**Krish Patel**

**CS M148 Homework 1**

1. **Data and bias**

(a)When using a non-probabilistic method of selecting the sample population, there could be a lot of different biases that could skew the data. For instance, it would create Response bias, where there is a difference between the respondents (In this case Reddit Users) and the non respondents(Non-users). This form of data might also exhibit Undercoverage bias with respect to non reddit users, and would also lead to convenience bias when it comes to taking the “responses” from Reddit users

(b) The tool here might be discriminating between women and men due to underlying correlations with other factors, for instance, due to underrepresentation of women in STEM may lead to biases against women in the AI model. Eliminating the gender might not work in this scenario as there may be other features in that may be correlated with the gender feature, such as experience, education (such as stats suggesting women having education rates lower than men in developing countries, a correlation that an AI tool might have picked up.

1. **Linear Regression: goodness of fit & Interpretation**

1.

(a)The values for the MSE loss:

A black and white math equation

Description automatically generated with medium confidence

Using the above formula, and plugging in the values (X = year, Y = population)

Mean x = 1910, mean y = 114.6 million

Based on that, β0 = (9504+2984+0+1456+14976)\*10^6/19400 = 1490721.6

β1 = 114.6 mil – 1910\*1490721.6 = -2732678351

(b)R is 1- 0.0671 = 0.9323, which is a pretty good R^2 value, This is better than

(c)The plot residual suggest that this is not a good model, given the large residual values at the extremes (x1 and x5), it can be deduced that the model only tends to fit the value nera the mean, and better suits a, for instance, quadratic model.

2.

There seems to be a moderate negative correlation between heart disease and wine collection, as we can see from the R^2 value of 0.71. However, the R^2 value between two variables or predictors doesn’t necessarily suggest a causational effect. Thus, we could say there is a correlation, however other factors may contribute to this data and we can’t rule other factors out and say that X causes Y.

3.

A screen shot of a computer program

Description automatically generated

(base) krishpatel@Krishs-Air Tennis % /opt/homebrew/bin/python3 /Users/krishpatel/Desktop/m148h1q3.py

For consumption based on income:

beta0: 4.270916376452803

beta1: 0.6199702093735578

R^2: 0.5815606212788424

For income based on working experience:

beta0: [35.4]

beta1: [7.58]

R^2: 0.6026979398313262

According to the values above, values, based on income, the R^2 is ~0.5815606212, i.e 58.15606212% of the variation in consumption can be explained by income(better than the average values of Y by 8.15606212%

For the income based on working experience, the relation has an R^2 of 0.602697939… which mean that 60.26979398313262% of the variation in income can be explained by the working experience in years.

4.

(a)

A screen shot of a computer program

Description automatically generated

(base) krishpatel@Krishs-Air Desktop % /opt/homebrew/bin/python3 /Users/krishpatel/Desktop/m148hw1q3.py

The model's coefficients are: [0.09996037]

The model's intercept is: 5.00574610922345

(b)

A screen shot of a computer screen

Description automatically generated

(base) krishpatel@Krishs-Air Desktop % /opt/homebrew/bin/python3 /Users/krishpatel/Desktop/m148hw1q3.py

The model's coefficients are: [0.09996037]

The model's intercept is: 5.00574610922345

The r^2 value is: 0.23351554974225897

The r^2 value is 0.233, thus suggesting that it performs worse than the average. It indicates that 23.3% of the y values are suggested by x.

4.

(a) (base) krishpatel@Krishs-Air Desktop % /opt/homebrew/bin/python3 /Users/krishpatel/Desktop/m148hw1q3.py

The model's coefficients are: [0.09996037]

The model's intercept is: 5.00574610922345

The r^2 value is: 0.23351554974225897

OLS Regression Results

==========================================================================

Dep. Variable: y R-squared: 0.245

Model: OLS Adj. R-squared: 0.244

Method: Least Squares F-statistic: 259.4

Date: Thu, 15 Feb 2024 Prob (F-statistic): 9.68e-51

Time: 14:25:04 Log-Likelihood: -1146.6

No. Observations: 800 AIC: 2297.

Df Residuals: 798 BIC: 2307.

