Homework 1

● Graded

Student

TIYA CHOKHANI

Total Points

78.5 / 92 pts

Question 1

Data & Bias **12** / 12 pts

1.1 (a) **6** / 6 pts

✔ **– 0 pts** Correct

– 3 pts Only one bias is provided with full explanation, or only the names of the bias are provided without any explanation.

– 3 pts Explanation of bias is provided, but specific kinds of bias (voluntary, response, etc.) are not given.

– 5 pts Only one bias is provided without any explanation.

– 6 pts No answers

1.2 (b) **6** / 6 pts

✔ **– 0 pts** Correct

– 3 pts Part 1 (why biased against low-income neighborhoods)

– 3 pts Part 2 (why bias against low-income is not eliminated; should mention how other non-location features could still be used by the AI to infer location)

– 6 pts No answer

Question 2

Linear Regression **34** / 41 pts

2.1 **Population** **5** / 6 pts

– **0 pts** Correct

– **1 pt** (a): β_0 or β_1 is wrong

– **0.5 pts** (a): Calculation error only for β_0 or β_1 (correct formula)

– **0.5 pts** (a) No population estimate

– **0.5 pts** (b): wrong R^2 value

– **0.5 pts** (b): does not mention that R^2 is not enough to judge if the estimated regression line fits

✔ – **1 pt** (b) No explanation

– **1 pt** (c): wrong conclusion or no explanation over conclusion

– **0.5 pts** (c) correct conclusion but wrong explanation

– **1 pt** (c): no plot or plot is wrong

– **6 pts** No answer

– **3 pts** No calculations shown

2.2 **Wine Consumption** **4** / 4 pts

✔ – **0 pts** Correct

– **0 pts** claim without explanation in correlation between heart disease and wine consumption.

– **2 pts** wrong/no conclusion in correlation between heart disease and wine consumption.

– **1 pt** claim without explanation in concluding relation (need at least one reason) / incorrect reasoning

– **2 pts** wrong conclusion concluding relation/no conclusion regarding causation

– **4 pts** No answer

2.3 **Business** **5** / 6 pts

– **0 pts** Correct

– **1 pt** (a): wrong β_1 or β_0 value for one of the models: each worth 0.5 point.

– **2 pts** (a): wrong β_1 or β_0 value for both models: each worth 0.5 point.

– **1 pt** (b): one wrong R^2 value

– **2 pts** (b): both wrong R^2 values

✔ – **1 pt** (b): incorrect / incomplete explanation. An explanation is considered complete if it mentions **either** of the following: (1) **both** the relationships between business growth and investment, and between investment and profit, **or** (2) the relationship between all three variables. Must also analyze potential reason.

– **2 pts** (b): no explanation

– **6 pts** No answer

2.4 **Experiment** **12** / 15 pts

– **0 pts** Correct

– **2 pts** (a): wrong β_1 or β_0 value (each 1 point)

– **1 pt** (a): wrong β_1 value

– **3 pts** (b): computation (1 point), Interpretation (1 point), weak relation (1 point)

– **1 pt** (b): wrong value of R^2 / R / β_1

– **3 pts** (c): correct conclusion with appropriate wordings (reject the null hypothesis, significant relation, etc.)

– **1 pt** (c): missing 1 component (rejecting null, or significant relation)

– **2 pts** (d): incorrect interpretation (1 point), meaningfully different from 0 (1 point)

– **4 pts** (e) spot the contradiction (1 point), correctly explain the reason (2 point), reasonable suggestion (1 point)

✔ – **2 pts** (e) incorrect explanation of the reason behind contradiction

✔ – **1 pt** (e) reasonable suggestion to mitigate this issue

– **15 pts** No answer

– **1 pt** The confidence interval is 2*standard error

– **1 pt** D) Confidence interval

– **1 pt** Wrong interpretation of R = 0.2

– **0.5 pts** Noted data was noisy, but some confusion over strength of relation between X and Y and improvement of model over guessing mean.

2.5 Earthquake 8 / 10 pts

- **– 0 pts** Correct
- **– 1 pt** Wrong R^2 value or no analysis
- **– 2 pts** Wrong Interval
- **– 2 pts** (a): Correct R^2 value (1 point), explain the meaning of R^2 (1 point). If the student argues that R^2 alone is not enough to make the conclusion, they should also get full score.
- **– 4 pts** (b): correct interval (2 point), interpretation (2 point)
- **– 4 pts** (c) correct conclusion (2 point), explanation no the decision (2 point)
- **– 10 pts** No answer
- ✔ **– 2 pts** (b) no explanation of Interval
- **– 2 pts** (c) no conclusion or wrong conclusion
- **– 2 pts** (c) no explanation
- **– 2 pts** (b) no interpretation
- **– 0.5 pts** Interval Half Right

Question 3

Interpretation of Coefficients in Linear Regression **15** / 20 pts

3.1 **(a)** **0** / 5 pts

– **0 pts** Correct choice and reasoning that the feature does not follow ordinal relationships.

– **2.5 pts** Incorrect choice but provided detailed reasoning that partly makes sense.

✔ – **5 pts** Incorrect answer of left blank.

– **1.5 pts** No explanation

3.2 **(b)** **5** / 5 pts

✔ – **0 pts** Correct. Also correct if not using X_2^C (or one of the indicator variable).

– **0.5 pts** Missing one term but otherwise correct. For example, missing the intercept term or the term for X_1.

– **1 pt** Missing two terms.

– **2 pts** Missing several terms but the listed terms are correct. For example, missing all interaction terms or missing all terms related to X_1.

– **3 pts** Missing several terms and the listed terms are only partly correct.

– **5 pts** Incorrect or left blank.

3.3 (c) **10** / 10 pts

✔ **– 0 pts** Correct. If the interpretations are correct/sensible, full credit can be awarded.

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Partly incorrect interpretation on the intercept terms. (β_0)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1.5 pts Missing/incorrect interpretation on the intercept terms. (β_0)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 0.5 pts Partly incorrect interpretation on the expected change of sales per unit of the fish's weight. ($\beta_1 X_1$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Missing/incorrect interpretation on the expected change of sales per unit of the fish's weight. ($\beta_1 X_1$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Partly incorrect interpretation on the base sale for fish type A term. ($\beta_2 X_2^A$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 3 pts Missing/incorrect interpretation on the base sale for fish type A term. ($\beta_2 X_2^A$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 2 pts Partly incorrect interpretation on the interaction terms (e.g. the expected change of sales per unit of the fish's weight with respect to each fish specie: $X_1 X_2^A$).

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 3 pts Missing/incorrect interpretation on the interaction terms (e.g. the expected change of sales per unit of the fish's weight with respect to each fish specie: $X_1 X_2^A$).

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 3 pts Missing details of explanation (e.g. only stating β_1 is the coefficient for variable X_1, etc)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 3 pts Explanation should contain phrases like "compared to type C (or whichever specie that is missing in the equation"

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 10 pts Question unanswered.

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Missing one interaction term

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 5 pts Missing explanation of what each β_i represents

– **2 pts** Missing explanation of β_i for interaction terms or for individual β_i terms

– **0.5 pts** Missing explanation of 1 β_i

Question 4

Bias, Variance and Regularization **8.5** / 9 pts

4.1 **(a)** **1.5** / 2 pts

– **0 pts** Correct

✔ – **0.5 pts** Plotted decision boundary is not overfitting.

– **0.5 pts** If did not mention that training error will converge while the test error will not.

– **0.5 pts** If did not mention that the model will not perform well.

– **0.5 pts** If no or wrong reasons for overfitting are listed.

4.2 **(b)** **2** / 2 pts

✔ – **0 pts** Correct.

– **0.5 pts** Plotted decision boundary is not underfitting.

– **0.5 pts** If did not mention that training error and test error both will not converge/high at the end of the training.

– **0.5 pts** If no or wrong reasons for underfitting are listed.

– **0.5 pts** If did not mention that the model will not perform well.

4.3 **(c)** **5** / 5 pts

✔ – **0 pts** Correct

– **0.5 pts** (1) The optimal solution for unregularized model is incorrect/unspecified (center of ellipses)

– **0.5 pts** (1) The optimal solution for regularized model is incorrect/unspecified (w* / intersection)

– **1 pt** (2) The L1 & L2 loss are not correctly identified

– **0.5 pts** (2) The reasons for identifying the L1 & L2 loss are missing/wrong.

– **1 pt** (3) Wrong/missing reasons for using L1/L2 loss

– **5 pts** Missing

Question 5

Logistic Regression **9** / 10 pts

5.1 **(a)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt wrong odds/odds value not mentioned
Correct answer: 1.

– 1 pt wrong probability

5.2 **(b)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt wrong X1 interpretation

– 1 pt Wrong x2 interpretation

– 1 pt Partially correct/Numerical value not mentioned

5.3 **(c)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt Wrong interpretation on β_0

– 1 pt Wrong interpretation on β_1

– 1 pt Wrong interpretation on β_2

– 1 pt Wrong interpretation either for odds/probability

5.4 **(d)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt Wrong equation of boundary/missing equation

– 2 pts Incorrect/Missing answer

5.5 **(e)** **1** / 2 pts

– 0 pts Correct

✔ **– 1 pt** Did not mention that coefficients are not unique when with or without multicollinearity

– 1 pt Unreasonable reasons for potential difficulties.

Questions assigned to the following page: 1.1 and 1.2

Introduction to Data Science (CS M148) Submit to Gradescope
Released: January 31, 2025 **Homework 1** Due: February 10, 2pm

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 (Refter to Week 1 Lecture 2, slide 30 for a list).

(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?

1.a) Yes, this exhibits undercoverage bias as only students who are part of this club are sampled & they're more likely to be taking care of the health & wellness. Since the survey is sent online there would also be voluntary bias where only student who really care about the topic participate. Additonally students may not respond giving way to non response bias.

b) AI systems learn & are trained on historical data so if the past lending actions were discriminatory the model would perpetuate these biases leading to more data being generated against low income grops which is then further fed into the model.

- Even if zip code data was removed other variables may still be highly correlated to geographical areas which could serve as proxies. The model could still infer the applicants location so it wouldn't help. If the training label are biased the algo will continue to learn these unfair patterns

No questions assigned to the following page.

2 Linear Regression: goodness of fit & Interpretation

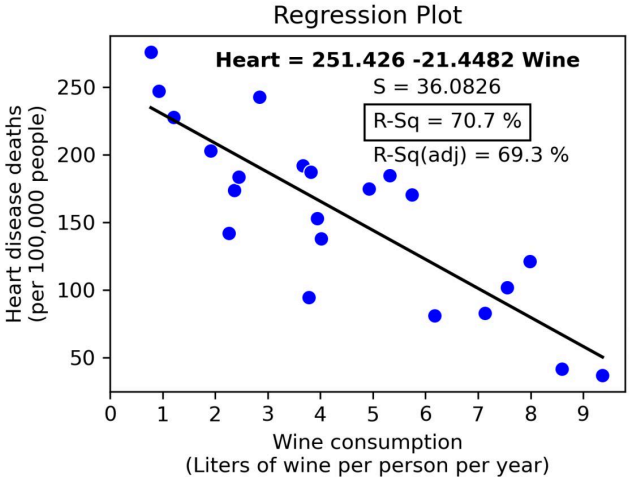
1- **(6 points)** US population was around 9 million in 1820, 40 million in 1870, 92 million in 1910, 151 million in 1950, and 281 million in 2000.

(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?

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

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

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.



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 β_0, β_1 for two *linear* classifiers that model: (i) Business Growth based on Investment, and (ii) Investment based on Profit.

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

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

No questions assigned to the following page.

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

(b) Is there a strong linear relation between x and y? Explain your reasoning.

(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). Analyze your result

(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 suggest 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 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.

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

(c) If you just experienced an aftershock lasting for 5 minutes, and you will leave the earthquake area in 50 minutes, can you determine if you will experience the next aftershock based on the data? Explain your reasoning.

