MA684 Midterm Project

Spring 2015

This is an individual project - please do your own work. Some discussion with other students around computer programming or clarification of the questions is allowed, but you should plan your analyses on your own and you should write up your results on your own. More substantial questions about the project should be addressed to the instructor. This project makes up a substantial proportion of your grade, so please provide an organized and edited write-up. Please write up your responses in paragraph form (on separate pages, not on the computer output). **Remember this is a statistics class – please provide statistical support for your responses and quantify your answers where appropriate (for example, if you are asked to describe an association between two variables, simply responding that there is a positive association is not a complete answer – you may want to report a correlation, give a t-statistic, and report a p-value as well as providing an interpretation).** Please attach computer output as an appendix, although ideally your write-up will be complete enough so we don’t have to refer to the computer output.

Question 1. (General context taken from a problem in Practical Statistics by Sincich, Levine, and Stephan) A forester is interested in measuring heights of relatively young trees in a couple of different environmental areas. Measuring heights of trees is difficult (they’re tall), and we would like to use the diameter of the tree (at chest height) to estimate the height of the tree. Data from n=35 spruce trees are included in the ‘SpruceTrees’ data set posted as a .csv file with this project. Variables in the data set are: 1) an id number, from 1 to 35; 2) the diameter of the tree, in inches; 3) the height of the tree, in feet. We will use regression to develop a formula predicting height from diameter.

1A (6 points). Find the scatter plot showing the association between height and diameter for this sample of trees. Given this scatter plot, do you think these data meet the assumptions of linear regression? As part of your response, list the assumptions of linear regression.

1B (6 points). Find the regression formula to predict the height of a tree from its diameter, reporting the slope and intercept, standard errors, t-statistics, and p-values from the regression (it’s OK to cut-and-paste the table from R into your answer).

Is there a significant association between the diameter and height? Explain.

How well does the diameter predict the height of a tree?

1C (6 points). Give an interpretation of the slope from this regression line. Find the 95% confidence interval for this slope. Explain the relationship between this confidence interval and the p-value for the slope.

1D (6 points). Based on this regression, what is the predicted height for a tree with a diameter of 10 inches? For a tree with a diameter of 12 inches? How close do you expect the true height of the tree to be to this estimate – that is, find and interpret an appropriate interval estimate for the height of a tree with a diameter of 10 inches, and for a tree with a diameter of 12 inches.

2. (Based loosely on a report in the Wall Street Journal ‘Exercise boosts brain power’, 2011). From the article: ‘As people age, the hippocampus, the brain’s memory center, loses 1% to 2% of its volume annually (on average, individual volumes may increase or decrease over time), affecting memory and possibly increasing the risk for dementia. A growing body of evidence has pointed to aerobic exercise as a low-cost hedge against neurocognitive decline.’

In this study, 300 healthy elderly adults (ages 55 to 85) were recruited and randomized to one of three groups (n=100 per group). The Control group agreed to be evaluated, but were not assigned to any intervention program. The Walking group (aerobic exercise) walked three days a week for 40 minutes a day. The yoga group (yoga and toning exercises, which are non-aerobic exercise) participated in group yoga sessions 3 days a week. Magnetic resonance imaging (MRI) was used to measure the volume of the hippocampus at study baseline and then again after 1 year. The dependent variable for this study is the percent change in hippocampus volume, where positive change values indicate an increase in hippocampus volume (e.g., 1.3 indicates a 1.3% increase in volume) and negative change values indicate a reduction in hippocampus volume (e.g., -1.7 indicates a 1.7% decrease in volume). Our primary study question is whether those who exercised had less decline in hippocampus volume than those in the control group.

Data for this study are saved in the attached ‘elders\_2015’ .csv file. Variables in the data set are: 1) studyid, an id number ranging from 1 to 300; 2) exercisegroup, coded 1 for those in the Control group, 2 for those in the Walking group, and 3 for those in the Yoga group; 3) age, in years, restricted to adults between the ages of 55 and 85; 4) sexf, coded 1 for females and 0 for males; 5) IQ, measured at the start of the study, as a general measure of cognitive ability, the mean IQ is expected to be around 100; and 6) hippochange, the percent change in the hippocampus volume, which should range roughly between -3 percent (indicating a 3% decrease in volume) and +2 percent (indicating a 2% increase in volume).

