MA684 Take Home Final Project 2015
Jiayuan Shi

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?

Solution: Using the 'eigenvalue greater than 1.0' rule, two components are needed to summarize these questions.

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

Solution: (2.01+1.51)/6 = 58.67%
Based on the cumulative proportion of variance explained from the eigenvalues table, 58.67% of the information in the original set of 6 questions is captured by these two components.

- Give an interpretation of these components.

Solution: The Rotated Component Matrix,

	PC1	PC2	h2	u2	com
PrivOwn	-0.06	0.82	0.67	0.33	1
GayMarriage	0.80	-0.02	0.64	0.36	1
Abortion	0.87	-0.05	0.76	0.24	1

GovResp 0.05 0.35 0.12 0.88 1
Compete 0.02 0.84 0.70 0.30 1
AssitSuicide 0.78 0.12 0.61 0.39 1

gives correlations between each initial variable and the rotated components. Items loading on each of the two components, and a name describing the common feature of these items:

Items	Name
GayMarriage, Abortion, AssitSuicide	Moral Values
PrivOwn, Compete	Political Values

Specifically, Components 1 Moral Values is related to:
- Should gay marriage be legal?
- Should abortion be legal?
- Should assisted suicide be legal?

Components 2 Political Values is related to:
- Should private ownership of services be increased?
- Does competition bring out the best in businesses and people?

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

Solution: The question GovResp (Should the government take more responsibility to see that all people are provided for?) is poorly represented by these components with loadings for both the components are less than 0.5. In addition, the 'h2' heading gives the communalities for each of the original variables, which describe proportion of variance of each initial variable captured by retained factors. These components capture only 12% of the variability in the question GovResp.

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.

Solution: Using "alpha()"command from "psych" package, I get the Cronbach's alpha for Moral Values scale is 0.7524, and for Political Values scale is 0.6133. This shows that the internal consistency of Moral Values scale is acceptable, suggesting that the items have relatively high internal consistency. However, the internal consistency of Political Values scale is questionable, and the items have relatively low internal consistency.

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.

Solution: To compare the political viewpoints of voters from Region 1 and Region 2, I get the subscales for the two regions respectively, and then conduct a two-sample t-tests in R. Given the p-values for the test statistic, the one comparing Political Values (p-value = 0.8664 > 0.05) is not significant. Therefore, mean of the subscale Political

Values in region 1 is similar as the mean of Political Values in region 2, so we cannot conclude that one region more conservative or liberal than the other from political viewpoints.

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:

Solution:
Description of the study sample, by lead exposure

	Low Lead Levels (n=373)	High Lead Levels (n=127)	p-value*
Age (mean ± sd)	8.5469 ± 1.4184	8.5511 ± 1.5872	0.9774
Sex (n, %)			0.7675
Male	176, 47.18%	58, 45.67%	
Female	197, 52.82%	69, 54.33%	
Race (n, %)			0.1268
White	198, 53.08%	54, 42.52%	
Black	65, 17.43%	26, 20.47%	
Hispanic	50, 13.40%	26, 20.47%	
Asian	60, 16.09%	21, 16.54%	

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?

Solution: For age, I conducted a t-test comparing the age in the high lead and the low lead groups. For sex and race, I conducted two chi-square tests.

For p-value of age, I ran a t test to check the association between age and lead level. The p-value is 0.9774, which means that the difference between means of ages in low and high lead level is not significant.

For p-value of gender, I ran a chi square test to check the association between gender and lead level. The chi square statistic is 0.0874 with 1 degree of freedom. The p-value for chi square statistic is 0.7675. So there is no significant effect from gender on the lead level.

For p-value of race, I also ran a chi square test to check the association between race and lead level. The chi square statistic is 5.7059 with 3 degree of freedom. The p-value for the chi square statistic is 0.1268. So there is no significant effect from race on the lead level.

No, there are not any differences between those with high vs. low lead levels that might effect the comparison of these two groups on ADHD or IQ, because given the above p-values, the associations between age and lead levels, sex and lead levels, and race and lead levels are all not significant. They will not affect the influence of lead levels 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.