Df Model: 1

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 5.0057 0.069 72.708 0.000 4.871 5.141

x 0.1000 0.006 16.105 0.000 0.088 0.112

==============================================================================

Omnibus: 0.149 Durbin-Watson: 2.134

Prob(Omnibus): 0.928 Jarque-Bera (JB): 0.204

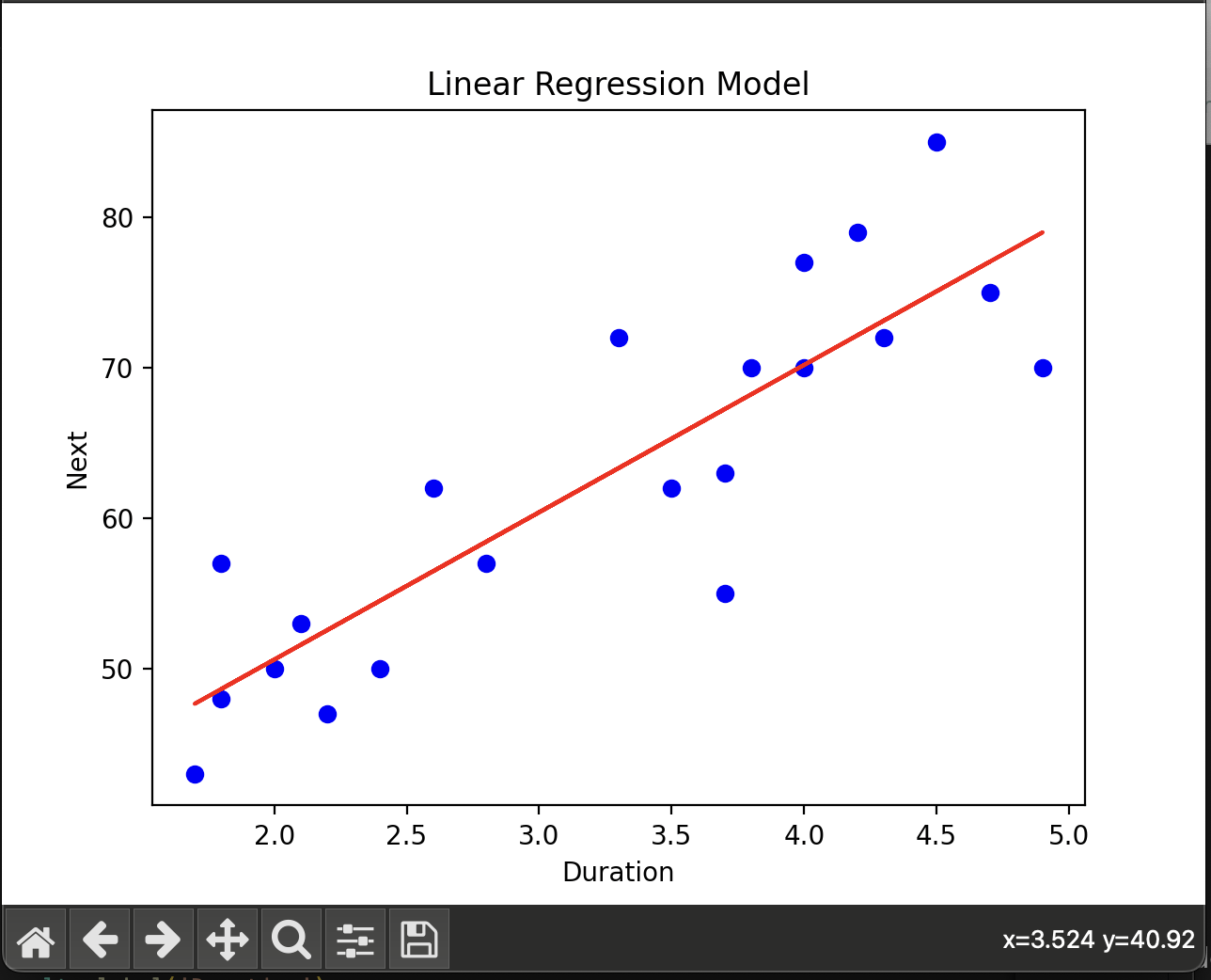
Skew: -0.029 Prob(JB): 0.903

Kurtosis: 2.947 Cond. No. 21.4

==============================================================================

According to the above graph, as the pvalue is zero, thus we reject the null hypothesis and we assume that x is statistically significant

5



A screen shot of a computer program

Description automatically generated

(base) krishpatel@Krishs-Air Desktop % python3 hw1csm148q4.py

Coefficient: 9.790068977336013

Intercept: 31.013109978150055

R^2: 0.7490804158751225

Based on the output, and the graph, it can be inferred that a linear regression model is probably the best model that would fit the data. With an R^2 value of 0.7491, 74.91% of the variation in the next eruptions could be explained by the duration of the least eruptions.

