Establish

Our dataset is related to the occurrence of heart disease in a study of 500 towns. Data was collected on the proportion of the population with heart disease and the proportion of the townsfolk that smoked or cycled. We want to use linear regression to understand if the proportion of a town's population that cycles reduces the proportion of a town's heart disease. Similarly, we want to understand whether the proportion of smokers across towns increases the proportion of heart disease.

Empathise

While the aim is precise, we don't know whether the collected data will produce a useful linear model to predict heart disease outcomes. One advantage is that we have only two independent variables. Therefore, feature selection is not needed. However, we must check whether our "independent" variables are dependent. We can validate this case using a technique called multicollinearity. Other issues we need to check for are observations that may cause inconsistent predictions as part of the derived model.

Explain

From the model output table, we see that both independent variables biking and smoking, have a very low pr>t value. These results indicate a high correlation with the dependent variable heart disease. Additionally, the adjusted R^2 value 0.98 indicates that the fitted model contains 98% of the collected data, which indicates that the model may be a good predictor. Furthermore, from the diagnostic plots, we note that the plots show that very few residuals have a significant Cooks distance; in other words, for the observations that don't pass through the linear model, they are not far away from the model line. The residuals themselves indicate that the distribution is not normally distributed, which could indicate a poor fit for specific observations. Finally, our multicollinearity check determined that our independent variables biking and smoking, are indeed independent with scores of 1, respectively.

Enlighten

While the model checks provide a good result in almost all cases (except in the normal distribution of residuals), checking whether the model can provide a reasonable prediction is the most important outcome. Overall, we found significant relationships between the frequency of biking to work, smoking and the frequency of heart disease. We found a 0.2% decrease in the frequency of heart disease for every 1% increase in biking and a 0.178% increase in the frequency of heart disease for every 1% increase in smoking. If you consider the model a linear equation, you could predict the

proportion of heart disease by knowing the proportion of towns that cycle or smoke. On average, our model predicts the proportions of heart disease with 94% accuracy. While the model is not perfect, it is pretty useful.