Solution:
1) Through the "table()" command in R, we can find that there are 398 children without ADHD, and 102 children with ADHD. Therefore, the percent of children with ADHD in this 500 children sample is 102/500, which is 20.4%.
A 95% confidence interval for this percentage is

$$\hat{p} \pm z_{.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .204 \pm 1.96\sqrt{\frac{.204(1-.204)}{500}} = (0.169, 0.239)$$

2) Through the "mean()" command in R, we can find the mean IQ for children in this sample is 99.172.
A 95% confidence interval for this mean is

$$\bar{x} \pm z_{.025}\mathrm{s} = 99.172 \pm 1.96 * 14.70316 = (70.35, 127.99)$$

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?

Solution:
To compare the means of IQ, I'll do a two-sample t-test. I'll use the equal variance version of the two-sample t-test, since the standard deviations for weight loss are quite similar in the two groups.

The null hypothesis for this test is that the two population means are equal, and I'll do a two-tailed test.

The test statistic is

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{97.1811 - 99.8498}{14.6718\sqrt{1/127 + 1/373}} = -1.7704$$

where $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(127-1)13.977^2+(373-1)14.8998^2}{127+373-2}} = 14.6718$

	High Lead Levels			Low Lead Levels		
	Mean	SD	n1	Mean	SD	n2
IQ	97.1811	13.9770	127	99.8498	14.8998	373

with n1+n2-2 =498 degrees freedom.
The two-tailed p-value for this test statistic (I used Excel, '=t.dist.2t(1.7704, 498)') is p=0.0773.

Since p is more than the conventional 0.05, we fail to reject the null. It is a non-significant difference between the means of IQ for those with high lead levels and for those with low lead levels. Based on this analysis, children with high lead levels and children with low lead levels have similar 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.

Solution: Including the 'relevel(factor(race),'1'))' command in the regression to set the white people as a reference race variable, so that I conduct a regression in R to predict IQ from age, sex, race, and lead exposure.
Table: Multiple regression predicting IQ

Variable	Slope	SE of Slope	p-value

Intercept	96.1188	3.9506	<2e-16
Age	0.3791	0.4498	0.3998
Sex Female	-0.1942	1.3091	0.8821
relevel(factor(race), ref = "1")2	2.0169	1.7909	0.2606
relevel(factor(race), ref = "1")3	-3.2688	1.9169	0.0888
relevel(factor(race), ref = "1")4	4.2298	1.8655	0.0238
lead	-2.5169	1.5039	0.0949

Interpretation:

- Controlling for other variables, with every unit increase in age, on average, children's IQ is expected to increase by 0.38. Given the p-value -.3998, this change is not significant.
- Controlling for other variables, females are 0.19 points lower in IQ than males on average. Given the p-value 0.88, this difference is not significant.
- Controlling for other variables, in all four races, only the difference between Asian and White is significant (p-value < 0.05). The average IQ of Asian children is 4.23 points higher than White children, and this difference is significant.
- Controlling for other variables, children with high lead levels are 2.52 points lower in IQ than children with low lead levels. Given the p-value, this difference is not significant.

- *How well does this model predict IQ?*

Solution: R^2 is 0.03197, which means only 3.197% of the variability in IQ can be explained by the model. The p-value for the F-statistic is 0.0133, which means that at least one of the independent variables in this model is significantly associated with children's 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.*

Solution: The p-value for the F-statistic is 0.0133, which means that at least one of the independent variables in this model is significantly associated with children's IQ.

From the coefficient of slope, controlling for other variables, on average, children with high lead levels are 2.52 points lower in IQ than children with low lead levels. The p-value for the slope of lead exposure is testing the null hypothesis that there is no association between the lead exposure and IQ. Since the p-value is 0.0949, more than 0.05, we fail to reject the null hypothesis. Therefore, controlling for age, sex and race, there is a non-significant association between lead exposure and IQ.

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

Solution: From the p-values for age, sex and race, only the variable Asians, with a small p-value 0.0238, is significantly associated with IQ, after controlling for the other variables in the model. IQ is 4.23 points higher on average for Asian children than White children, controlling for all other variables in the model.

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.

Solution:
The predicted IQ for a white,10 year old male without lead exposure is, 96.1188 + 10*0.3791 = 99.9098

The predicted IQ for a white, 10 year old male with lead exposure is, 96.1188 + 10*0.3791 - 2.5169 = 97.3929

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?

