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 choice) Suppose I am interested in the linear association between weight prior to pregnancy (in pounds) and birth weight of children (in grams). I choose to adjust for marital status (married yes/no) and years of education a birth parent has had in my model. Fitting this model in R, I obtain the following results:

> mod <- lm(data = births, bwt ~ married + education + wpre)

> summary(mod)

Call:

lm(formula = bwt ~ married + education + wpre, data = births)

Residuals:

Min 1Q Median 3Q Max

-3066.22 -300.39 28.65 340.36 1766.84

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2693.4458 77.3291 34.831 < 2e-16 \*\*\*

married 138.4739 28.6603 4.832 1.44e-06 \*\*\*

education 9.7485 4.5048 2.164 0.0306 \*

wpre 3.2328 0.3157 10.241 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 544.7 on 2496 degrees of freedom

Multiple R-squared: 0.05297, Adjusted R-squared: 0.05184

F-statistic: 46.54 on 3 and 2496 DF, p-value: < 2.2e-16

Which of the following is the correct interpretation of the coefficient that corresponds to my statistical question?

* 1. We estimate that the difference in mean birth weight between two groups of unmarried birth parents with zero years of education, differing by one pound in pre-pregnancy weight is 3.23 grams, with the higher pre-pregnancy weight group having higher birth weights
  2. We estimate that the difference in mean birth weight between two groups of birth parents with the same marital status and years of education, differing by one pound in pre-pregnancy weight is 3.23 grams, with the higher pre-pregnancy weight group having higher birth weights
  3. We estimate that the difference in mean birth weight between two groups of birth parents with the same marital status and pre-pregnancy weight, differing by one year in education is 9.75 grams, with the higher education group having higher birth weights
  4. We estimate that the difference in mean birth weight between two groups of birth parents differing by one pound in pre-pregnancy weight is 3.23 grams, with the higher pre-pregnancy weight group having higher birth weights

1. I want to study if wearing face masks causes an overall reduced risk of death. I sample people around the country. Which of the following is not a confounder?
   1. Attitudes about the importance of staying healthy
   2. Age
   3. Covid Risk
   4. State of Residence
2. (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?
   1. Classical: Scatterplot of residuals vs. fitted values, histogram of residuals.
   2. Large sample size and robust SEs: Scatterplot of residuals vs. fitted values, histogram of residuals.
   3. Classical: Scatterplot of residuals vs. fitted values, histogram of residuals.
   4. Large sample size and robust SEs: Scatterplot of residuals vs. fitted values.
   5. Classical: Scatterplot of residuals vs. fitted values, histogram of residuals.
   6. Large sample size and robust SEs: None.
3. I am studying the association between sepal length and species in iris flowers. I run a regression of sepal length (in centimeters) on the species (a categorical variable with three levels: setosa, virginica, versicolor) and see the following output:

> mod <- lm(data = iris, Sepal.Length ~ Species)

> summary(mod)

Call:

lm(formula = Sepal.Length ~ Species, data = iris)

Residuals:

Min 1Q Median 3Q Max

-1.6880 -0.3285 -0.0060 0.3120 1.3120

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.0060 0.0728 68.762 < 2e-16 \*\*\*

Speciesversicolor 0.9300 0.1030 9.033 8.77e-16 \*\*\*

Speciesvirginica 1.5820 0.1030 15.366 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5148 on 147 degrees of freedom

Multiple R-squared: 0.6187, Adjusted R-squared: 0.6135

F-statistic: 119.3 on 2 and 147 DF, p-value: < 2.2e-16

Use this output to produce an estimate of the mean difference in sepal length (in centimeters) comparing flowers of species versicolor and flowers of species setosa. Don’t worry about a confidence interval.

1. We are studying the impact of hours of physical activity per week and blood pressure. We also have data on hours worked per week. As people who work longer hours might have less time for physical activity and also have higher blood pressure due to stress, we are wondering if hours worked is a confounder. We run a simple linear regression with blood pressure as the outcome and hours worked as a predictor of interest. The association is not significant at a 0.05 level. Based on this, we do not adjust for hours worked in our final analysis.
   1. This decision makes sense
   2. This decision does not make sense
2. (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?
3. The intercept in your model **is not** scientifically relevant, but the intercept in your friend's model **is** scientifically relevant.
4. The intercept in your model **is** scientifically relevant, but the intercept in your friend's model **is not** scientifically relevant.
5. The intercepts in both models **are** scientifically relevant.
6. The intercepts in both models **are not** scientifically relevant.
7. I want to see if there is an association between the neighborhood of Seattle someone lives in and how many days in the last year they had a headache. I sample people from U District, Ballard, Beacon Hill, and Belltown. I fit the following model:

What is my null hypothesis?

* 1. At least one of is equal to 0
  2. At least one of is equal to 0

1. When we fit a multiple regression model in R, we need to specify which predictor is our predictor of interest and which, if any, of the other covariates are confounders.
2. Which of the following scientific questions is asking about effect modification?
   1. What is the association between brain atrophy score and history of stroke, adjusting for age?
   2. Is the association between brain atrophy score and history of stroke different for people of different ages?
   3. Is there an association between brain atrophy score and either history of stroke or age?