Demand Modeling - 1.202

Problem Set 1
Due Date: March 8, 2019

1 Case Study 1: Aggregate Trip Generation Rates (50 points)

1.1 Objective of the case study

The objective of this case study set is to make you familiar with linear regression models. In this case study, you have to estimate a linear regression model that predicts trips per occupied dwelling unit.

1.2 Software

The reference software is **R**. However, you can use any other regression software of your choice that has enough functionality to complete the assigned tasks (e.g. MATLAB, SAS, STATA). Please check with your TA if you want to use a software application different from **R**. **R** can be downloaded at http://www.r-project.org/.

1.3 Data

You are given a set of aggregate data for 57 traffic analysis zones in the Chicago Area collected in a survey conducted in the 60s. For each of the 57 zones you have the information on the average trips per occupied dwelling unit, the average car ownership, the average household size, and three zonal social indices. The data file (CS1_data.csv) is available at the Stellar course website. The detailed description of the variables is enclosed in the appendix.

1.4 Your tasks

1. Give a brief description of the cause/effect relationships you think are relevant in predicting trip generation rates and formulate a hypothesis (a priori beliefs of the relationship). You may find it useful to do some preliminary statistical analysis on the data before this (in a similar manner as in Case Study 0).
2. Estimate the model to predict the average trips per occupied dwelling unit to reflect the hypotheses.
3. Test different specifications and perform statistical tests comparing the specifications.
4. Select the best model specification.

1.5 Report Content

Your final report should include:

1. Description and discussion of the cause/effect relationships you think are relevant for trip generation rates;
2. Description of no more than three model specifications which you believe reflect your a priori (i.e., part 1) considerations;
3. Presentation of your "best" specification (it is up to you to define "best") and a discussion of your selection criteria;
4. A discussion of the similarities and differences between the causal inferences from your "best" specification and your a priori considerations; and
5. The programming code (if applicable) you wrote to estimate the regressions.

1.6 Some Comments

1. The following criteria will be applied for grading:
   a) Your understanding of the problem (e.g. causal relationships);
   b) Your understanding of linear regression analysis;
   c) Your utilization of the regression software (evidence you estimated a regression model, etc.);
   d) Your understanding of regression statistics and hypothesis testing (explain what the statistics mean).
2. Remember that you must always examine and comment on your results. Computer outputs without explanations are not acceptable (hand-written notes are fine).

Appendix

The variables available for your specification are:

Name	Description
TODU	Trips per Occupied Dwelling Unit: Trips refer to the daily frequency of person-trips via motor vehicle (auto driver or passenger) or public transit made from a dwelling unit by members of that dwelling unit. All trips whose origins were other than "from home" were ignored.
ACO	Average Car Ownership: Cars per dwelling unit.
AHS	Average Household Size: Number of residents per dwelling unit.
SRI	Social Rank Index: This index contains two elements: (i) the proportion of blue-collar workers, defined as the ratio of craftsmen, operatives, and laborers to all employees; and (ii) educational level as measured by the proportion of persons 25 years and older completing eight or fewer years of schooling. The social rank index is inversely related to both ratios; hence, it attains a maximum value where no residents fall into the blue-collar jobs category, and all adult residents have more than eight years of education
UI	Urbanization Index: This index contains three elements: (i) fertility rate, defined as the ratio of children under five years of age to the female population of childbearing age; (ii) female labor force participation rate, meaning the percentage of women who are in the labor force, and (iii) incidence of single family units or simply the percentage of single units to total dwelling units. The degree of urbanization index would be increased by (a) lower fertility rate, (b) higher female labor force participation rate, and (c) lower proportion of single dwelling units. In a sense, this index measures in a rather negative way the degree of attachment to the home. High values for this index imply less attachment to the home because of fewer children, higher likelihood of women being employed, and less permanency of dwelling unit type in terms of average tenure.
SI	Segregation Index: This index is defined as the proportion of an area's residents who belong to certain minority groups, such as non-whites, foreign-born Eastern Europeans, etc. It measures the extent to which these minority groups live in relative isolation.

2 Simulation Exercise (10 points)

Let ε be Gumbel distributed so that its CDF and PDF are given as:

$$F(\varepsilon) = \exp\left[-e^{-\mu(\varepsilon-\eta)}\right], \mu > 0$$
$$f(\varepsilon) = \mu e^{-\mu(\varepsilon-\eta)} \exp\left[-e^{-\mu(\varepsilon-\eta)}\right]$$

where η is a location parameter and μ is a positive scale parameter. (See page 104 in Discrete Choice Analysis (1985) or page 164 in the DCA book for the properties of Gumbel distribution). The goal of this exercise is to use simulation to support the stated properties. Please do not use any built-in function for generating Gumbel distributed numbers, but write your own code.

a) **[3 points]** For a Gumbel distributed random variable $\varepsilon \sim EV(\eta, \mu)$, the mode is η, the mean is $\eta + \frac{\gamma}{\mu}$ where γ is Euler's constant 0.5772, and the variance is $\frac{\pi^2}{6\mu^2}$.

