Lab Eight: Goodness-of-fit and Independence Tests; Simple Linear Regression

STA 111 (Summer Session I)

Lab Objective

The purpose of the lab is to perform chi-squared goodness of fit and independence tests using R and to also gain some experience with simple regressions.

Introduction

You can use the RStudio pre-installed on the lab computers or on your personal laptops.

Before continuing, download the lab report template by typing:

```
download.file("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2016/Labs/Lab.Rmd", destfile ="Lab.Rmd")
```

Note: remember to use the right double quotation marks when you copy commands into R!

Lab Procedures

1. Goodness-of-Fit

In the U.S., you are supposed to be tried by a jury of your peers. Does this really happen in practice? A study in the UCLA Law Review (1973) of grand juries in Alameda County, California, compared the demographic characteristics of a random sample of jurors with the general population. Below are the data for age and educational level. Only persons 21 and over are considered; the population data are known from the Public Health Department.

Age	County-wide %	# of Jurors
21-40	42	5
41-50	23	9
51-60	16	19
> 61	19	33
Total	100	66

Educational Level	County-wide %	# of Jurors
Elementary	28.4	1
Secondary	48.5	10
Some college	11.9	16
College degree	11.2	35
Total	100.0	62

You can manually create your own data frames of the data in the tables or load them from the course directory; they are saved as *"JurorsAge.txt"* and *"JurorsEducation.txt"*.

Questions:

1. Test whether the juries appear to be randomly selected with respect to the distribution of *age* in the county. Report the expected number of jurors in each age group, the value of the chi-squared test statistic and its degrees of freedom, the p-value, and your conclusion.

   For a goodness-of-fit test, type `chisq.test(Data$var1, p=Data$var2, rescale.p = TRUE)` where **Data** is your data frame, **var1** is the variable you wish to test and **var2** is the variable containing the proportions/percentages/distribution you wish to test. In this section, County-wide would be your **var2**.

2. Perform the goodness of fit test for the *education* data by hand. Show in your report the null hypothesis, the four pieces of the test: all values of (observed - expected)^2/expected, the degrees of freedom, the p-value, and your conclusions. Use R to check your answer, but all the by-hand work must appear to get full credit.

2. Independence

Do people's opinions of their appearance change with age? In a survey reported in Newsweek magazine (Spring/Summer 1999), 747 randomly selected women were asked, "How satisfied are you with your overall appearance?" The numbers of women who chose each of four answers are shown in the table below.

Age	Very	Somewhat	Not Too	Not At All
Under 30	45	82	10	4
30 - 49	73	168	47	6
Over 50	106	153	41	12

It is often easier to start with detailed data rather than aggregated data as in the table. The detailed data is saved in the course directory as *"Appearance.txt"*.

Questions:

3. Test the null hypothesis that women's satisfaction with their appearance is not associated with age. Report the value of the chi-squared test statistic, its degrees of freedom, the p-value, and your conclusion.

For an independence test between two variables, type `chisq.test(table(Data$var1,Data$var2))` where **Data** is your data frame containing the detailed data, **var1** is the first variable and **var2** is the second.

Sometimes, the warning message "Chi-squared approximation may be incorrect" is shown below the results. This warning message is usually due to small cell values in the contingency table, and the p-value calculations can thus be inaccurate. There are a number of solutions/ways around this problem but you can simply ignore it in this lab.

3. Linear Regression

Many macroeconomic studies use cross-sectional data (i.e., data from the same time frame) from countries around the world. Of particular interest is the factors related to Gross National Income (GNI), which essentially is the amount of money the country produces from all sources. Load in the *"Countries2010.txt"* data. It contains economic data for 175 countries from around the world as of 2010. All monetary values are expressed in U.S. dollars. The variables include:

Variable	Description
GNI	GNI per capita - GNI divided by the number of people in the country
BirthRate	Number of births per 1000 people in the country
DeathRate	Number of deaths per 1000 people in the country
InfantDeaths	Number of infant deaths per 1000 people in the country
FertilityRate	Average number of births per woman
LifeExpectancyMales	Average age at death for men
LifeExpectancyFemales	Average age at death for women
Region	Worldwide Region
CountryName	Name of country
CountryCode	3-letter code of country

4. Does a normal curve describe the distribution of per capita GNI well?

5. Which numerical variable has the strongest correlation with per capita GNI?

   Note that region, countrycode, and countryname are not numerical variables so correlations involving them make no sense and you need to remove them. In our case, the first 3 variables are string variables, so type `cor(Countries2010[,-c(1:3)],use = "complete.obs")` to exclude them when computing the correlations and add the **use** option as well to specify that missing entries should be ignored.

6. What is the regression equation for predicting per capita GNI (y) from birth rate (x)? Interpret your results.

For linear regression in R, type `Model = lm(y ~ x, data = Data); summary(Model)` where **y** is the response variable, **x** is the independent variable and **Data** is the data containing both **x** and **y**.

7. Now create a new variable in the data frame to be the (natural) logarithm of per capita GNI. Does a normal curve describe the distribution of logarithm of per capita GNI well?

   The `log` command computes logarithms and by default, natural logarithms.

8. What is the regression equation for predicting the (natural) logarithm of per capita GNI (y) from birth rate (x)? Interpret your results.

   The last two questions can be done easily in R but I want you to do them by hand.

9. If a country has a birth rate of 30 people per 1000, can you use the regression equation to predict the per capita GNI? If you think so, write down the estimated per capita GNI (take "e" raised to the predicted log(per capita GNI)). If you think not, explain why not in at most one sentence. For that country, what is the probability that it's per capita GNI exceeds 9000?

10. If a country has a birth rate of 80 people per 1000, can you use the regression equation to predict the per capita GNI? If you think so, write down the estimated per capita GNI (take "e" raised to the predicted log(per capita GNI)). If you think not, explain why not in at most one sentence.

This ends the lab. Remember to turn in your lab reports on Sakai.