasets, we expect a linear growth-response of sales with respect to weight over a given range. Hence, we want to model the outcome $Y$ (sales) as a linear function of the weight $X_1$ and the fish species $X_2$.

(a) **(5 points)** As fish type is a categorical feature, we need to first convert it through encoding. When processing each dataset individually, which of the following encodings will be more preferable? Explain your reasoning for both Datasets A and B.

(1) Create one variable $X_2 = \{1, 2, 3\}$. Specifically, for Dataset A, assign $X_2 = 1$ for Tuna, $X_2 = 2$ for Swordfish, and $X_2 = 3$ for Blobfish. For Dataset B, assign $X_2 = 1$ for Canned-Salmon, $X_2 = 2$ for Commercial-Grade Salmon, and $X_2 = 3$ for Sashimi-Grade Salmon.

(2) Create three indicator variables to represent each fish type. For Dataset A, these variables would be: $X_2^{\text{Tuna}}$, $X_2^{\text{Swordfish}}$, and $X_2^{\text{Blobfish}}$, where each variable is set to 1 if the fish is of that type and 0 otherwise. Similarly, for Dataset B, define $X_2^{\text{Canned}}$, $X_2^{\text{Commercial}}$, and $X_2^{\text{Sashimi}}$ using the same encoding.

(Dataset A: 2) In Dataset A, a one-hot encoding will be more preferable. Since the categorical feature "fish species" does not have an ordinal relationship in Dataset A, it will be better to use one-hot encoding and will be easier to interpret.

(Dataset B: 1 or 2) In Dataset B, both encoding 1 and 2 are acceptable. The "fish species" feature has an ordinal relationship in quality between canned, and commercial grade and sashimi grade. Answers that use encoding 1 must indicate the existence of an ordinal relationship in Dataset B.

(b) **(5 points)** For both Dataset A and Dataset B, based on the encoding you chose how do you model the weight of the fish on the sales of different fish species? **Hint.** Use $\beta_0$, $\beta_1$, ... to denote the coefficients and write the model in the form of $Y = \beta X + \ldots + \epsilon$.
**[Note:]** This model that uses p variables will be marked as correct only for this homework:
Dataset A, Best Answer:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{Tuna} + \beta_3 X_2^{Swordfish} + \beta_4 X_1 X_2^{Tuna} + \beta_5 X_1 X_2^{Swordfish} + \epsilon$$

Also acceptable for Dataset A, only for this homework (using more coefficients and harder to

interpret):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{Tuna} + \beta_3 X_2^{Swordfish} + \beta_4 X_2^{Blobfish} + \beta_5 X_1 X_2^{Tuna} + \beta_6 X_1 X_2^{Swordfish} + \beta_7 X_1 X_2^{Blobfish} + \epsilon$$

**Dataset B:**

One hot encoding follows the same form as Dataset A. With an ordinal encoding it is of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \epsilon$$

Coefficients can take any notations, no need to strictly be as written. For example, $\beta_2^{Tuna}$ is also correct.

(c) **(10 points)** How do you interpret each coefficients in your model? Your answer should include interaction terms.

**[Important Note:]** The following mode that uses $p-1$ binary variable for $p=3$ categories is the right choice for interpretation. **Use this version in the exams and future homeworks**:
When the fish species is Blobfish: $Y = \beta_0 + \beta_1 X_1 + \epsilon$
When the fish species is Swordfish: $Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 + \epsilon$
When the fish species is Tuna: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 + \epsilon$

- $\beta_0$ represents the average base sale for Blobfish when weight is 0.
- $\beta_1$ represents the expected change in sales for Blobfish per unit increase in weight.
- $\beta_2$ Represents the average difference in base sales between Tuna and Blobfish when weight is 0.
- $\beta_3$ Represents the average difference in base sales between Swordfish and Blobfish when weight is 0.
- $\beta_4$ Represents the average additional change in sales for Tuna per unit increase in weight, compared to Blobfish.
- $\beta_5$ Represents the average additional change in sales for Swordfish per unit increase in weight, compared to Blobfish.