Question assigned to the following page: 2.1

Part 2: Linear Regression

Q1.

```
import numpy as np
import matplotlib.pyplot as plt


years = np.array([1820, 1870, 1910, 1950, 2000])
populations = np.array([9, 40, 92, 151, 281])  # population in millions


x_bar = np.mean(years)
y_bar = np.mean(populations)

beta1 = np.sum((years - x_bar) * (populations - y_bar)) / np.sum((years - x_bar)**2)

beta0 = y_bar - beta1 * x_bar

print("Estimated slope (beta1):", beta1)
print("Estimated intercept (beta0):", beta0)
```

```
Estimated slope (beta1): 1.4907216494845361
Estimated intercept (beta0): -2732.678350515464
```

```
#predicting for 2020
year_2020 = 2020
pop_pred_2020 = beta0 + beta1 * year_2020
print("\nPredicted US population in 2020 (in millions):", pop_pred_2020)
```

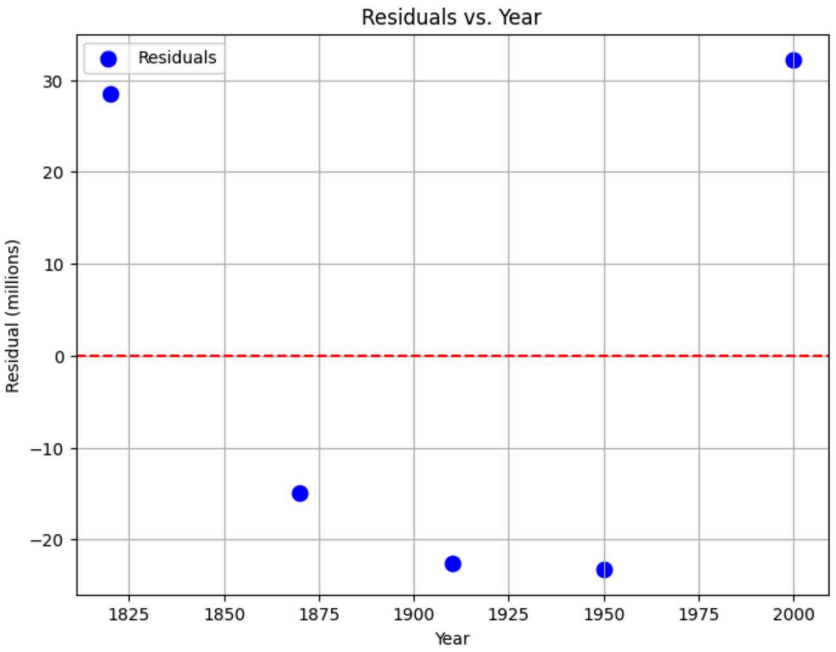
```
Predicted US population in 2020 (in millions): 278.5793814432991
```

```
pop_pred = beta0 + beta1 * years

# Compute the residual sum of squares (SS_res) and the total sum of squares (SS_tot)
SS_res = np.sum((populations - pop_pred) ** 2)
SS_tot = np.sum((populations - y_bar) ** 2)

R2 = 1 - SS_res / SS_tot
print("\nCoefficient of determination (R^2):", R2)
```

```
Coefficient of determination (R^2): 0.9323216115302541
```

```
residuals = populations - pop_pred

plt.figure(figsize=(8, 6))
plt.scatter(years, residuals, color='blue', s=80, label='Residuals')
plt.axhline(0, color='red', linestyle='--')
plt.xlabel("Year")
plt.ylabel("Residual (millions)")
plt.title("Residuals vs. Year")
plt.legend()
plt.grid(True)
plt.show()
```

Questions assigned to the following page: 2.1, 2.2, and 2.3



The R^2 value is around 0.93, which might suggest that the linear model explains a high proportion of the variance in the given data points. However, note that there are only 5 data points and the growth in population over time appears to be not strictly linear. In the plot of the residuals versus year the data points arent randomly scattered around 0 so this may show systematic patterns, indicating that the linear model does not capture the true relationship well and we should explore other models. Realistically, an exponential model that for population growth might be more appropriate.

Q2.

The regression plot indicates a negative correlation between wine consumption and heart disease deaths. The R-squared value is 70.7%, meaning that approximately 70.7% of the variation in heart disease deaths can be explained by wine consumption. This suggests a relatively strong correlation.

However, correlation does not imply causation. While the data shows an inverse relationship, it does not prove that drinking more wine directly reduces heart disease risk. Other confounding factors could influence heart disease rates.

Thus, while the data suggests a strong correlation, it would be incorrect to conclude that drinking more wine causes a reduction in heart disease risk without further research, including controlled experiments or longitudinal studies.

Q3.

```
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
import pandas as pd


data = {
    'Investment':      [106, 94, 35, 44, 42, 64, 65, 66, 69, 105, 109, 115, 119, 137],
    'Business Growth': [27, 41, 50, 15, 20, 28, 43, 50, 51, 75, 83, 80, 85, 95],
    'Profit':          [8, 9, 2, 3, 1, 3, 5, 3, 5, 7, 10, 12, 4, 8]
}

df = pd.DataFrame(data)
df
```

Question assigned to the following page: 2.3

	Investment	Business Growth	Profit
0	106	27	8
1	94	41	9
2	35	50	2
3	44	15	3
4	42	20	1
5	64	28	3
6	65	43	5
7	66	50	3
8	69	51	5
9	105	75	7
10	109	83	10
11	115	80	12
12	119	85	4
13	137	95	8

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
X1 = df[['Investment']]  # Predictor: Investment (as a 2D array)
y1 = df['Business Growth']  # Response: Business Growth

model1 = LinearRegression()
model1.fit(X1, y1)

beta0_model1 = model1.intercept_
beta1_model1 = model1.coef_[0]
R2_model1    = model1.score(X1, y1)

print("\nModel 1: Business Growth = β₀ + β₁ * Investment")
print("Estimated β₀ (intercept):", beta0_model1)
print("Estimated β₁ (slope):", beta1_model1)
print("R² for Model 1:", R2_model1)
```

```
Model 1: Business Growth = β₀ + β₁ * Investment
Estimated β₀ (intercept): 1.1010393451500633
Estimated β₁ (slope): 0.6218679052717087
R² for Model 1: 0.5893613869599095
```

```
X2 = df[['Profit']]  # Predictor: Profit (as a 2D array)
y2 = df['Investment']  # Response: Investment

model2 = LinearRegression()
model2.fit(X2, y2)

beta0_model2 = model2.intercept_
beta1_model2 = model2.coef_[0]
R2_model2    = model2.score(X2, y2)

print("\nModel 2: Investment = β₀ + β₁ * Profit")
print("Estimated β₀ (intercept):", beta0_model2)
print("Estimated β₁ (slope):", beta1_model2)
print("R² for Model 2:", R2_model2)
```

```
Model 2: Investment = β₀ + β₁ * Profit
Estimated β₀ (intercept): 40.00000000000002
Estimated β₁ (slope): 7.624999999999996
R² for Model 2: 0.6054927516610777
```

Model 1 (Business Growth vs. Investment): $R^2 \approx 0.5894$ This indicates that approximately 58.94% of the variation in Business Growth is explained by the variation in Investment. This is a moderate level of explanatory power.

The positive slope (0.622) shows that as Investment increases, Business Growth tends to increase. For every unit increase in Investment, Business Growth is predicted to increase by about 0.622 units.

Model 2 (Investment vs. Profit): $R^2 \approx 0.6055$ This means about 60.55% of the variation in Investment is explained by the variation in Profit.

The positive slope (7.625) indicates that higher Profit is associated with higher Investment. For every unit increase in Profit, Investment is predicted to increase by approximately 7.625 units.

Question assigned to the following page: 2.4

The moderately high R^2 for both the models suggests that Investment is a strong, but not exclusive, determinant of Business Growth and the same for Profit and Investment. However, the remaining approx 40% of the variability is likely due to other factors not captured by Investment or Profit alone. We should probably do some further analysis into these relationships and see if multiple variable regression might be better.

