Ethan Wong
305319001

CS M148 HW1

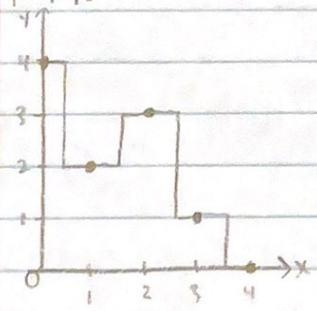
1. a)
- Voluntary Bias - only people who want to share their opinion will post online
- Response Bias - people who share their opinion online tend to have strong feelings on the issue
- Under-coverage bias - not all students use Reddit
- Over-coverage bias - students could make multiple posts on multiple accounts, causing their opinion to be magnified
- Non-response bias - not every student with a Reddit account will make a post about the issue

b) i) The tool probably saw that successful applicants were mostly male, as there were just more males applying/getting hired. The model then began looking at features of these males and prioritized those (such as 'gender' feature). This caused applicants with 'gender = female' to get discriminated against.

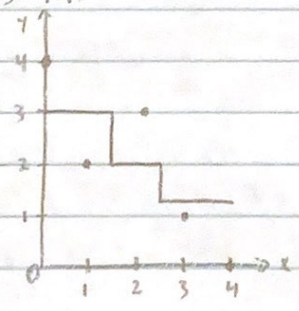
ii) Dropping the 'gender' field from the data may not entirely resolve this issue, as there can be other information in an application that hints at an applicant's gender. For instance, if they attended an all-girls school or were in a woman's sports team, the gender can be inferred.

2. a)

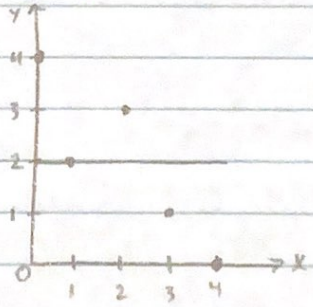
1-NN



3-NN



5-NN



$(0,4)$ $\frac{4+2+3}{3} = 3$

$(1,2)$ $\frac{4+2+3}{3} = 3$

$(2,3)$ $\frac{2+3+1}{3} = 2$

$(3,1)$ $\frac{3+1+0}{3} = 4/3$

$(4,0)$ $\frac{3+1+0}{3} = 4/3$

2 b) (0,3,3) (1.8,2) (3.8,1)

1-NN: (0.3,3) (1.8,2) (3.8,1)
(0.3,3) (1.8,2) (3.8,1)
$[(3-3)^2 + (2-2)^2 + (1-1)^2] \div 3 = 0.00$ MSE

3-NN: (0.3,3) (1.8,2) (3.8,1)
(0.3,2) (1.8,2) (3.8,2)
$[(3-2)^2 + (2-2)^2 + (1-2)^2] \div 3 = 0.67$ MSE

5-NN: (0.3,3) (1.8,2) (3.8,1)
(0.3,1.2) (1.8,1.2) (3.8,1.2)
$[(3-1.2)^2 + (2-1.2)^2 + (1-1.2)^2] \div 3 = 1.31$ MSE

K=1 yields the best results as the MSE = 0.00. However, these perfect predictions will not generalize well as the model is overfitted specifically to this dataset. Realistically, K=3 is probably the best option since the MSE is relatively low and it will generalize better.

c) If using 1-NN on training data, then $R^2 = 1$. This is a bad model since it is overfitted to the data. If using n-NN (where n = number of data points) then the $R^2 = 0$. This is also a bad model as it is underfitted to the data.

3. 1a) Intercept: $\hat{\beta}_0 = -2.7327 \times 10^9$ Slope: $\hat{\beta}_1 = 1.4907 \times 10^6$

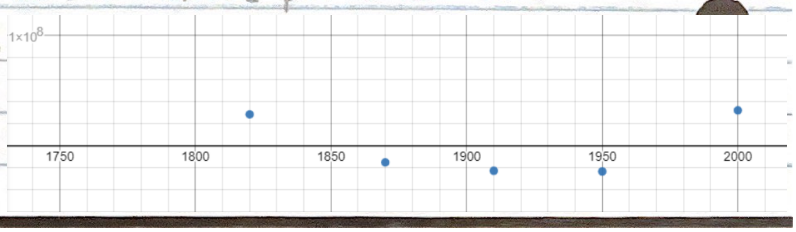
$\hat{y} = 1490700x - 2732700000$

Population in 2010: 2.6367×10^8 people

1b) The R^2 is 0.9323, which means that the estimated regression line fits the data very well. A higher R^2 could be achieved if not using a straight line, but a curve instead to fit the data.