2A (6 points). As a description of the study sample, complete the following table:

Table 1A. Description of the study sample

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Mean | Standard Deviation | Minimum | Maximum |
| Age  IQ  Hippo Change |  |  |  |  |

Table 1B. Description of the study sample

|  |  |  |
| --- | --- | --- |
| Variable | N | % |
| Sex  Male  Female |  |  |

2B (6 points). As a preliminary analysis, find and interpret the correlations between hippochange (the change in hippocampus volume) and age, and between hippochange and IQ. Based on these correlations, are age and/or IQ significantly associated with hippochange?

Our primary goal is to determine whether there is an association between exercise and hippochange, but first we will look at a multiple regression predicting hippochange from age, sex, and IQ (this model should not include exercise group). Questions 2C – 2E ask about this preliminary model.

2C (10 points). Complete the following tables summarizing the results of this multiple regression (Hippochange has a small scale, please report 3 decimal places for slopes and standard errors throughout Question 2). It is OK to cut-and-paste R output for these tables, if the R tables give the requested information. For Table 2b, there should be 1 row in the table for the Overall Model, and One row for the Residual.

Table 2a. Multiple regression predicting percent change in hippocampus volume

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Slope | SE of  Slope | p-value |
| Intercept  Age  Sex Female  IQ |  |  |  |

Table 2b. The ANOVA for the regression predicting change in hippocampus volume

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | df | Sum of Squares | Mean  Squares | F | p-value |
| Model  Residual |  |  |  |  |  |
| Total |  |  |  |  |  |

Report and interpret the R2 for this regression model. Report the F-statistic and p-value from the ANOVA table for the regression. What can you conclude from this p-value?

2D (6 points). What can you say about the associations between age, sex, and IQ and change in hippocampus volume, based on this regression? Explain.

2E (6 points). Find and interpret the standardized slopes from this regression. Include in your response an explanation of what the standardized slopes from a multiple regression measure.

Our primary goal is to examine the association between exercise and hippochange, controlling for age, sex, and IQ. We are specifically interested in whether those in the walking group (aerobic exercise) or the yoga group (non-aerobic exercise) had more positive (or less negative) hippochange than those in the control group. Run a multiple regression predicting change in hippocampus volume from age, sex, IQ, and exercise group, and use this regression model to answer questions 2F – 2 H.

2F (10 points). Provide a Table 3, similar to Table 2a from Question 2C, reporting slopes, standard errors, and p-values from this regression (it’s OK to cut-and-paste a table from R). Report and interpret the R2 from this regression. Which of these independent variables are significantly related to hippochange in this analysis?

2G (6 points). Our primary interest in this analysis is in whether or not either Walking exercise or Yoga exercise has a positive benefit on change in hippocampus volume, compared to the No Exercise group. Interpret the results of this multiple regression analysis with a focus on this question - include as part of your answer an interpretation of slopes relating to exercise, and confidence intervals for these slopes.

2H (7 points). As another way of describing the effect of exercise, controlling for age, sex, and IQ, give a (Type II) partial R2 for the addition of exercise group, after controlling for age, sex, and IQ. Also give and interpret the partial F test for exercise group, after controlling for age, sex, and IQ (report the F statistic, degrees freedom, and p-value from this test).

2I (7 points). As a final analysis, one investigator wants to look at the interaction between age and exercise group. Run a multiple regression predicting hippochange from age, sex, IQ, exercise group, and the interaction between exercise group and age. Give a Table 4, similar to Tables 2a and 3, summarizing this analysis (it’s OK to cut-and-paste from R). What is being tested by the interaction between exercise group and age? What can you conclude about the effects of exercise from this analysis?