**[Important Note:]** This model that uses $p=3$ binary variables will be marked as correct for this homework. Although this model will have a satisfactory performance, it is not the right choice for interpretation (since we have two extra coefficients $\beta_4, \beta_7$ that can be inferred from the rest). For future exams and homework, use the model form and interpretation shown above:
When the fish species is Tuna,

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) X_1 + \epsilon$$

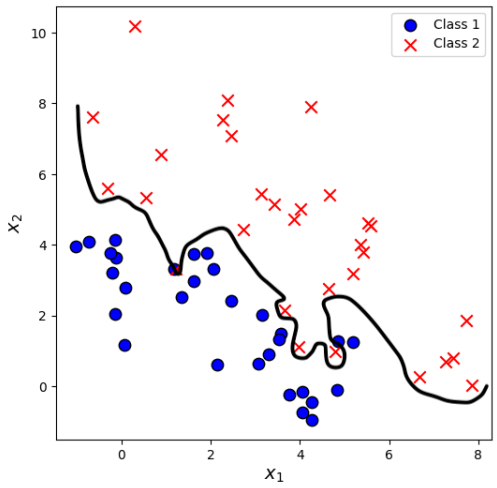
- $\beta_0$ represents the average base sale for all fish (when weight is 0).
- $\beta_1$ represents the expected change of growth of sale for all fish per unit change of the amount of weight.
- $\beta_2$ represents the average extra sales for Tuna compared to the average base sale for all fish for weight 0.
- $\beta_3$ represents the average extra sales for Swordfish compared to the average base sale for all fish for weight 0.
- $\beta_4$ represents the average extra sales for Blobfish compared to the average base sale for all fish for weight 0.
- $\beta_5$ represents the average expected additional change of the amount of growth of sales per unit change of the amount of weight for Tuna.
- $\beta_6$ represents the average expected additional change of the amount of growth of sales per unit change of the amount of weight for Swordfish.
- $\beta_7$ represents the average expected additional change of the amount of growth of sales per unit change of the amount of weight for Blobfish.

## 4 Bias, Variance and Regularization

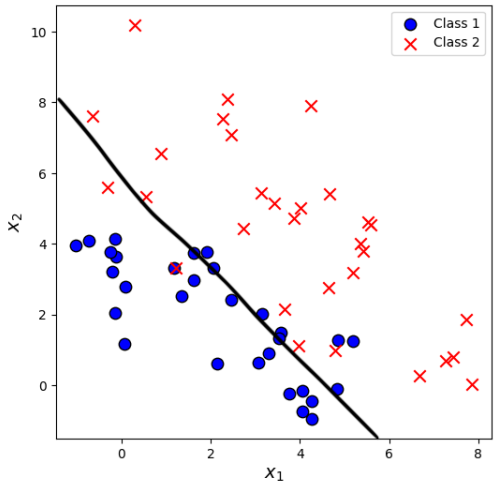
(a) **(2 points)**

The figure below shows a labeled dataset. Could you draw a decision boundary where the classifier is overfitting? How will the training and test errors change over time? What are some potential reasons for overfitting to occur? How do you think the model will perform on the test set?

Solution: Figure (a) training error will converge while test error will not. the test error will be larger than the training error. (b) there are many potential reasons, such as: 1. the model is too complex; the training dataset is too small; lack of proper regularization; imbalanced dataset; overtraining, etc. (c) the model will perform poorly.



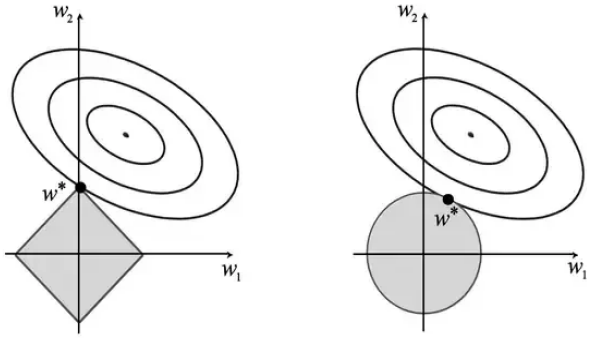
(b) **(2 points)** Now, could you draw a decision boundary where the classifier is underfitting? How will the training and test errors change over time? What are some potential reasons for underfitting to occur? How do you think the model will perform on the test set?