```
from io import StringIO

data = """0,6.397809
0,5.532191
0,5.123259
0,6.319232
0,7.136257
0,7.177463
0,3.704488
0,5.221000
0,7.075686
0,8.261649
0,4.351055
0,5.145795
0,6.980260
0,4.040791
0,4.275448
0,5.168821
0,6.383496
0,6.989982
0,5.913795
0,5.916315
0,7.696846
0,5.895029
0,5.043379
0,4.657653
0,5.987102
0,5.727850
0,7.600570
0,6.186357
0,6.125445
0,4.811354
0,5.610431
0,5.259359
0,5.887925
0,5.223274
0,3.258889
0,5.875271
0,7.214710
0,5.633613
0,6.119648
0,6.023233
0,5.926017
0,8.454964
0,7.295902
0,3.928739
0,3.968419
0,5.527882
0,3.928266
0,5.337388
0,3.593133
0,6.523350
1,5.437469
1,6.392609
1,6.034174
1,7.421561
1,4.303619
1,7.948469
1,5.155663
1,4.546045
1,7.205488
1,7.704003
1,7.869729
1,5.806137
1,3.341998
1,6.979537
1,4.220869
1,5.541374
1,4.596337
1,6.771773
1,4.925673
1,4.240992
1,5.244899
1,5.984139
1,5.828767
1,5.264250
1,4.380315
```

No questions assigned to the following page.

```
1,6.535146
1,7.763487
1,5.679092
1,4.836219
1,6.060370
1,3.444512
1,5.860655
1,7.662640
1,4.949771
1,6.109094
1,4.525310
1,5.857258
1,3.888716
1,3.645613
1,5.634810
1,3.754914
1,6.291489
1,7.848660
1,2.200337
1,5.764076
1,7.543617
1,5.405727
1,6.078614
1,6.418119
1,6.079127
2,6.052743
2,6.631594
2,5.866366
2,7.034462
2,3.995913
2,6.207806
2,4.206560
2,3.407804
2,6.301974
2,5.379103
2,5.252079
2,5.610732
2,4.963746
2,6.332934
2,6.825209
2,6.348738
2,7.124977
2,6.142856
2,5.011918
2,5.575881
2,5.000899
2,5.258590
2,4.864159
2,7.404670
2,6.419191
2,5.139899
2,5.916134
2,5.276078
2,8.976365
2,4.775443
2,6.754184
2,5.811295
2,5.477042
2,4.621183
2,4.959969
2,4.789153
2,5.195110
2,4.975681
2,6.048538
2,7.067669
2,5.420221
2,4.983800
2,5.220958
2,3.869172
2,4.610222
2,5.836233
2,5.830253
2,6.508360
2,4.793240
2,7.260913
3,6.595780
3,5.616119
3,6.719450
3,8.018110
3,7.051658
3,5.620945
3,5.095760
```

No questions assigned to the following page.

```
3,5.974558
3,4.495460
3,3.147199
3,6.035785
3,6.343267
3,4.692331
3,5.245720
3,7.269477
3,6.346542
3,3.894446
3,3.780490
3,7.419155
3,5.626811
3,8.373469
3,5.866250
3,6.852891
3,6.088649
3,5.271324
3,7.006189
3,7.900945
3,4.081222
3,4.204586
3,4.066615
3,6.931005
3,4.287889
3,6.230199
3,5.755036
3,7.299670
3,5.916281
3,4.679732
3,7.588978
3,5.668266
3,5.756166
3,5.998381
3,5.149174
3,5.745471
3,6.492545
3,6.074674
3,6.635681
3,6.162054
3,7.909955
3,7.319753
3,5.776466
4,4.683694
4,6.409731
4,6.545456
4,7.950362
4,7.396000
4,5.299221
4,4.794351
4,7.749124
4,5.795096
4,5.449844
4,5.765562
4,9.107632
4,5.989961
4,8.131689
4,5.895014
4,5.368511
4,6.142832
4,5.685852
4,7.938775
4,7.250659
4,4.394926
4,3.717136
4,7.094292
4,3.452336
4,5.892423
4,6.484538
4,7.077693
4,5.940706
4,7.259181
4,5.439107
4,6.636834
4,7.114068
4,5.498542
4,5.464169
4,6.975248
4,5.606841
4,7.638433
4,5.240799
4,5.447849
```

No questions assigned to the following page.

```
4,6.387571
4,5.730335
4,5.583096
4,6.529100
4,5.529919
4,6.885087
4,6.555909
4,6.802902
4,6.089306
4,4.632391
4,3.258059
5,5.445032
5,6.688775
5,4.566164
5,8.209309
5,6.591598
5,6.610729
5,8.008455
5,6.309230
5,7.457600
5,7.920924
5,6.305227
5,7.267295
5,5.719276
5,7.343328
5,4.006989
5,5.036844
5,4.761128
5,4.562326
5,6.935394
5,5.377649
5,4.788505
5,5.230463
5,7.446357
5,5.638407
5,5.741285
5,6.377323
5,7.575796
5,5.683282
5,5.825364
5,3.520843
5,5.156648
5,4.609719
5,6.241852
5,5.417844
5,5.963189
5,5.793820
5,5.553347
5,5.216368
5,6.728124
5,6.675112
5,5.125440
5,4.646518
5,6.281689
5,7.936931
5,5.819929
5,5.846428
5,4.814833
5,7.153393
5,5.758462
5,6.376159
6,6.452149
6,6.622496
6,7.412613
6,6.197177
6,6.537556
6,6.021881
6,7.263037
6,5.036601
6,4.658770
6,5.271508
6,5.835501
6,5.984419
6,5.771633
6,6.842897
6,6.805799
6,5.800254
6,4.121984
6,7.123685
6,5.580381
6,7.265037
6,5.797143
```

No questions assigned to the following page.

