Problem Set 6

Statistics 104

Due April 16, 2020 at 11:59 pm

Problem set policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, be sure to include the code in your solution.*

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Problem 1.

This problem uses data from the Prevention of REnal and Vascular END-stage Disease (PREVEND) study, which took place between 2003 and 2006 in the Netherlands. Clinical and demographic data for 4,095 individuals are stored in the `prevend` dataset in the `oibiostat` package.

Body mass index (BMI) is a measure of body fat that is based on both height and weight. The World Health Organization and National Institutes for Health define a BMI of over 25.0 as overweight; this guideline is typically applied to adults in all age groups. However, a recent study has reported that individuals of ages 65 or older with the greatest mortality risk were those with BMI lower than 23.0, while those with BMI between 24.0 and 30.9 were at lower risk of mortality. These findings suggest that the ideal weight-for-height in older adults may not be the same as in younger adults.

Explore the relationship between BMI (`BMI`) and age (`age`), using the data in `prevend.samp`, a random subset of 500 individuals from the larger `prevend` data.

a) Create a plot that shows the association between BMI and age. Based on the plot, comment briefly on the nature of the association.

b) Fit a linear regression model to relate BMI and age.

   i. Write the equation of the linear model.

   ii. Interpret the slope and intercept values in the context of the data. Comment on whether the intercept value has any interpretive meaning in this setting.

   iii. Is it valid to use the linear model to estimate BMI for an individual who is 30 years old? Explain your answer.

   iv. According to the linear model, estimate the average BMI for an individual who is 60 years old.

   v. Based on the linear model, how much does BMI differ, on average, between an individual who is 70 years old versus an individual who is 50 years old?

c) Create residual plots to assess the model assumptions of linearity, constant variability, and normally distributed residuals. In your assessment of whether an assumption is reasonable, be sure to clearly reference and interpret relevant features of the appropriate plot.

i. Assess linearity.

ii. Assess constant variance.

iii. Assess normality of residuals.

iv. Suppose that a point is located in the uppermost right corner on a Q-Q plot of residuals (from a linear model). In one sentence, describe where that point would necessarily be located on a scatterplot of the data.

d) Conduct a formal hypothesis test of no association between BMI and age, at the $\alpha = 0.05$ significance level. Summarize your conclusions.

e) Report the R^2 of the linear model relating BMI and age. Based on the R^2 value, briefly comment on whether you think the estimated average BMI values calculated in part b) are accurate.

Problem 2.

The data file `malebirths.csv` contains data for the proportion of male births (annually) in four countries: Denmark, the Netherlands, Canada, and the United States. This problem explores the relationship between proportion of male births and time (as measured by year).

a) Create a scatterplot for each country that includes the regression line. Be sure the plots are clearly labeled and that each plot has the same bounds on both axes.

b) For each country, assess whether there is evidence of a significant association between proportion of male births and year. Report b_1, the standard error of b_1, the related t-statistic, and the related p-value for each model.

c) Based on a visual inspection of the plots, explain why the United States can have the t-statistic with the largest magnitude even though its slope does not have the largest magnitude.

d) Why might it be reasonable to expect that the standard error of the slope would be the smallest for the United States? *Hint*: Consider that each y-observation can be thought of as an average of `0`s and `1`s for all births (in a year).

Problem 3.

This problem uses data from the National Health and Nutrition Examination Survey (NHANES), a survey conducted annually by the US Centers for Disease Control (CDC). The data can be treated as if it were a simple random sample from the American population. The dataset `nhanes.samp.adult.500` in the `oibiostat` package contains data for 500 participants ages 21 years or older that were randomly sampled from the complete NHANES dataset that contains 10,000 observations.

Regular physical activity is important for maintaining a healthy weight, boosting mood, and reducing risk for diabetes, heart attack, and stroke. In this problem, you will be exploring the relationship between weight (`Weight`) and physical activity (`PhysActive`) using the data in `nhanes.samp.adult.500`. Weight is measured in kilograms. The variable `PhysActive` is coded `Yes` if the participant does moderate or vigorous-intensity sports, fitness, or recreational activities, and `No` if otherwise.

a) Explore the data.

   i. Identify how many individuals are physically active.

   ii. Create a plot that shows the association between weight and physical activity. Describe what you see.

b) Fit a linear regression model to relate weight and physical activity. Report the estimated coefficients from the model and interpret them in the context of the data.

c) Report a 95% confidence interval for the slope parameter and interpret the interval in the context of the data. Based on the interval, is there sufficient evidence at $\alpha = 0.05$ to reject the null hypothesis of no association between weight and physical activity?

d) Report and interpret an approximate 95% prediction interval for the weight of an individual who is physically active.

e) Suppose that upon seeing the results from part c), your friend claims that these data represent evidence that being physically active promotes weight loss. Do you agree with your friend? Explain your answer.

f) In the context of these data, would you prefer to conduct inference using the linear regression approach or the two-sample t-test approach? Explain your answer.

g) Suppose that the estimated slope coefficient from the model were positive (and statistically significant). Propose at least two possible explanations for such a trend.

