Multiple choice problems (2 points each)

1

The average prevalence of lung cancer across the entire state is 31.6 cases per 100,000 individuals. A doctor wants to determine how closely the prevalence of lung cancer for her patients from a rural area of that state aligns with the state wide average. She calculates a 90% confidence interval of $[37.2, 40.1]$ for the average prevalence (per 100,000 individuals) of lung cancer based on a random sample of her patients and patients from nearby doctors over the last three years. Which of the following statements are valid interpretations of this confidence interval? Circle all that apply.

(a) This confidence interval indicates that the prevalence of lung cancer in this area is likely much higher than the state-wide average.

(b) If another doctor were to conduct the same study on a new random sample of patients from the region, there is a 90% chance that he would calculate the same confidence interval.

(c) If this study were to repeated with new random samples, the resulting confidence intervals would contain the true prevalence of lung cancer in this area 90% of the time.

(d) There is a 10% chance that the interval $[37.2, 40.1]$ does not capture the true prevalence of lung cancer (per 100,000 individuals) for this area.

5

In which of the following situations do we need to asses the normality and randomness conditions for inference in regression? Circle all that apply.

(a) A scientist studying a random sample of sparrows at Kent Island wants to determine how the average weight (in grams) of the birds changes for each additional mm in wing length.

(b) Someone interested in selling their house creates a SLR model using the list price of local houses as a predictor of the final sale price of each house once sold. They want to predict the actual sale price of their home for a list price of $189,000.

(c) A nutritionist wants to determine if an increase in sugar content (in grams) of breakfast cereal corresponds to a positive (non-zero) change in calories per serving while also obtaining an upper and lower bound on the size of the increase.

(d) A used car salesperson wants to determine if an increase in the number of miles on a car has a statistically significant impact on the sale price of the vehicle.

Q no. 6

1) Choosing which relationship to model:

We start this step by first identifying our response and predictor variables. Since we are interested in the relationship between the credit card payments and APR at the time, in this case the payment is the response variable while the predictor variable is APR.
We can look at a scatterplot to see the association between the two variables and if the association is linear.

2) Fitting the Data

In this step, we choose the slope and intercept for the line that best summarizes the relationship of the variables. We use a technique called least square regression to find estimates of the parameters B_0 and B_1. We rely on computer for this step. The fitted model is represented by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Here $\hat{y}$ is the response variable, $\hat{B}_0$ is the intercept, $\hat{\beta}_1$ is the slope and x is predictor model.
A key tool for fitting a model is to compare the values it predicts for actual values of the response variable. This discrepancy is called the residual.

3) Assessing the fit

In this step we use different tools to check if the conditions that we initially held regarding linearity, constant variance, normality, zero mean for residuals is indeed true. We make use of residuals vs fitted values plot to examine these criteria. Likewise, we use Normal Quantile Plot to ensure normality of the residuals.

4) Using the model

Since the motivation for investigating the data was to see the association between credit card payments and APR, we can construct a confidence interval for β_1 using the standard error to see how much is the relationship between the card payments and APR.

Q no.

8b) If the diamonds were measured in grams instead of carats then all of the plots would stay the same. Since, the change in unit of measurement is a linear transformation the resulting response variable would also change in relation to the predictor variable. Therefore there would be no change in the plots.

9d) The R output which can help us determine if model is good fit for data are:

(i) $R^2 = 0.43$ which means 43% variation in the data is explained by our model.

(ii) P value of the test for $\beta_1 = 0.00225$
This means there is evidence in our data for association between speed of the run and calories burned.