```
6,6.172956
6,5.440177
6,5.796555
6,5.816434
6,5.097917
6,5.763866
6,7.324661
6,6.773812
6,6.971123
6,6.450124
6,5.191527
6,6.935975
6,6.544650
6,6.075886
6,3.257355
6,6.548657
6,6.973000
6,5.759773
6,5.739847
6,5.671172
6,7.225243
6,6.889530
6,8.066126
6,4.020999
6,5.266090
6,5.065575
6,6.565765
6,8.101550
6,6.426256
7,4.854768
7,6.216267
7,5.516220
7,6.567409
7,7.288177
7,5.843946
7,6.404716
7,7.087657
7,5.970034
7,4.963026
7,4.580497
7,7.390320
7,7.263860
7,4.333863
7,6.261767
7,4.864749
7,6.045964
7,5.446123
7,5.726309
7,7.529437
7,5.211484
7,6.049766
7,5.334143
7,4.772034
7,6.402024
7,6.577651
7,5.715244
7,3.737089
7,5.996669
7,6.117982
7,7.430054
7,5.812813
7,6.967022
7,3.490324
7,5.367875
7,4.582016
7,5.494792
7,5.337199
7,8.537116
7,5.021267
7,6.193359
7,3.175836
7,5.518272
7,5.949558
7,5.892715
7,7.391536
7,6.529599
7,6.587681
7,6.445373
7,4.616642
8,6.513526
8,6.842956
8,5.801251
```

No questions assigned to the following page.

```
8,5.498790
8,6.869183
8,7.170215
8,6.727892
8,5.032628
8,7.074022
8,6.807990
8,5.997180
8,5.956650
8,7.443587
8,7.020246
8,4.659949
8,6.036737
8,6.563669
8,6.275421
8,6.197023
8,7.915371
8,8.355189
8,6.111192
8,6.816535
8,6.824060
8,6.834732
8,4.950058
8,6.145697
8,6.349865
8,6.563733
8,6.606444
8,8.130963
8,6.774511
8,5.939746
8,6.099638
8,6.015866
8,7.300601
8,2.550620
8,6.954689
8,4.444515
8,5.966893
8,6.785532
8,4.618662
8,7.522828
8,8.233653
8,5.913177
8,7.867970
8,7.281066
8,5.237486
8,6.855732
8,5.958054
9,7.711350
9,8.118175
9,6.549778
9,7.825989
9,8.102060
9,5.933953
9,6.062767
9,4.627150
9,5.750256
9,7.666728
9,7.730165
9,5.510727
9,6.564653
9,5.110511
9,6.280675
9,5.770042
9,7.236429
9,5.743896
9,6.592464
9,5.652575
9,7.291522
9,7.472144
9,7.363622
9,8.593219
9,5.342703
9,7.958807
9,5.016629
9,6.719463
9,5.618554
9,5.157983
9,6.036042
9,6.719279
9,5.646246
9,4.290551
9,7.000037
```

No questions assigned to the following page.

```
9,4.901637
9,5.746860
9,5.939277
9,6.266862
9,3.522050
9,6.038469
9,7.141624
9,4.648545
9,4.419259
9,7.037319
9,7.342750
9,6.755474
9,4.412591
9,7.290403
9,4.675565
10,4.768821
10,4.306492
10,7.189551
10,6.979083
10,6.033462
10,4.609813
10,5.848156
10,5.483386
10,4.332899
10,5.355768
10,6.338265
10,6.265044
10,6.865431
10,5.857124
10,6.532105
10,8.034613
10,5.844667
10,5.702362
10,7.778033
10,6.465588
10,7.419702
10,7.487725
10,7.613101
10,5.569379
10,6.674224
10,5.966037
10,6.045095
10,7.098283
10,4.879928
10,7.667796
10,5.200935
10,8.669504
10,5.176135
10,6.227697
10,6.694601
10,6.359037
10,4.764879
10,7.362724
10,6.836608
10,5.089343
10,5.371397
10,7.147577
10,5.873209
10,5.996439
10,5.973283
10,6.251001
10,4.325648
10,6.501863
10,6.800652
10,8.637065
11,5.346709
11,7.292967
11,6.067954
11,5.976624
11,7.228380
11,6.702282
11,5.952825
11,5.876282
11,7.240486
11,7.865036
11,6.017012
11,6.453768
11,5.889622
11,8.475720
11,4.753543
11,6.662296
11,6.957428
```

No questions assigned to the following page.

```
11,5.365771
11,9.498353
11,7.281119
11,4.511932
11,7.306444
11,6.282418
11,8.353334
11,6.476530
11,7.629290
11,6.571432
11,5.071208
11,6.097613
11,6.310751
11,7.043477
11,8.020770
11,6.583157
11,5.196676
11,8.852052
11,6.256870
11,5.088272
11,6.572631
11,5.837927
11,4.589601
11,8.199952
11,4.298361
11,7.777320
11,5.852809
11,7.612523
11,6.056613
11,7.362545
11,4.959005
11,7.233636
11,7.232901
12,7.619955
12,7.054643
12,6.320752
12,7.267876
12,6.312286
12,6.566037
12,8.974518
12,5.571554
12,6.312367
12,6.555815
12,7.823489
12,8.142946
12,7.445550
12,6.841224
12,7.088835
12,7.161542
12,6.498820
12,6.794918
12,7.319604
12,3.300026
12,6.956785
12,7.883189
12,5.472913
12,9.306229
12,6.330116
12,6.865971
12,7.368100
12,7.407018
12,8.116464
12,7.696744
12,6.005147
12,6.257281
12,6.727378
12,9.334161
12,7.070539
12,5.853977
12,6.529160
12,5.196538
12,6.736471
12,5.693305
12,8.269342
12,5.917299
12,5.278401
12,7.183820
12,5.652891
12,6.368623
12,6.008941
12,5.627343
12,8.090436
```

No questions assigned to the following page.

```
12,7.461486
13,7.693589
13,6.257702
13,7.744619
13,7.311703
13,7.875391
13,6.122809
13,6.837792
13,6.806125
13,6.214897
13,7.369171
13,7.754729
13,7.286703
13,6.902166
13,5.451980
13,6.177205
13,6.636800
13,7.874417
13,5.082905
13,4.795976
13,6.833707
13,6.904340
13,8.368183
13,7.317523
13,7.809924
13,5.678490
13,7.105152
13,7.674128
13,7.949699
13,7.472361
13,6.142934
13,4.251015
13,8.090834
13,5.267301
13,6.947123
13,5.460649
13,7.869510
13,6.760633
13,9.105385
13,8.963889
13,8.083833
13,5.841549
13,6.974577
13,5.584014
13,7.158578
13,7.565899
13,7.108529
13,5.263981
13,6.997354
13,7.573751
13,5.610311
14,6.237389
14,6.131223
14,5.994580
14,8.019872
14,5.956501
14,7.886078
14,6.855142
14,5.784224
14,5.404865
14,5.756748
14,6.073893
14,7.370506
14,7.624782
14,6.733186
14,7.955994
14,5.919343
14,6.564062
14,7.040765
14,8.103792
14,8.881194
14,7.012190
14,8.294164
14,7.521010
14,6.425617
14,7.163449
14,6.479069
14,7.407884
14,7.631348
14,8.327372
14,4.756984
14,6.005232
```

No questions assigned to the following page.

```
14,5.643417
14,6.146373
14,7.301133
14,6.195839
14,6.004494
14,7.581055
14,4.971226
14,10.119446
14,6.459920
14,6.968818
14,7.126947
14,7.842329
14,5.100421
14,5.784779
14,6.541277
14,5.851824
14,7.450815
14,6.242531
14,7.458605
15,7.627919
15,6.673848
15,5.620349
15,6.377470
15,6.442333
15,9.313435
15,5.785823
15,6.415104
15,6.576781
15,7.283558
15,6.252783
15,5.774086
15,7.186105
15,6.846120
15,5.416302
15,8.832227
15,7.989839
15,6.715754
15,5.868314
15,8.091849
15,5.088671
15,6.893911
15,7.275766
15,7.614611
15,6.728275
15,8.253085
15,6.712404
15,8.026599
15,5.989956
15,4.744729
15,6.620084
15,4.781579
15,7.219749
15,6.455483
15,6.176526
15,7.797944
15,8.238902
15,8.310269
15,7.941082
15,5.483794
15,6.670603
15,6.869635
15,6.529573
15,7.717752
15,8.187883
15,7.829160
15,8.597585
15,7.923699
15,7.433536
15,6.290273
16,9.689810
16,7.867752
16,9.985470
16,5.952494
16,9.021099
16,9.383097
16,4.838715
16,6.563203
16,8.013861
16,6.678713
16,7.802244
16,6.906424
16,8.949388
```

No questions assigned to the following page.

```
16,7.801343
16,7.931438
16,6.624877
16,9.008992
16,6.663159
16,6.141185
16,8.648319
16,5.682638
16,8.566025
16,5.776162
16,6.604790
16,6.416065
16,7.117943
16,7.504245
16,6.271917
16,7.657747
16,6.021499
16,7.419968
16,6.713184
16,7.398270
16,7.435792
16,7.864164
16,6.661872
16,6.574176
16,6.303328
16,7.690940
16,7.950119
16,4.641334
16,5.968819
16,5.510555
16,5.784332
16,7.289911
16,6.519936
16,7.756091
16,5.227972
16,5.556293
16,6.179343
17,8.028692
17,7.017837
17,4.746682
17,5.712006
17,5.190014
17,7.658967
17,7.230618
17,7.175504
17,7.887340
17,7.085356
17,5.859214
17,5.022408
17,7.891128
17,7.071442
17,8.269575
17,7.840899
17,6.510453
17,6.941205
17,6.065902
17,6.516448
17,7.008896
17,8.492190
17,8.535440
17,8.203192
17,8.017457
17,7.152813
17,7.516174
17,8.516708
17,8.233885
17,6.078529
17,7.843514
17,8.026006
17,8.486107
17,6.645236
17,6.845529
17,7.894089
17,7.006143
17,7.705775
17,5.926802
17,4.845711
17,6.373856
17,6.796450
17,6.778314
17,7.076330
17,8.118656
```

No questions assigned to the following page.

```
17,6.211551
17,7.725399
17,7.251464
17,7.200669
17,7.788442
18,6.666637
18,6.486846
18,8.790382
18,7.094035
18,7.526212
18,8.501291
18,8.722691
18,8.063106
18,7.929114
18,7.798321
18,5.968685
18,8.711229
18,7.328857
18,7.082121
18,7.101074
18,7.951326
18,7.068263
18,6.430582
18,7.682770
18,8.453293
18,5.501784
18,7.321223
18,7.528392
18,5.794485
18,6.197142
18,6.822105
18,6.937092
18,8.416033
18,8.497546
18,8.011089
18,5.294739
18,6.766145
18,7.645582
18,6.760501
18,8.255788
18,4.739490
18,7.760863
18,7.190156
18,7.356833
18,7.866600
18,8.058920
18,8.538983
18,9.322918
18,7.436637
18,5.827420
18,6.696146
18,6.772958
18,9.289395
18,7.032109
18,8.726424
19,9.333089
19,8.213519
19,7.794728
19,8.329071
19,6.204730
19,8.216593
19,7.138537
19,7.688442
19,5.424175
19,6.388854
19,6.082015
19,9.305728
19,6.932674
19,7.481222
19,8.341112
19,7.897842
19,7.899985
19,8.833647
19,8.678515
19,7.108617
19,7.636059
19,7.820535
19,7.808243
19,6.831329
19,6.907262
19,8.993926
19,5.031005
```

Question assigned to the following page: 2.4

```
19,8.334561
19,7.389548
19,7.223534
19,6.209655
19,7.072414
19,8.571289
19,10.091109
19,7.789440
19,9.035410
19,8.290691
19,7.526787
19,6.854107
19,5.798311
19,9.338998
19,8.716375
19,7.510946
19,8.050681
19,8.266458
19,6.765266
19,8.999146
19,6.035795
19,7.693754
19,8.080392"""

df = pd.read_csv(StringIO(data), header=None, names=["X", "Y"])

print(df)
```

```
       X         Y
0      0  6.397809
1      0  5.532191
2      0  5.123259
3      0  6.319232
4      0  7.136257
..   ..       ...
995  19  6.765266
996  19  8.999146
997  19  6.035795
998  19  7.693754
999  19  8.080392