Solution: I use the "table()"command in R to find the numbers of children for high lead levels and low lead levels, and for those with ADHD and without ADHD like this,

	ADHD	
lead	With ADHD	Without ADHD
low lead levels	311	62
high lead levels	87	40

Then I calculate: 40/(87+40)= 0.3150, 62/(311+62)= 0.1662

Hence, the percent with ADHD for those with high lead levels is 31.5%, and the percent with ADHD for those with low lead levels is 16.62%.

To compare these percentages, I'll do a chi-square test in R. The null hypothesis for this test is that the two percentages are equal.

I get the chi-square statistic 12.908, and so the p-value is 0.0003, less than 0.05, which indicates significant difference between the two percentages at the 95% confidence level. Based on this analysis, the percent of children with ADHD is higher for those with high lead levels compared to those with low lead levels.

	High Lead Levels		Low Lead Levels	
	Proportion	n1	Proportion	n2
With ADHD	31.50%	127	16.62%	373

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.

Solution: Since ADHD is a dichotomous variable, I conduct a logistic regression in R to predict ADHD from age, sex, race, and lead exposure here. The following are partial results from the logistic regression.

Variable	Coefficient	Standard Error	p-value
Intercept	-0.8066	0.6780	0.2342
Age	-0.0248	0.0778	0.7494
Sex Female	-0.6074	0.2331	0.0092
relevel(factor(race), ref = "1")2	-0.2254	0.2983	0.4500
relevel(factor(race), ref = "1")3	-0.9684	0.3816	0.0112
relevel(factor(race), ref = "1")4	-1.4020	0.4300	0.0011
lead	0.9741	0.2478	8.44e-05

I use "exp()" command to get the odds ratios for each variables, and get their 95% confidence intervals.

Variable	Odds Ratio	95% CI	p-value
Intercept	0.4464	(0.1168, 1.6746)	0.2342
Age	0.9755	(0.8372, 1.1363)	0.7494
Sex Female	0.5447	(0.3433, 0.8576)	0.0092
relevel(factor(race), ref = "1")2	0.7982	(0.4373, 1.4145)	0.4500
relevel(factor(race), ref = "1")3	0.3797	(0.1707, 0.7719)	0.0112
relevel(factor(race), ref = "1")4	0.2461	(0.0975, 0.5381)	0.0011
lead	2.6487	(1.6271, 4.3072)	8.44e-05

Given the 95% CI for the odds ratios, sex($p=0.0092<0.05$), Hispanic($p=0.0112<0.05$), Asian($p=0.0011<0.05$), and lead($p<0.05$) are significantly associated with ADHD. Also, 95% CIs for OR of these variables do not contain null value of 1.0, indicating ADHD significantly associates with sex, Hispanic, Asian and lead.

- How well does this model predict ADHD?

Solution: The model chi-square = 505.90 – 469.02 = 36.88, with 6 degrees of freedom, p-value < 0.01. Therefore, at least one of the variables in the model is significantly associated with ADHD.
The C-statistic for the logistic regression model predicting ADHD is 0.6914, which indicates a weaker predictive model.

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

Solution: Children with high lead levels have 2.65 times the odds of having ADHD than those with low lead levels, controlling for all other variables in the model. Given the p-value less than 0.05 for the odds ratio, we are 95% confident that this difference is significant.

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

Solution: From the p-values, except lead exposure, the variable sex($p=0.0092<0.05$), Hispanic($p=0.0112<0.05$) and Asian($p=0.0011<0.05$) are significantly associated with ADHD, after controlling for the other variables in the model.

Females have 0.55 times the odds of having ADHD than males, controlling for other variables in the model. Hispanics have 0.38 times the odds of having ADHD than Whites, controlling for other variables in the model. Asians have approximately 0.25 times the odds of having ADHD than Whites, controlling for other variables in the model.

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.

Solution:
The predicted ADHD for a white,10 year old male without lead exposure is,
-0.80657-0.02484*10 = -1.00657
exp(-1.00657) / (1+exp(-1.00657)) = 0.258
The predicted probability of ADHD for a white, 10 year old male without lead exposure is 0.258.

The predicted ADHD for a white, 10 year old male with lead exposure is,
-0.80657-0.02484*10+0.97409 = -0.08088
exp(-0.03248)/(1+exp(-0.03248)) = 0.492
The predicted probability of ADHD for a white, 10 year old male with lead exposure is 0.492.