Biostatistics 514/518 Winter 2020
Homework 1 (3 problems)

1. Read the brief article from "news@nature.com" and answer the following questions.
   a) Is the study an observational study or a clinical trial?
   b) Critique the article's presentation of the study results.

On this (and all) homework assignments, code and raw software output is unacceptable. Instead, use software output to (thoughtfully) prepare tables or other summaries of results. Tables should be appropriate for inclusion in a scientific report, with appropriate rounding. (Most students report too many digits in tables and should round numbers more – think about whether extra digits add useful information.)

2. A law of physics applies to the length of a string hanging from a fixed point with a mass of weight x hanging from the end. As long as the weight x does not break the string, the law saws that the length of the string is given by

$$Y=\alpha + \beta x+ \varepsilon$$

The independent random error ε is known empirically to be Normally distributed with variance σ^2 (the same variance for any weight x).

a) What is the interpretation of α?
b) What is a sensible interpretation for the parameter β? Think of a word in English that captures what β measures. Do not answer "slope."
c) Would you expect β to be larger when the "string" is a broken rubber band or when the "string" is the rope from a flagpole? Assume both strings are measured in the same units (let's say inches).

- Suppose a study is done with one string. The type of string is known to break if more than about 100 ounces is suspended. The study is done by choosing a set of weights ranging from 10 to 50 ounces and measuring the string after hanging the weight from the string. The linear model is fit to the data using classical regression procedures, producing estimates for α and β.

d) Comment on the reasonableness of using the fitted model to estimate the length of the string with a 30 ounce weight attached.
e) Comment on the reasonableness of using the fitted model to estimate the length of the string with an 80 ounce weight attached.
f) Comment on the reasonableness of using the fitted model to estimate the length of the string with a 120 ounce weight attached.
g) Comment on the reasonableness of using the fitted model to estimate $E(Y|x=-10)$.
h) Suppose one of the weights used in the study weighed 20 ounces, so there is one measurement for the length of the string with a 20 ounce weight. Assuming all

the premises of this exercise are all true, which do you prefer as your estimate of the length of the string with a 20 ounce weight (and why): the experimental value with the 20 ounce weight, or the quantity $\hat{\alpha} + 20\hat{\beta}$?

i) Download the "string" data from the class website and estimate the length of the string and its parameter β. If reasonable, give confidence intervals for these parameters, making their interpretations clear. If reasonable, give any important p-values, making clear what is being tested and interpreting the results. Do not perform a task if you believe it is invalid or lacks any plausible interest or relevance, and instead explain why it is invalid or irrelevant.

j) Using software, estimate the error variance σ^2. Estimate the error standard deviation σ.

k) Verify the equality (3.12) (page 43) using these data.

3.

The class website contains the 'WCGS' dataset that was collected to identify risk factors for coronary heart disease (CHD).

In 1960-1961, 3,154 healthy, middle-aged men were entered into the Western Collaborative Group Study, a long-term study of coronary heart disease (CHD). The risk of coronary heart disease mortality was studied for several variables measured at baseline, i.e., Type A/B behavior, systolic blood pressure, serum cholesterol level, cigarette smoking status, and age. Unfortunately, there is no codebook for the dataset. Fortunately, the variable descriptions in the dataset are largely informative.

Consider the association between height and weight among middle-aged men.

a) To investigate the association between height and weight, Chris decides to perform a two-sample t-test to compare the weights for every height in the dataset (heights are given in integer inches). How many statistical tests will Chris perform with this approach? If Chris applies the Bonferroni correction to significance level 0.05, what significance threshold will he use for each individual t-test?

b) Angela decides to dichotomize the heights into two groups and perform a two-sample t-test comparing weights on the dichotomized heights. She considers every possible cutpoint, and identifies the cutpoint that gives her the most significant results (smallest p-value). She reports the results of this t-test and p-value. Describes the merits or issues you see in Angela's approach.

c) Perform a simple linear regression of weight on height using the wcgs data. Using 1-3 carefully written sentences, summarize results in language suitable for a scientific publication.

Make a scatterplot of weight on the vertical axis and height on the horizontal axis. If any additional figures are helpful to answer the following questions, make and present those figures also.

d) Based on these limited data, do you believe there is a linear relationship between weight and height in this population? Why or why not? What implication does this have for your summary in item (c)?

e) Do you think the homoscedasticity assumption of classical linear regression is reasonable, or do you see evidence of heteroscedasticity?

f) Comment on the reasonableness of using the fitted linear regression to estimate the mean weight of men from this population who are 70 inches tall.

g) Comment on the reasonableness of using the fitted linear regression to estimate the mean weight of men from this population who are 82 inches tall. (For reference, a taller-than-average professional basketball player might be 82 inches tall.)

h) Comment on the reasonableness of using the fitted linear regression to estimate the mean weight of boys who are 42 inches tall.

i) Convert height from inches to centimeters and weight from pounds to kilograms and fit a new simple linear regression model. Compare/contrast the p-values for the regression parameters to the first regression model and comment.