[1000 rows x 2 columns]
```

```
X1 = df[['X']]
y1 = df['Y']

model4 = LinearRegression()
model4.fit(X1, y1)

beta0_model4 = model4.intercept_
beta1_model4 = model4.coef_[0]
R2_model4    = model4.score(X1, y1)

print("Estimated β₀ (intercept):", beta0_model4)
print("Estimated β₁ (slope):", beta1_model4)
print("R² for Model 1:", R2_model4)
```

```
Estimated β₀ (intercept): 5.541119861714286
Estimated β₁ (slope): 0.09751500687218047
R² for Model 1: 0.20082546227127673
```

```
import statsmodels.api as sm

X = sm.add_constant(df["X"])  # adds the intercept term
model = sm.OLS(df["Y"], X).fit()


beta0 = model.params["const"]
beta1 = model.params["X"]

print("\n(a) Fitted Linear Model:")
print("Estimated β₀ (intercept):", beta0)
print("Estimated β₁ (slope):", beta1)

r_squared = model.rsquared
print("\n(b) R² =", r_squared)
```

```

(a) Fitted Linear Model:
Estimated β₀ (intercept): 5.541119861714282
Estimated β₁ (slope): 0.09751500687218076

(b) R² = 0.2008254622712764
```

Questions assigned to the following page: 2.4 and 2.5

R^2 =0.2008 means that approximately 20.08% of the variability in Y is explained by X. R^2 of 0.20 is relatively low; this suggests that the linear model does not explain much of the variation in Y. In other words, while there is a statistically detectable relationship between X and Y, many other factors likely influence Y that are not included in the model.

```
p_value = model.pvalues["X"]
print("\n(c) p-value for β₁:", p_value)
```

```
(c) p-value for β₁: 1.470018500836489e-50
```

The p-value < 0.05 so we reject the null hypot (H0) and we can conclude there's a statistically significant relationship

```
se_beta1 = model.bse["X"]
CI_lower = beta1 - 2 * se_beta1
CI_upper = beta1 + 2 * se_beta1
print("\n(d) Standard Error for β₁:", se_beta1)
print("95% Confidence Interval for β₁: ({:.4f}, {:.4f})".format(CI_lower, CI_upper))
```

```
(d) Standard Error for β₁: 0.0061576870999215235
95% Confidence Interval for β₁: (0.0852, 0.1098)
```

Since we're 95% certain the value of β_1 lies between (0.0852, 0.1098) and the entire intervak is <1 we would not consider β_1 to be meaningfully different than 0.

e) Although the slope is statistically significantly different from 0 and the model explains about 20% of the variance in Y, the estimated effect size is very small. Our 95% confidence interval for β_1 lies below the practical significance threshold of 1. Thus, while we detect a statistically significant linear relationship, the effect is too small to be considered meaningful in practice.

```
data = """2.0,53
1.8,58
3.7,56
2.2,50
2.1,56
2.4,53
2.6,65
2.8,60
3.3,75
3.5,64
3.7,64
3.8,72
4.5,87
4.7,78
4.0,77
4.0,72
1.7,44
1.8,51
4.9,74
4.2,81
4.3,73"""

df = pd.read_csv(StringIO(data), header=None, names=["Duration", "Next"])

df.head()
```

	Duration	Next
0	2.0	53
1	1.8	58
2	3.7	56
3	2.2	50
4	2.1	56

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
import statsmodels.api as sm

X = sm.add_constant(df["Duration"])  # adds the intercept term
model = sm.OLS(df["Next"], X).fit()
```

Questions assigned to the following page: 3.1 and 2.5

```
# Get estimated coefficients and R²
beta0 = model.params["const"]
beta1 = model.params["Duration"]
R2 = model.rsquared

print("\n(a) Fitted Linear Model:")
print("Estimated β₀ (intercept):", beta0)
print("Estimated β₁ (slope):", beta1)
print("R²:", R2)

# (b) 95% prediction interval
new_obs = pd.DataFrame({"Duration": [5]})
new_obs = sm.add_constant(new_obs, has_constant='add')
pred = model.get_prediction(new_obs)
pred_summary = pred.summary_frame(alpha=0.05)  # 95% prediction interva

print("\n(b) Prediction for Duration = 5 minutes:")
print(pred_summary)
```

```
(a) Fitted Linear Model:
Estimated β₀ (intercept): 33.560815731973776
Estimated β₁ (slope): 9.6797480827728
R²: 0.7482018074174728

(b) Prediction for Duration = 5 minutes:
        mean   mean_se  mean_ci_lower  mean_ci_upper  obs_ci_lower  \
0  81.959556  2.628451      76.458144      87.460968      68.10739

   obs_ci_upper
0     95.811722
```

The R² value for this model is approximately 0.748 (or 74.8%), which indicates that about 75% of the variability in the time until the next aftershock is explained by the duration of the current aftershock. R² of around 0.75 is considered to be quite high, suggesting that the linear model does a good job of capturing the relationship between the two variables.

The model predicts that when the duration is 5 minutes, the expected time until the next aftershock is about 81.96 minutes. The standard error (2.63 minutes) measures the uncertainty in the estimated mean prediction.

Questions assigned to the following page: 3.2 and 3.3

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^{Tuna}, $X_2^{\text{Swordfish}}$, and X_2^{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^{Canned}, $X_2^{\text{Commercial}}$, and X_2^{Sashimi} using the same encoding.

(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 β_0, β_1, ... to denote the coefficients and write the model in the form of $Y = \beta X + \ldots + \epsilon$.

(c) **(10 points)** For Dataset A, how do you interpret each of the coefficients in your model? Your answer should include interaction terms (for example, $\beta X_i X_j$). **Hint.** When doing interpretation, try to discuss by cases. For example, when the fish species is Tuna/Swordfish/Blobfish.