Solution: Figure (a) Both the training and test errors would not converge, and they will be high at the end of the training. (b) there are many potential reasons, such as: 1. the model is too simple; not enough training epochs; the data does not have good features, etc. (c) the model will perform poorly.

(c) **(5 points)** One strategy to reduce variance and improve generalization is regularization. In figure below, the contour lines represent the loss function. Could you explain (1) where is the optimal solution, with or without regularization? (2) which one is with L1 regularization, and which one is with L2 regularization? Why? (3) Under what conditions would you use these regularizations?

Solution: (1) Without regularization: The optimal solution is at the center of the contour lines, which represents the point where the loss function is minimized. With regularization: The optimal solution is shifted to the point where the contour lines intersect with the constraint imposed by the regularization (the shaded region). This encourages a simpler model by penalizing large weights. (2) The left figure represents L1 regularization (Lasso) because the shaded area is a diamond shape, meaning it encourages sparsity by setting some parameters to exactly zero. The right figure represents L2 regularization (Ridge) because the shaded area is a circular shape, meaning it uniformly shrinks all parameters rather than setting them to zero. (3) I would use L1 regularization when feature selection is needed, as it forces some weights to be exactly zero, effectively removing less important features. I would use L2 regularization in almost all other scenarios.




## 5 Logistic Regression

Suppose we fit a multiple logistic regression: $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

(a) **(2 points)** Suppose we have $p = 3$, and $\beta_0 = 2, \beta_1 = 5, \beta_2 = -2, \beta_3 = -3$. When $X_1 = 4, X_2 = 8, X_3 = 2$, what are the odds and probability of the event that $Y = 1$? Solution: The odds are 1 and the probability is 0.5 .

(b) **(2 points)** Suppose we increase the $X_1$ value by 2, how does it change the log odds and odds of the event that Y = 1? What if instead, we decrease the $X_2$ value by 1? Solution: One unit increase in $X_1$ increases log odds by 5 , so two units will increase the log odds by 10, which in turn causes odds to be multiplied by $e^{10}$ (odds increase). Similarly, when $X_2$ is decreased by one unit, log odds are increased by 2 and odds are multiplied by $e^2$.

(c) Explain how increasing or decreasing $\beta_0, \beta_1$ or $\beta_2$ affect our predictions. Solution: Increasing $\beta_0$ will increase our odds and therefore increase our predicted probability (similarly, decreasing it will decrease odds/probability). If $\beta_1$ is increased, our predicted odds/probability will increase if the feature is positive (opposite if the feature is negative). If $\beta_2$ is increased, our predicted odds/probability will decrease if the feature is positive (opposite if the feature is negative). If $\beta_3$ is increased, our predicted odds/probability will decrease if the feature is positive (opposite if the feature is negative). Answer could alternatively describe how the probability curve is shifted when $\beta_0$ changes and how the curve becomes steeper or less steep based on the other $\beta$ values.

(d) What is the formulation of the decision boundary? Which points are on the decision boundary? Solution: Our decision boundary can be found by setting $P(Y = 1) = 0.5$, so $log(\frac{P(Y=1)}{1-P(Y=1)}) = 0$. Our boundary then becomes $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0$

(e) Suppose we fit another two logistic regression models: one with only $X_1$ and the other one with only $X_2$, and we observe that the coefficients of $X_1$ and $X_2$ in the two models are different than those specified in part (a). Explain what is the potential reason and why it could be problematic that the coefficients are different than those specified in part (a). Solution: There is multicollinearity between these features. This makes it problematic to interpret the coefficients, as the coefficients may not be unique. Also, multicollinearity affects our confidence intervals for these features.