(base) krishpatel@Krishs-Air Desktop % /opt/homebrew/bin/py

thon3 /Users/krishpatel/Desktop/hw1csm148q4.py

Coefficient: 9.790068977336013

Intercept: 31.013109978150055

R^2: 0.7490804158751225

Prediction for next eruption (5 minutes): 92.44699282237437

95% Prediction Interval for next eruption (5 minutes):

[84.85963825068345, 100.03434739406529]

95% Prediction Interval:

Observation 1: [46.21343208991126, 54.973063775732875]

Observation 2: [43.82432648289788, 53.44614179181185]

Observation 3: [64.1679542330674, 70.30477615551916]

Observation 4: [48.57471384712223, 56.527809609456305]

Observation 5: [47.39806796960891, 55.74644169150242]

Observation 6: [50.89883633164737, 58.11971471586556]

Observation 7: [53.1733856861581, 59.76119295228925]

Observation 8: [55.38284966340692, 61.46775656597482]

Observation 9: [60.5160253693267, 66.12464983739106]

Observation 10: [62.389829193821264, 68.16687360383092]

Observation 11: [64.1679542330674, 70.30477615551916]

Observation 12: [65.02585638752065, 71.40488779653312]

Observation 13: [70.63858902805751, 79.49825172426668]

Observation 14: [72.16269585143522, 81.89017249182339]

Observation 15: [66.69015699171221, 73.65661478327597]

Observation 16: [66.69015699171221, 73.65661478327597]

Observation 17: [42.62155473449234, 52.690899744750176]

Observation 18: [43.82432648289788, 53.44614179181185]

Observation 19: [73.6664890543838, 84.30240687980923]

Observation 20: [68.29930514774566, 75.96349421817695]

Observation 21: [69.08759003338906, 77.13322312800074]

Based on this

The

**3.Interpretation of Coefficients in Linear Regression**

1. It would be better to convert by doing one hot encoding, as using one variable would not capture the different changes that the fish species would have on the weight correctly. Thus (2) would be the best option
2. The equation that I can derive from it is B0 + B1X1 + B2X2(A) +B3X2(B)+B4X2(C) + B5X1X2(A)+ B6X1X2(B)+ B7X1X2(C) + E
3. To interpret these coefficients in my model, I would use different circumstances, where X2(N) is either true or false, N is either species A, B, or C
4. When species A, the price of the fish is calculated as B0 + B1X1 + B2 +B5X1 = (B0+B2)+(B1+B5)X1.

When species B, the price of the fish is calculated as B0 + B1X1 + B3 +B6X1 = (B0+B3)+(B1+B6)X1.

When species A, the price of the fish is calculated as B0 + B1X1 + B4 +B7X1 = (B0+B4)+(B1+B7)X1.

This is how we can interpret the model

The (B1 + BN)X1 defines the sensitivity of the fish species to the price with respect to the base (when the weight is zero). Therefore, an increase in 1 unit wight, increases the price of the fish by (B1 + BN), where N is either B5, B6, B7 based on the fish species.

**4.Bias, Variance and Regularization**

(a)According to the following figures, the left model has the highest bias as it doesn’t take into account of the distribution properly. This is a result of using a suboptimal decision boundary. Increasing the complexity of the model should solve this issue, This would give a high error rate on the test set.

The right model has high variance, as it fits to the outliers as well, giving a very unique shape to the curve. This model is likely to perform worse on the test set, as it fits to the noise (overfitting) and regularizing this model will help with increasing the model’s performance on the test set.

The center model seems to accurately model the classification, as the outliers don’t affect the decision boundary. Thus, theoretically, this should perform best out of all 3 models when used in the test set.

(b) According to the graphs, the black line mostly uses l1 regression, as you only get sparse solutions when using l1 regularization(LASSO), which can inferred from the model above.

**5. Logistic Regression**

(a) log of odds = 3 + 5 -8 which is equal to 0. That means that there is an equal probability that Y =1 (probability is 0.5)

(b)Increasing the value of X1 by 1 increases log odds by 0 +5 = 5. Thus the odds ratio increases by e^5. However, when the value of the X1 decreases by 1 to 0, the log of odds is -5. Thus, the odds are e^(-5). [P(Y=1)/P(Y=0)].

(c)Increasing the value of B1 by 1 unit increases the odds of P(y=1) by e^X1. Increasing the value of B2 by 1 unit increases the log of odds e^X2

(d)The formulation of the decision boundary is B0 + B1X1 + B2X2 = 0, which gives the decision boundary. This is because at this value, the log of odds is 0, thus there is an equal ratio of P(Y=1) and P(Y=0).

(e)This mean that there is dependence (some sort of correlation) between the two vairables. Usually the features that we choose are meant to be independent of each other(and thus ideally should have a correlation of 0. However, this means that some of coefficients of one feature are leaking into the other, and thus may not be optimal when working on the test data(or unseen data). In this case, PCA would be something that would help this.