Here β's for A & β's for B

b) For A: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{\text{Tuna}} + \beta_3 X_2^{\text{Swordfish}} + \beta_4 X_2^{\text{Blobfish}} + \beta_5 X_1 X_2^{\text{Tuna}} + \beta_6 X_1 X_2^{\text{Swordfish}} + \beta_7 X_1 X_2^{\text{Blobfish}} + \epsilon$

For B: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{\text{Canned}} + \beta_3 X_2^{\text{Commercial}} + \beta_4 X_2^{\text{Sashimi}} + \beta_5 X_1 X_2^{\text{Canned}} + \beta_6 X_1 X_2^{\text{Commercial}} + \beta_7 X_1 X_2^{\text{Sashimi}} + \epsilon$

where Y is the num of sales

c)
- β_0 is the baseline sales when weight is 0 (theoretical)
- β_1 is the general avg incr in sales for 1 unit incr in weight across all species
- $\beta_2, \beta_3, \beta_4$ Additional avg incr to sales when Tuna, Swordfish or Blobfish or Canned, Commercial or Sashimi
- $\beta_5, \beta_6, \beta_7$ interaction terms that capture the avg 'extra incr' in change of sales per one unit change in weight for different species

when Tuna: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{\text{Tuna}} + \beta_5 X_1 X_2^{\text{Tuna}} + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_5 X_1 + \epsilon$

β_2 is the avg additional amt to sales when Tuna

β_5 is the avg 'extra incr' in change of sales per one unit change in weight for Tuna fish vs others (Blowfish & Swordfish)

when Sashimi: $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_2^{\text{Sashimi}} + \beta_7 X_1 X_2^{\text{Sashimi}} + \epsilon = (\beta_0 + \beta_4) + (\beta_1 + \beta_7) X_1 + \epsilon$

β_4 is the avg additional amt to sales when Sashimi

β_7 is the avg 'extra incr' in change of sales per one unit change in weight for Sashimi fish vs other (Canned & Commercial)

$Y = \beta_0 + \beta_4$ when weight of Sashimi is 0

$\beta_1 + \beta_7$ is the avg incr in sales per unit incr in weight of Sashimi

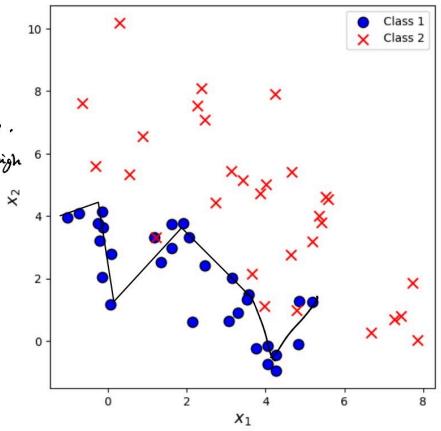
Questions assigned to the following page: 4.1 and 4.2

4 Bias, Variance and Regularization

(a) **(2 points)**

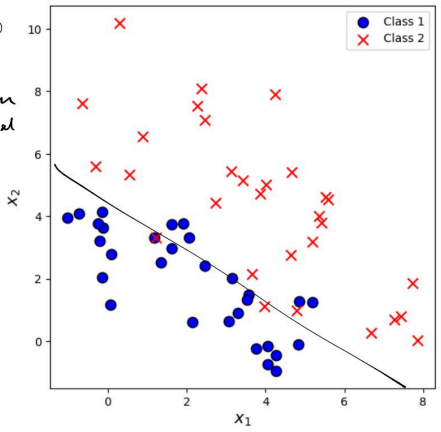
The figure below shows a labeled dataset. Could you draw a decision boundary where the classifier is overfitting? How will the training and test errors change over time? What are some potential reasons for overfitting to occur? How do you think the model will perform on the test set?

a) Overfitting will lead to very low training error because the training data has been memorised but the test error will be high because the decision boundry is to sensitive to variations & noise. Overfitting could occur because the polynomial degree is too high or we have too many predictors



(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?

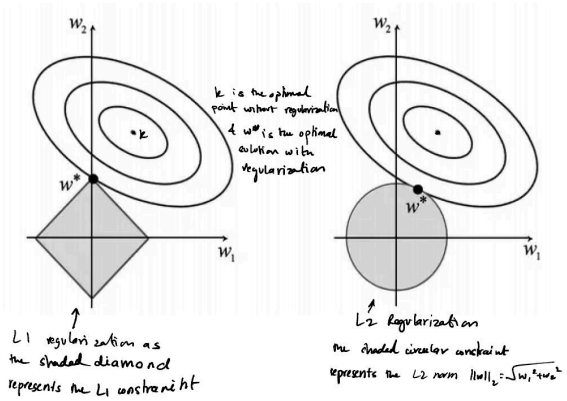
b) An underfitted model would have high training & testing error. It could be underfitted for several reasons such as the model is to simple linear like the line I drew or there's not enough predictors or the polynomial degree is too high



(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

Question assigned to the following page: 4.3

which one is with L2 regularization? Why? (3) Under what conditions would you use these regularizations?



- we use L1 when feature selection is imp & when there is a lot of irrelevant & redundant features
- we use L2 when we want to reduce variation while keeping all the features & when features are correlated & we dont want to arbitrarily remove same.

Questions assigned to the following page: 5.1, 5.2, 5.3, 5.4, and 5.5

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 $Y = 1$?

(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?

(c) **(2 points)** Explain how increasing or decreasing $\beta_0, \beta_1\beta_2, \beta_3$ affect our predictions.

(d) **(2 points)** What is the formulation of the decision boundary? Which points are on the decision boundary?

(e) **(2 points)** 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).

a) $\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

$= 2 + 5(4) + (-2)(8) + (-3)(2)$

$= 2 + 20 - 16 - 6 = 0$

$\frac{P(y=1)}{1-P(y=1)} = e^0 = 1 \Rightarrow P(y=1) = 1 - P(y=1)$

$2(P(y=1)) = 1$

$P(y=1) = 1/2 = 0.5$

b) Incr X_1 by 2 means $\beta_1 X_1$ incr by 10 so log odds $= 0+10 = 10$ & odds $= \exp(10)$

Decr X_2 by 1 means $\beta_2 X_2$ incr by 2 $(-2(-1) = 2)$ so log odds $= 0+2 = 2$ & odds $= \exp(2)$

c) Incr β_0 shifts the log-odds upwards bc it's the intercept, i.e. incr the overall probability of $y=1$

Incr β_1 increases the effect of X_1, making $y=1$ more sensitive to X_1

Decr β_2 or β_3 (so making them more neg) will lower the prob of $y=1$

d) The decision boundary is at $P(y=1) = 0.5$ so $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = 0$

$2 + 5X_1 - 2X_2 - 3X_3 = 0$ this is a linear decision boundary in (X_1, X_2, X_3) space

e) When we fit other logistic regression models with only X_1 or only X_2, the coeffs are different bc multicollinearity or confounding variable can be present when X_1, X_2, X_3 are considered together their coefficient account for their combined effect if we remove a variable the others will make up for some parts of its effect, changing it's original value This can be problematic as it can lead to misinterpretation or biased predictions