Homework 5 Vedant Desai

May 3, 2023

1a) According to the model, a one million dollar increase in healthcare expenditure predicts an increase of 0.3269 years of life expectancy.

1b) The model explains 70.7% of variation

1c)The model will predict a value of 78.3453 for that observation

1d)Yes we can reject the null hypothesis that there is no relationship between GDP and life expectancy as the p-value (0.00) is less than the threshold for rejection(0.05)

1e)The regression coefficient for fertility will always be between -5.174 and -3.743.

1f) The non-linearity won't affect us from doing the regression but our interpretation and inferences of data's validity should be checked and inspected.

1g)We are conducting multiple regression in this question.

1h)In this question the researcher is carrying out a statistical approach to regression.

2a) The above research is an observational study as the researcher is simply observing the trend and not affecting the variables.

2b) If the researcher had selected on the dependent variable, it will lead to the implication that they had chosen observations for their analysis based on their outcome or response variable. This leads to a selection bias. This can cause distorted or misleading results and compromise the validity of the research design.

2c)In this study the life expectancy can affect the expenditure on health. Because if people are living longer the healthcare spending of a country will also increase

2d)Confounders are a concern in this study.A plausible confounder can be political stability and state of the country(war or peace) as it affects all the independent and the dependent varaible as well.

2e)The scientist is selecting on the dependent variable and causing a selection bias. The exclaimatory scientist is also causing something called anecdotal fallacy.

2f)This would be an ecological fallacy.

2g)The patients in Bellevue hospital can be the collider. It might cause distorted associations.

3a)In Supervised ML both inputs and output data sets are considered, whereas in unsupervised ML only input data set is used to find structures based on that data.

3b)One problem with 100 percent accuracy is that when classes are very unbalanced, it's easy to be accurate. 100 percent accuracy can lead to overfitting as well i.e the model may fail to generalize well to new data.

3c)RMSE calculates the average error(distance) between the predicted value of the model and the actual value of the data set, whereas R-squared measures the proportion of variablity of the dependent variable in the model.

3d) Because it helps in avoiding ties when making a prediction.

3e)Its optimal to choose the intial centroids which are representative of the data and are well spread out. It would also help while choosing to run the k-means algorithm multiple times with different centroids to choose the best one based on the ouutputs.

3f)We can use kNN with continuous variables by turning them into categorical variables(i.e bins or categories)

3g) The assumption about documents that word order doesn't matter.

3h)In reinforcement learning, behaving "optimally" means that choosing actions that maximize the expected overall reward over time

4a)The subjects were the 6500 volunteers.

4b) The study is deemed unethical as the subjects were not deciding what happens to their private data.The researchers defended the accustions by stating that they gathered the data with user's consent

4c)From the specific perspective of the law and public interest, the study might be deemed unethical on the grounds of privacy concerns and accountablity for actions

5a) The article notes dutch nationality as a risk factor for determining whether an applicant had potentially committed fraud.

5b)Because of the non-dutch nationality, people from ethnic minorities were disproportionately inpacted.

5c)The bias was not resolved because the civil servants that were required to conduct the review were not given any information on why the system had generated a high risk score (such 'black box' systems led to the bias being unresolved)

5d)Because of an incentive for the tax authorities to seize as many funds as possible regardless of the veracity of the fraud accusations, as they had to prove the efficiency of the algorithmic decision-making system.

[ ]: