Quiz 4

1. (Short answer) Suppose I am interested in studying whether a new training for dentists can improve patient experience when receiving a dental filling for a cavity. I select 100 dentists in Seattle and randomize half of them to receive the training, and half to receive no training. For each dentist, I randomly select 10 of their patients and survey them on the level of discomfort (on a scale of 1 to 10) experienced while receiving a filling. For my statistical analysis, I run a linear regression of patient discomfort level on whether or not the dentist performing the procedure received the training.

Based on the information above, do you believe that the independence assumption of linear regression is satisfied? Give one sentence explaining your rationale.

1. (Multiple response) Suppose I am interested in estimating the association between forced expiratory volume (FEV) (in liters per second) and log-transformed age in children. I run the following code in R:

Text

Description automatically generated

Text

Description automatically generated

Which of the following are correct interpretations of the slope coefficient corresponding to the association between forced exhalation volume (liters per second) and log-transformed age in children, in either of the two models I fit? Choose all correct answers.

1. The difference in average FEV comparing two groups of children that differ in age by 30% is 0.549 liters/second, with the older group having the higher FEV.
2. The difference in average FEV comparing two groups of children that differ in age by 1 log year is 2.09 liters/second, with the older group having the higher FEV.
3. The difference in average FEV comparing groups of children that differ in age by a multiplicative factor of 2 is 4.27 liters/second, with the older group having the higher FEV.
4. The difference in average FEV comparing groups of children that differ in age by a multiplicative factor of 2 is 1.45 liters/second, with the older group having the higher FEV
5. The difference in average FEV comparing two groups of children that differ in age by 1 log year is 8.11 liters/second, with the older group having the higher FEV.
6. (Short answer) Suppose I am interested in understanding the linear association between a person's monthly rent (in US dollars) and diastolic blood pressure. I hypothesize that people with a higher monthly rent will have higher average diastolic blood pressure. I want to be able to make statements about the association between these two variables using *multiplicative* differences, so I decide to use the following model:

My friend prefers models where the intercept interpretation is for groups of individuals who have the mean value of our predictor of interest. She suggests fitting the following model instead:

Explain to your friend why the model she suggests is actually *impossible* to use. (Hint: if we mean-center rent, some observations will have covariate values with rent – mean(rent) < 0)

1. (Short answer) Suppose I survey a random sample of 10,000 students at UW, and ask them about their pets. The format of my survey question is as follows:

Indicate which of the following pets you have. If you have more than one, please only choose the *oldest* pet that you have. For example, if you have a 6 year old dog and a 7 year old cat, select "Cat." If you have a 4 year old turtle and a 10 year old iguana, select "Other."

1. Dog
2. Cat
3. Turtle
4. Other
5. No pets

You believe that the results of your survey question should provide an *accurate* estimate of the proportion of people who have each pet (in any of the five categories) at UW because you surveyed a random sample of 10,000 students. Your friend believes that your results will not provide an accurate estimate of the proportion of people who have each pet, one of the reasons being that you are likely going to overestimate the proportion of students in the population who have turtles and cats. Who is correct about the accuracy of the results of your survey, and why?

1. (Multiple choice) In the births dataset, the variable "smoke" is recorded as 0's and 1's, where a 1 indicates that a birth parent smoked during pregnancy and a 0 indicates that a birth parent did not smoke during pregnancy. Suppose your friend creates a new variable for smoking using 1's and 2's, where a 2 indicates that birth parent smoked during pregnancy and a 1 indicates that birth parent did not smoke during pregnancy. You and your friend both fit linear regression models with birth weight as the outcome, and your respective smoking variables as predictors: you use smoke with 0's and 1's, they use smoke with 1's and 2's. Which of the following statements is true?
2. The slope estimates for both models are the same, and the intercepts are the same.
3. The slope estimates for both models are the same, but the intercepts are different. The intercept in your model is scientifically relevant, but the intercept in your friend's model is not scientifically relevant.
4. The slope estimates for both models are the same, but the intercepts are different. The intercepts in both models are scientifically relevant.
5. The slope estimates and the intercepts in both models are different.
6. (Multiple choice) Which diagnostic plot(s) are needed for the *classical* linear regression assumptions, and which diagnostic plot(s) are needed for the linear regression assumptions if we have a large sample size and use robust standard errors?
7. Classical: Scatterplot of residuals vs. fitted values, histogram of residuals.

Large sample size and robust SEs: Scatterplot of residuals vs. fitted values, histogram of residuals.

1. Classical: Scatterplot of outcomes vs. fitted values, histogram of residuals.

Large sample size and robust SEs: Scatterplot of outcomes vs. fitted values.

1. Classical: Scatterplot of residuals vs. fitted values, histogram of residuals.

Large sample size and robust SEs: Scatterplot of residuals vs. fitted values.

1. Classical: Scatterplot of residuals vs. fitted values, histogram of residuals.

Large sample size and robust SEs: None.

1. (Short answer) Suppose I survey birth parents and ask them the total number of children they have ever had, and if they have had at least one child, I also ask them the total number of those children who have died. This type of data is known as Summary Birth History data, because we only know totals for each birth parent as opposed to date of birth and date of death for each child (which is known as Full Birth History data). For individuals who have had at least one child, I compute the proportion of child deaths as the total number who died divided by the total number of children they have ever had. What type of variable is the proportion of child deaths for each birth parent who has given birth to at least one child? Explain your reasoning.
2. (Multiple choice) Suppose I fit a linear regression model,

And I obtain estimates for the intercept and slope:

Which of the following is the correct interpretation of the intercept?

* 1. We estimate that mean birth weight for birth parents who are 1 year old is 2549.6 grams.
  2. We estimate that mean birth weight for birth parents who are 0 years old is 2549.6 grams

1. (True/False) The regression model cannot satisfy the linearity assumption for linear regression because is in the model.
2. (Multiple choice) Suppose I want to test whether there is a linear association between height and weight. I write my regression model as

Which of the following is the appropriate null and alternative hypothesis for answering my statistical question?