Problem 4.

The file `low_bwt.Rdata` contains information for a random sample of 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts.

The dataset contains the following variables:

- `birthwt`: the weight of the infant at birth, measured in grams
- `gestage`: the gestational age of the infant at birth, measured in weeks
- `momage`: the mother's age at the birth of the child, measured in years
- `toxemia`: recorded as `Yes` if the mother was diagnosed with toxemia during pregnancy, and `No` otherwise
- `length`: length of the infant at birth, measured in centimeters
- `headcirc`: head circumference of the infant at birth, measured in centimeters

The condition toxemia, also known as preeclampsia, is characterized by high blood pressure and protein in urine by the 20^{th} week of pregnancy; left untreated, toxemia can be life-threatening.

a) Fit a linear model estimating the association between birth weight and toxemia status.

   i. Write the model equation.

   ii. Report a 95% confidence interval for the slope and interpret the interval.

b) Using graphical summaries, explore the relationship between birth weight and toxemia status, birth weight and gestational age, and gestational age and toxemia. Summarize your findings.

c) Fit a multiple regression model with toxemia and gestational age as predictors of birth weight.

   i. Evaluate whether the assumptions for linear regression are reasonably satisfied.

   ii. Interpret the coefficients of the model, and comment on whether the intercept has a meaningful interpretation.

   iii. Write the model equation and predict the average birth weight for an infant born to a mother diagnosed with toxemia with gestational age 31 weeks.

   iv. The simple regression model and multiple regression model disagree regarding the nature of the association between birth weight and toxemia. Briefly explain the reason behind the discrepancy. Which model do you prefer for understanding the relationship between birth weight and toxemia, and why?

Problem 5.

The National Health and Nutrition Examination Survey (NHANES) is a yearly survey conducted by the US Centers for Disease Control. This question uses the `nhanes.samp.adult.500` dataset in the `oibiostat` package, which consists of information on a subset of 500 individuals ages 21 years and older from the larger NHANES dataset.

Poverty (`Poverty`) is measured as a ratio of family income to poverty guidelines. Smaller numbers indicate more poverty, and ratios of 5 or larger were recorded as 5. Education (`Education`) is reported for individuals ages 20 years or older and indicates the highest level of education achieved: either 8th Grade, 9 - 11th Grade, High School, Some College, or College Grad. The variable `HomeOwn` records whether a participant rents or owns their home; the levels of the variable are `Own`, `Rent`, and `Other`.

a) Create a plot showing the association between poverty and educational level. Describe what you see.

b) Fit a linear model to predict poverty from educational level.

   i. Interpret the model coefficients and associated p-values.

   ii. Assess whether educational level, overall, is associated with poverty. Be sure to include any relevant numerical evidence as part of your answer.

c) Create a plot showing the association between poverty and home ownership. Based on what you see, speculate briefly about the home ownership status of individuals who responded with `Other`.

d) Fit a linear model to predict poverty from educational level and home ownership. Comment on whether this model is an improvement from the model in part b).

Problem 6.

A survey of Harvard students was conducted to measure a number of variables of interest. The file `exercise.csv` contains the self-reported number of hours of exercise per week for $n = 225$ students split across the four class years, along with a number of binary variables, where `1` indicates "Yes" and `0` indicates "No".

In this problem, you will work with the following 10 predictor variables: class year (`classyear`), sex (`sex`), concentration (`conc`), hair color (`hair`), vegetarian (`vegetarian`), wears glasses or contacts (`glasses`), on an athletic team (`athlete`), regularly drinks coffee (`coffee`), height (`height`), and exercise hours per week (`exercise`).

The dataset has been randomly split into "halves", with `exercise_half1.csv` containing data for 112 students and `exercise_half2.csv` containing data for 113 students.

a) Using `exercise_half1.csv`, fit a regression model to predict resting heart rate (`heartrate`) in beats per minute (bpm) from the 10 predictors listed above. Name this `model1`. Which slope coefficients are significant at the $\alpha = 0.10$ signifcance level?

b) It may be the case that a predictor that shows up as significant from a regression model is not actually associated with the response on the population level; i.e., represents an instance of Type I error. For each of the significant predictors from part a), briefly explain whether you think the a model fit using the second half of the dataset will also demonstrate that the predictor is significantly associated with heart rate.

c) Using `exercise_half2.csv`, fit a regression model to predict resting heart rate (`heartrate`) in beats per minute (bpm) from the 10 predictors listed above. Name this `model2`. Are any of the slope coefficients identified as significant from Model 1 are also significant in Model 2? If so, which one(s)?

d) The code shown in the template creates a plot to visualize the results of the two models. It plots a horizontal line connecting the t-statistics for each coefficient in the two models and compares it to the critical t-value and 0 (plotted as dashed vertical lines). You are not expected to know how to recreate a similar plot.

   i. Briefly describe the most interesting features of the plot.

   ii. Explain what might have caused the differences between Model 1 and Model 2; i.e., why were some predictors significant in one model and not the other?