   Assume that $\eta = 3$ and $\mu = 1.2$. Simulate a Gumbel distribution, estimate the mode, mean and variance and check if they match the stated properties of the distribution. Plot a histogram of ε.

b) **[5 points]** Another property is that if $\varepsilon_j \sim EV(\eta_j, \mu)$, for $j = 1, \dots, J$, and that if ε_j are independent with the same scale parameter μ, then $\varepsilon = \max_{j=1,\dots,J} \varepsilon_j$ and $\varepsilon \sim EV(\eta, \mu)$, where $\eta = \frac{1}{\mu} ln \sum_{j=1}^{J} e^{\mu\eta_j}$. In this case, the variance is still $\frac{\pi^2}{6\mu^2}$, while the expected value is $E[\varepsilon] = \eta + \frac{\gamma}{\mu}$.

   Assume that $\eta_j = 5$ for all j, and $\mu = 1.2$. Simulate Gumbel distributions with $J = 4$, and check if the expected value and the variance of the simulated distributions match the stated properties of the distribution. Plot a histogram of ε.

c) **[2 points]** Now assume that $\eta_1 = 8, \eta_2 = 8.2, \eta_3 = 7.5, \eta_4 = 7.8$, and $\mu = 1.2$. Show that the simulation results match the stated properties of the distribution. Plot a histogram of ε.

3 Supplemental Problems (40 points)

1. **[10 points]** A common strategy for handling a case in which an observation is missing in the data for one or more independent variables is to fill those missing variables with 0's and add a variable to the model that takes the value 1 for that one observation and 0 for all other observations. Show that this 'strategy' is equivalent to discarding the observation as regards the computation of the intercept and the slope parameter. In other words, show that the ordinary least squares estimates for a and b in model (1) are equal to the ordinary least squares estimates for α and β, respectively, in model (2). Model (1) is estimated using the first $N-1$ observations where data on the independent variable X are available. Model (2) is estimated for all N observations, where data on the independent variable X are available for the first $N-1$ observations but are missing for the N^{th} observation. Z_n is a dummy variable equal to 1 for the N^{th} observation and is equal to zero for all other observations. ε_n is the error term of model (1), and v_n is the error term of model (2).

   $Y_n = a + bX_n + \varepsilon_n, \quad n = 1,\dots,N-1$..(1)

$$Y_n = \alpha + \beta X_n + \gamma Z_n + \nu_n, n = 1,\ldots,N \ldots\ldots (2)$$

2. **[14 points]** Consider a model that predicts the accident rates for different states in the U.S.:

$$Y_n = \alpha + \beta X_n + \varepsilon_n, n = 1, \ldots, 50$$

where n denotes a state, X_n is the proportion of vehicles exceeding speed limit (55 miles per hour) on the highways of state n, and Y_n is the number of fatalities per million vehicle miles. The sample averages are: 0.6 and 1.0 for X and Y respectively. The sum of squared deviations of X from its average is 10, the sum of squared deviations of Y from its average is 1, and the sum of cross products of deviations of X and Y from their respective averages is 2.

a) Compute the OLS estimates of α and β, their standard errors, the sum of squared residuals, and R^2.
b) Test the null hypothesis that $\beta = 0$.
c) Test the hypothesis that strict adherence to the speed limit would halve the average accident rate.
d) The validity of your answers to parts a), b), and c) are based on the assumptions for OLS estimation and hypothesis testing that we've discussed in class. Are these assumptions reasonable for the described data set? Why or why not? Provide examples to illustrate your reasoning.

3. **[6 points]** Prove that in a least squares regression of Y on a constant and multiple X's, the regression hyperplane passes through the sample averages assuming that the intercept is included. (The regression hyperplane is the set of all points $(y, x_1, x_2, \ldots, x_k)$ that are possible from the estimated model. For the bivariate case, the hyperplane is the fitted line.)

4. **[10 points]** Pay-As-You-Drive (PAYD) leasing, a car leasing program that charges a small fixed price when the car is leased and a variable price based on mileage traveled, has been in service for one year.

Your task is to design a sampling strategy and choose the appropriate sample size to estimate the proportion of people who use PAYD within ± 2 percentage points (in share points) at a 90% confidence level.

a. Compute the required sample size for a random sample.

b. Users may be stratified according to their household mileage level (low mileage vs. high mileage). The proportion of low mileage households to high mileage households is 40:60. The penetration rate for low mileage households is believed to be less than 15%, and the penetration rate for high mileage households less than 5%. Design an optimal stratified sample that meets the precision requirements specified above.