1c) The residuals are in a parabola, which means this model isn't ideal. For a linear model, we want to see that the plot of residuals exhibits a random pattern.

ATTACH PHOTO



3. 2) From the plot and its relatively high R^2 value, we can see that drinking more wine is indeed strongly correlated with reduced heart disease. However, a strong correlation does not necessarily imply causation. There could be other lifestyle choices among various countries (such as proper diet and exercise) that truly explain the difference in heart disease deaths.

3a) $\overset{y}{\text{Consumption}}$ based on $\overset{x}{\text{Income}}$

↳ Intercept: $\hat{\beta}_0 = 4.27$ Slope: $\hat{\beta}_1 = 0.62$ $R^2 = 0.5816$

$\overset{y}{\text{Income}}$ based on $\overset{x}{\text{Working Experience}}$

↳ Intercept: $\hat{\beta}_0 = 35.4$ Slope: $\hat{\beta}_1 = 7.58$ $R^2 = 0.6027$

3b) Consumption based on Income: $R^2 = 0.5816$

Income based on Working Experience: $R^2 = 0.6027$

- There is a slight association between consumption and income that indicates that individuals who earn more money tend to also spend more money.
- There is a slight association between income and working experience that indicates individuals who have more experience tend to make more money (most likely due to larger skillset).

4a) Intercept: $\hat{\beta}_0 = 4.53$ Slope: $\hat{\beta}_1 = 0.0992$

4b) For this linear model, we observe $R^2 = 0.2107$. This is quite low, which shows a weak linear relationship between x and y.

4c) β_1 has a standard error of 0.006, and a t-value of 16.323. The value $P > |t|$ is 0.000, which is less than the standard p-value threshold of 0.05. Thus, we reject the null hypothesis and say there is actually a significant linear relationship between x and y.

4d) 95% confidence interval: $[0.0872, 0.1112]$

This confidence interval does not suggest that β_1 is meaningfully different from 0.

4e) I believe that this contradiction is the result of a large sample size (1000 points). The sample slope is significantly different from 0 while not being meaningfully different from 0. When analyzing data, it would probably be smart to look at a scatter plot that comes with the linear regression to help visualize the data.

5a) The linear model gives $R^2 = 0.749$. This is a pretty good linear model since $R^2 > 0.5$ by a decent amount, but it is still far from perfect ($R^2 = 1$).

5b) We get a 95% prediction interval of $\{47.237, 73.529\}$ minutes. This result indicates that we are 95% confident that the time until the next eruption falls between 47.237 and 73.529 minutes.

5c) No. You can not determine if you can see the eruption as 50 minutes falls in the range of 47.237 minutes to 73.529 minutes. You can not be certain that the eruption will occur in the next 50 minutes.

4. a) Option 2 is preferable. This option ensures that there is no ordinal relationship between the three fish species (ie 1 is more similar to 2 than it is to 3). Option 1 does imply this ordinal relationship, so Option 2 (one-hot encoding) makes better sense.

b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^A + \beta_3 X_2^B + \beta_4 X_1 X_2^A + \beta_5 X_1 X_2^B + \epsilon$

$$= \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 + \beta_4 X_1 + \epsilon & \text{fish species A} \\ \beta_0 + \beta_1 X_1 + \beta_3 + \beta_5 X_1 + \epsilon & \text{fish species B} \\ \beta_0 + \beta_1 X_1 + \epsilon & \text{else / otherwise} \end{cases}$$

c) β_0 : expected market sales of Fish Species C at 0 weight

β_1 : expected market sales for Fish Species C per unit change in weight

β_2 : expected difference in market sales between Fish Species A and C at 0 weight

β_3 : expected difference in market sales between Fish Species B and C at 0 weight

β_4 : expected difference in market sales between Fish species A and C per unit change in weight

β_5 : expected difference in market sales between Fish species B and C per unit change in weight

5. a) The accuracy would be 99%. Accuracy is not a very useful measurement in this case because it is an extremely unbalanced dataset. It is smarter to use precision since it looks at the number of correct positives over the total number of positive guesses. This ratio is less sensitive to unbalanced datasets, making it much more suited to this situation.

b) The problem with this dataset is that it is extremely unbalanced. A method that may help resolve this problem is to augment the dataset with synthetic data samples of the minority class.