CE 472 **–** Computer Applications in Highway and Traffic Engineering

Homework 2 – Statistics with R

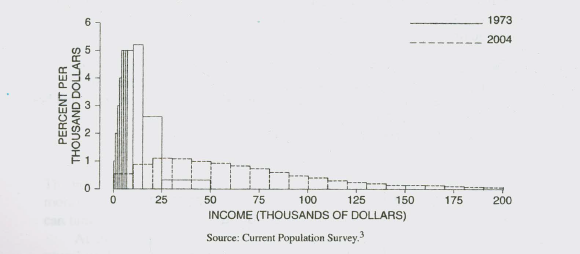
Instructions:

1. All HW submissions must be through moodle / e-mail (No prints! Save paper!) and shall include the following:
   1. A well-edited document with answers to the questions below.
   2. An edited R-script file in a document with the final correct commands. Please include the output from each command immediately after the command line.
2. The HW is due on Sep 14th 2019.
3. Ads for ADT–India Security Systems Claim:

*When you go on vacation, burglars go to work…. According to police statistics over 25% of home burglaries occur between May and August.*

Do the statistics prove that burglars go to work when other people go on vacation? Answer yes or no, and explain briefly.

1. The figure below compares the histograms for family incomes in the U.S. in 1973 and in 2004. It looks as if family income went up by a factor of 4 over 30 years. Or did it? Discuss briefly.



1. In a survey, people were asked how many vehicles they owned. Results are shown below for people age 25-39, by educational level.
2. Is the number of vehicles owned discrete or continuous?
3. Draw histograms for these data (You may take “5 or more” as 5 – very few people had more than 5 vehicles)
4. What do you conclude?

*Distribution of people age 25-39 by educational level and number of vehicles (percent):*

|  |  |  |
| --- | --- | --- |
| Number of vehicles | People who are high school graduates (%) | People with college degrees (%) |
| 0 | 30.2 | 47.9 |
| 1 | 21.8 | 19.4 |
| 2 | 28.4 | 22.7 |
| 3 | 13.7 | 8.0 |
| 4 | 4.4 | 1.5 |
| 5 or more | 1.5 | 0.5 |

Note: Percent may not add to 100 % due to rounding

1. At the University of California, Berkeley, Statistics 2 is a large lecture course with small discussion sections led by teaching assistants. As a part of study, at the second-to-last lecture one term, the students were asked to fill out anonymous questionnaires rating the effectiveness of their teaching assistants (by name) and the course, on the scale

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| poor | fair | good | very good | excellent |

The following statistics were computed.

* The average rating of the assistant by the students in each section
* The average rating of the course by the students in each section
* The average score on the final for the students in each section

The results are shown below (sections are identified by letter). Draw a scatter diagram for each pair of variables – there are three pairs – and find the correlations.

|  |  |  |  |
| --- | --- | --- | --- |
| Section | Avg. rating of assistant | Avg. rating of course | Avg. score on final |
| A | 3.3 | 3.5 | 70 |
| B | 2.9 | 3.2 | 64 |
| C | 4.1 | 3.1 | 47 |
| D | 3.3 | 3.3 | 63 |
| E | 2.7 | 2.8 | 69 |
| F | 3.4 | 3.5 | 69 |
| G | 2.8 | 3.6 | 69 |
| H | 2.1 | 2.8 | 63 |
| I | 3.7 | 2.8 | 53 |
| J | 3.2 | 3.3 | 65 |
| K | 2.4 | 3.3 | 64 |

The data are section averages. Since the questionnaires were anonymous, it was not possible to link up student ratings with scores on an individual basis. Student ability may be a confounding factor. However, controlling for pre-test results turned out to make no difference in the analysis. Each assistant taught one section. True or False, explain:

