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.

Estimation : Finds unknown values for subject of interest

Statistical inference: Use probability distribution of subject of interest to make probabilistic conclusion

A is estimating the average weight
B is estimating the price of a home
} They are both estimation

(6)

1. I would plot the predictor variable (payments made on a credit card) vs the response variable

2. We would use the least square line to produce a slope and intercept to model the data

$$y = \beta_0 + \beta_1 x$$

3. We would use the residual vs fits and Normal quantile plot to access the necessary condition: linearity, uniform spread, normality, zero mean, independence, randomness. If these conditions hold we can use the model. If not we need to rexpress our values perhaps by logarithm

4. We can use our model for estimates or inference assuming all conditions are met. Inference includes confidence intervals for the slope / mean, prediction intervals or Hypothesis test

8A

a) Based on these plots, what conclusions can we make about the conditions necessary for inference with a SLR model?

- In the normal probability plot we see some small skewedness in the tail
- There isn't a clear sense of uniform spread in the residual plot; some data point were closer to zero whereas some were much farthe.
- In the histogram which we can use for testing normality we see some unusual distribu indicative of an outlier

– Using each of these plot we can conclude that at least the normality and uniform spread conditions have not been met. I am also tempted to say the linearity condition as well but for now I am only sure of the normality and uniform spread condition. That said as these conditions do not hold the model ~~cannot~~ is not a good representation of the

New answer

- The normal quantile plot has a slight tail skewed but overall it shows the normality condition holds. The histogram also shows some unusual distribution indicating an outlier
- The residual vs fit was a bit difficult to analyze but it seems to show that both the linearity and uniform spread conditions are met.

- Independence and randomness: we can't really test for these conditions by looking at these graphs. These condition depend on how the experiment was designed and implemented. As of now we assume they hold true

Overall all condition necessary to the model the data have been met.

9a

a) What is the estimate for the standard deviation of the number of calories burned based on this linear model?

The estimate for the standard deviation is 30.84

The estimate of the standard deviation is 30.84 Calorie This value represent the average distance the observed values fall from the regression line.

(9c)

c) Suppose, on average, for any person within the same age group as our runner, every mph increase in running speed corresponds to 100 additional calories burnt. Describe a procedure to determine if our runner's rate of burning calories is different from this average for all people in the age group. Make sure you define any symbols you use. You do not need to actually determine the answer, just describe the statistical procedure you would use.

we could do a confidence interval to find the true average calorie burn per MPH run. If we see a CI centered around β_1 containing our observed value of 100 then we would know this value was not statistically significant and could not reject the null hypothesis. If the interval does not catch the observed 100 however, we know this difference was statistically significant.