Problem 7.

Do men and women think differently about their body weight? To address this question, you will be using data from the Behavioral Risk Factor Surveillance System (BRFSS).

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States collected by the Centers for Disease Control and Prevention (CDC). As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about diet and weekly physical activity, HIV/AIDS status, possible tobacco use, and level of healthcare coverage.

The `cdc.sample` dataset contains data on 500 individuals from a random sample of 20,000 respondents to the BRFSS survey conducted in 2000, on the following nine variables:

- `genhlth`: general health status, with categories `excellent`, `very good`, `good`, `fair`, and `poor`
- `exerany`: recorded as `1` if the respondent exercised in the past month and `0` otherwise
- `hlthplan`: recorded as `1` if the respondent has some form of health coverage and `0` otherwise
- `smoke100`: recorded as `1` if the respondent has smoked at least 100 cigarettes in their entire life and `0` otherwise
- `height`: height in inches
- `weight`: weight in pounds
- `wt.desire`: desired weight in pounds
- `age`: age in years
- `gender`: gender, recorded as `m` for male and `f` for female

a) Create a variable called `wt.discr` that is a measure of the discrepancy between an individual's desired weight and their actual weight, expressed as a proportion of their actual weight:

$$\text{weight discrepancy} = \frac{\text{actual weight} - \text{desired weight}}{\text{actual weight}}$$

b) Fit a linear model to predict weight discrepancy from age and gender. Interpret the slope coefficients in the model.

c) Investigate whether the association between weight discrepancy and age is different for males versus females.

   i. Fit a linear model to predict weight discrepancy from age, gender, and the interaction between age and gender. Write the model equation.

   ii. Write the prediction equation for males and the prediction equation for females.

   iii. Is there statistically significant evidence of an interaction between age and gender? Explain your answer.

d) Comment on whether the results from part c) suggest that men and women think differently about their body weight. Do you find the results surprising; why or why not? Limit your response to at most five sentences.

Problem 8.

In chronic diseases such as multiple sclerosis, medications are approved for sale after the US Food and Drug Administration evaluates both the effectiveness of the medication in delaying the progression of the disease and the extent to which the medication provides symptom relief. An earlier study showed that two treatments (labeled A and B) were equally effective in delaying the progression of disease. This problem examines simulated data from a study comparing how well the treatments relieve disease symptoms.

A total of 410 participants were randomly assigned to one of the two treatments. The effect of disease symptoms on daily living was assessed by responses to the EuroQol 5D-5L (EQ-5D-5L), a questionnaire widely used to study health-related quality of life; higher scores are indicative of better quality of life. The EQ-5D-5L was administered to each participant before the start of treatment and again after 6 months of treatment. Investigators planned to compare the mean change in the EQ-5D-5L score between the two groups in order to assess whether the extent of symptom relief differs by treatment.

The data from the study are contained in the file `quality_of_life.Rdata`. The dataset `quality.of.life` contains the following variables:

- `treatment.group`: The treatment to which the participant was assigned, coded `A` or `B`.
- `pre.treatment.score`: The EQ-5D-5L score just before treatment started.
- `post.treatment.score`: The EQ-5D-5L score after 6 months of treatment.

Do the data from this study provide definitive evidence that the two treatments are different in their ability to offer symptom relief?

a) Outline a plan for analyzing the data. Specify how the variables in the dataset will be used, in addition to any variables that might be created. Clearly indicate which methods will be used in the analysis. If a hypothesis test will be conducted, state the null and alternative hypotheses and use significance level $\alpha = 0.05$.

b) Conduct the analysis and summarize the numerical results. Be sure to provide an answer to the original question in the context of the data.

c) The sponsors of treatment A decide to do a second study comparing A and B, but to save money, they decide to administer the EQ-5D-5L only once, 6 months after the start of the therapy. Earlier studies have shown that the population average post-treatment score on A is 64.3, with standard deviation 8.5. The sponsors believe that treatment B will have a population average post-treatment score of 63 and also have standard deviation 8.5. How large should this second study be to have power 0.80 of rejecting the null hypothesis of no treatment difference between A and B, when $\alpha = 0.05$?

d) When the investigators behind the first study learn about the design of the second study, they send a message to the study team: “You may be saving some money by not administering the questionnaire before treatment starts, but your study will be much larger than if you were to use our design.” What justifies that message? Provide a specific explanation.