1. On the average, those sections that liked their TA more did better on the final.
2. There was almost no relationship between the section’s average rating of the assistant and the section’s average rating of the course.
3. There was almost no relationship between the section’s average rating of the course and the section’s average score on the final.
4. Use the historicvmt.xls[[1]](#footnote-1) file for the following analysis. The file is a time series of data of vehicle miles travelled in the United States for every month starting from 1970 till 2009.
   1. Import the data into R. Convert file into appropriate CSV format. Use the read.csv and read.table command with header = TRUE.
   2. Explore the data. Plot the time series and comment on the trend.
   3. Use *aggregate* command to aggregate data by each year (use FUN = sum) and each month (use FUN = mean). Plot the time series for aggregated data by each year and a bar plot for aggregated data by each month.
   4. Check the MER\_T09\_04.csv[[2]](#footnote-2) dataset. This data set provides the price of gasoline (petrol) in US $ every month. The data contains price of different types of gasoline. The final set of data is an average price. Select the average price data from 1976 onwards and copy to clipboard. Then use the command *read.table(file="clipboard").*
   5. Explore the data. Use the correlation command to identify any correlation between vehicle miles travelled and gasoline price. Interpret the result.
5. The data below provides speed values (km/h) collected at a location on an interior road. Using R, determine all measures of central tendency and spread you know. Plot a histogram of the data. What is the 85th percentile Speed value in the data? Is the data normally distributed?

38.4, 34.5, 25.6, 27.9, 30.2, 43, 42.3, 29.9, 41.5, 49.1, 30.2, 39.6, 28.5, 39.1, 40.3, 36.6, 34.9, 21.3, 42.5, 49.7, 37.9, 42.1, 19.5, 45.6, 54.2, 43.92, 43.35, 39.89, 40.86, 34.41, 47.22, 37.41, 44.26, 41.88, 40.57, 41.13, 32.47, 46.50, 41.51, 40.59, 39.40, 39.39, 36.25, 34.42, 41.68, 36.98, 41.06, 40.38

A traffic calming measure was introduced at the location. The new speed values obtained were:

40.97 30.38 26.82 47.27 33.78 22.05 25.41 47.08 17.75 15.21 27.46

32.67 9.72 32.31 33.84 40.50 19.38 24.90 49.12 20.01 38.61 40.55

24.09 40.41 30.30 44.48 31.21 37.06 26.33 30.36

Has there been a significant reduction in speeds?

1. Speed.txt file contains data of walking speeds. The first column contains walking speeds of individuals, column 2 classifies individuals based on age (c – child, y – young, m – middle age, o – old), column 3 classifies individuals based on the luggage they carry (n – no luggage, s – small luggage, h – heavy luggage), and column 4 classifies individuals based on gender. Using ANOVA comment on the effect of different categories of individuals on their walking speeds. Are there any interaction effects?
2. The table below gives the frequency distribution of the number of vehicles arriving at an intersection in 1 minute intervals (cycle length of intersection = 1 min). Fit a Poisson distribution to the data. Comment on the goodness of fit.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | >8 |
| Frequency of Occurrence | 2 | 3 | 15 | 18 | 24 | 10 | 6 | 2 | 1 | 0 |

1. The attached file, hhcomp.csv, is extracted from 2001 US National Household Travel Survey (NHTS) database. Data from two states (TX, MD) are presented. The attached dictionary file explains the variable used.
   1. Perform exploratory data analysis to better understand the dataset. Provide your insights and interpretation on the results from the data analysis.
   2. Develop regression models using CNTTDHH as the dependent variable. What is the best overall model you can develop? Make sure you include only significant variables in the final model. Also watch out for variables that may be highly correlated.
   3. Incorporate dummy variables in the regression equation to accommodate non-linear effects.
   4. Interpret a few of the parameters (including the dummy variable parameter) in the regression equation.
   5. Are the residuals normally distributed? Use qqplot, qqline, as well as Shapiro and K-S tests. What are the implications / remedies?

1. <http://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm> [↑](#footnote-ref-1)
2. <http://www.eia.gov/totalenergy/data/monthly/query/mer_data.asp?table=T09.04> [↑](#footnote-ref-2)