Biostatistics 515/518 Winter 2020
Homework 2 (7 problems)

Questions 1-3 investigate whether the distribution of cerebral atrophy scores differs by age in the population of older patients from which the MRI data were sampled.

1. Provide suitable descriptive statistics related to the distribution of atrophy scores by age.
2. Perform a linear regression analysis to investigate whether there is a trend in mean atrophy scores across age groups.
    a. What is an interpretation of the intercept from the regression model you fit? Provide statistical inference pertaining to the hypothesis that the true intercept is 0. Do you think this inference is reasonable? Why or why not?
    b. Interpret the slope from the regression model, along with statistical inference pertaining to the hypothesis that the true slope is 0. Do you think this inference is reasonable? Why or why not?
    c. Using the fitted regression model from your analysis, estimate the mean atrophy for 70-year-olds and for 90-year-olds. Do you think these are reasonable estimates? Why or why not?

3. Perform a regression analysis to investigate whether there is a trend in the geometric mean of cerebral atrophy scores across age groups.
    a. What is an interpretation of the intercept from the regression model you fit? Provide statistical inference pertaining to the hypothesis that the true intercept is 0. Do you think this inference is reasonable? Why or why not?
    b. Interpret the slope from the regression model, along with statistical inference pertaining to the hypothesis that the true slope is 0. Do you think this inference is reasonable? Why or why not?
    c. Using the fitted regression model from your analysis, estimate the geometric mean atrophy for 70-year-olds and for 90-year-olds. Do you think these are reasonable estimates? Why or why not?

4. We like to have variables measured in units that are sensible for the context of an investigation and the interpretation of results. Consider performing a regression analysis in the following two contexts. What units would you prefer for the predictor?

a. Response: DSST; predictor: Age; population: Americans over 65.
    Measure age in years or in months? (Explain your choice.)
b. Response: Weight; predictor: Age; population: Infants 2 years old and younger
    Measure age in years or in months? (Explain your choice.)

5. Often a regression predictor is a variable that is familiar to those seeing the analysis. For example, if the predictor is cholesterol levels then cardiologists will have familiarity with the typical range of cholesterol levels, which will help them interpret the results. Sometimes a predictor in a regression model will be a variable for which the audience has little familiarity. One option for such variables is to scale the variable by the sample

SD. Briefly describe how regression output would be summarized in this situation and why this is useful.

Problems 6 and 7 use the FEV dataset, which can be found on the class website.

6. Create a binary smoking indicator coded as 1 for kids who smoke and 0 for kids who do not smoke. For the response variable FEV, perform four statistical analyses:

- two-sample t-test assuming equal variances comparing FEV for smokers and non-smokers
- two-sample t-test not assuming equal variances comparing FEV for smokers and non-smokers
- linear regression of FEV on the smoking indicator using model-based standard error estimates.
- linear regression of FEV on the smoking indicator that indicates smoking status using robust standard error estimates.

For the four analyses you performed, make a table summarizing their results for estimating the following quantities:

(a) Point estimate of mean FEV among non-smokers
(b) Point estimate of mean FEV among smokers
(c) Point estimate of difference in mean FEV – smokers compared to non-smokers
(d) Estimated standard error for difference in mean FEV (smokers compared to non-smokers)
(e) 95% confidence interval for difference in mean FEV (smokers compared to non-smokers)
(f) Wald p-value for test of the null hypothesis that smoking kids and non-smoking kids have the same mean FEV

(Organize your table in a way that you think best presents the results.)

For each of (a)-(f) comment on results that are the same or different. If results are different, qualify whether they are slightly different or substantially different.

For item (d), comment on the nature of any large differences you see.

7. The general scientific question is whether the deleterious effects of smoking can be seen in children who smoke. We use the FEV data to investigate whether there is an association between lung capacity and smoking in children. Fit a simple regression model where smoking status is the predictor of interest and FEV (a measure of lung volume) is the response. Interpret the results. Explain why this analysis might be unsatisfactory for addressing the scientific question.