MA684 Take Home Final Project 2015

This is an individual project – it is OK to talk with other students about the project, but you should perform the analyses on your own, and write up your results on your own. Questions about the content of the project or around computer programming should be addressed to Prof. Heeren. Please write up your results in paragraph form – do not simply annotate computer output. This is a statistics class, so please present appropriate statistical detail – identify the statistical methods that you use, explain how you reach your conclusions, report test statistics along with p-values to make it clear what information is being reported. Please report your results in the context of the problem. If your write-up is complete, we should not need to refer to your computer output, but please submit your computer output as an appendix with the project.

1. Based on data from Jeremy Albright, Univ. of Indiana. A survey was conducted of registered voters across the US, asking about their political and moral values. We will focus on a series of 6 questions gauging opinion about 6 contentious topics that were meant to measure whether respondents had a more liberal or conservative view:

1) Should private ownership of services be increased?

2) Should gay marriage be legal?

3) Should abortion be legal?

4) Should the government take more responsibility to see that all people are provided for?

5) Does competition bring out the best in businesses and people?

6) Should assisted suicide be legal?

All of these questions were answered on a 0 to 10 point scale, with 0 indicating ‘completely disagree’ and 10 indicating ‘completely agree’. Our ultimate research question is to explore differences in voter attitudes across different regions of the country.

Data on 230 respondents, from two different regions of the country, are included in the accompanying ‘VoterValues2015’ data sets. The 8 variables in the data set are:

1) a StudyId number, ranging from 1 to 230,

2) through 7) responses to the 6 questions above, in the order listed, and

8) region, coded as 1 for respondents in Region 1, and 2 for respondents in Region 2.

1A. We wish to summarize the information from these 6 questions. Carry out a principal components analysis with a varimax rotation, on these 6 questions.

- How many components are needed to summarize the information in these 6 questions? How   
 did you determine the number of components?

- How well do these components capture the information in these 6 questions?

- Give an interpretation of these components.

- Are any of the 6 original questions poorly represented by these components?

1B. Based on the results of the principal components analysis in 1A, create sub-scales summarizing the information from these 6 questions by summing responses to sub-sets of these 6 questions. Describe the internal consistency of these scales using Cronbach’s alpha.

1C. Using these sub-scales created in 1B, compare the political viewpoints of voters from the two regions represented in the sample (report the statistical method used to compare voters from the two regions, and presenting results in a table would be nice). Is one region more conservative or liberal than the other? Explain.

Question 2. (Based on an example in Sullivan, Introductory Biostatistics, and Pastor and Reuban, Racial and Ethnic Differences in ADHD and LD in Young School-Age Children, Public Health Reports). A study examined the association between lead exposure and ADHD in a group of inner-city children between the ages of 6 and 11. Hypothetical data on 500 children are in the attached ‘LeadStudy’ file. Variables in the data set are:

1) kidid, an id number ranging from 1 to 500,

2) age, in years, ranging from 6 to 11,

3) sexf, coded 1 for females and 0 for males,

4) race, a categorical variable coded 1 for Whites, 2 for Blacks, 3 for Hispanics, and 4 for Asians,

5) lead, coded 1 for those with high blood lead levels and 0 for those with low blood lead levels,

6) ADHD, coded 1 for those with Attention Deficit Hyperactivity Disorder, and 0 for those without ADHD

7) iq, a composite IQ score, expected to have a mean of 100 and a standard deviation of 15 in the general population.

Our primary research questions are whether children with high lead levels are 1) more likely to have ADHD than children with low lead levels, and 2) have lower IQ than children with low lead levels.

Question 2A. To describe the children in the study, complete the following table:

Description of the study sample, by lead exposure

|  |  |  |  |
| --- | --- | --- | --- |
|  | Low Lead Levels  (n= ???) | High Lead Levels  (n=???) | p-value\* |
| Age (mean + sd)  Sex (n, %)  Male  Female  Race (n, %)  White  Black  Hispanic  Asian |  |  |  |

For the categorical characteristics in the table, please give the number of children in each category (e.g., the number of males in the low lead group) and also the percent of children in each category (e.g., the percent of the low lead group that is male). Also please give a footnote to the table explaining the statistical procedure you used to find the p-value reported in the table.

Are there any differences between those with high vs. low lead levels that might effect the comparison of these two groups on ADHD or IQ?

Question 2B. As a preliminary look at our two outcome measures, find

1) the percent of children with ADHD in this sample, and give a 95% confidence interval for this percentage, and

2) the mean IQ for children in this sample, and give a 95% confidence interval for this mean.

Question 2C. First, carry out a series of analyses looking at the association between high lead levels and IQ (Questions 2C – E). As an unadjusted analysis, find the mean (and standard deviation) of IQ for those with high lead levels, and for those with low lead levels (presenting results in a table would be nice). Find a p-value comparing these means (and report the statistical method you used to find this p-value). Based on this analysis, do children with high lead levels have lower IQ, on average?

Question 2D. We are concerned that other factors (age, sex, race) might be associated with IQ, and so we want to control for these variables when comparing those with high vs. low lead levels on IQ. For our primary analysis, conduct an analysis modeling the association between IQ and age, sex, race, and lead exposure (and report the statistical method you used for this analysis). Based on this model:

- Provide a table summarizing the results of this analysis, with a focus on the associations between the predictors in the model and IQ.

- How well does this model predict IQ?

- Based on this model, describe the association between lead exposure and IQ, including both a description of the association and a statement about significance.

- What other variables in the model are significantly associated with IQ? Describe these significant associations.

Question 2E. Using the results of your analysis in 2D (including all variables, regardless of significance) calculate the predicted IQ for:

- a white, 10 year old male without lead exposure, and

- a white, 10 year old male with lead exposure.

Question 2F. These next questions (2F – H) focus on analyses looking at the association between high lead levels and ADHD. As an unadjusted analysis, find the percent with ADHD for those with high lead levels, and for those with low lead levels. Find a p-value comparing these percentages (summarizing results in a table would be nice, and report the statistical method you used for this analysis). Is the percent of children with ADHD higher for those with high lead levels compared to those with low lead levels?

Question 2G. We are concerned that other factors (age, sex, race) might be associated both with ADHD, and so want to control for these factors when comparing those with high vs. low lead levels. Conduct an analysis modeling the association between ADHD and age, sex, race, and lead exposure (and report the statistical method used). Based on this model:

- Provide a table summarizing the results of this analysis, with a focus on the associations between the predictors in the model and ADHD.

- How well does this model predict ADHD?

- Describe the association between lead exposure and ADHD, including both a description of the association and a statement about significance.

- What other variables in the model are significantly associated with ADHD? Describe these significant associations.

Question 2H. Using the results of your analysis in 2C (including all variables, regardless of significance) calculate the predicted probability of ADHD for:

- a white, 10 year old male without lead exposure, and

- a white, 10 year old male with lead exposure.