Assignment 2: Simple linear regression for modeling household trip generation

For this assignment you will **extend the household trip generation** model developed in week 3 class (30.10.2024). Basically, I want you to try **two more variables** in addition to worker count (WRKCNT) and vehicle count (HHVEHCNT). For this assignment you will need to use only **the household file, i.e., hh_ga.csv**.

You can try the following two variables:

1. Income given by the column named HHFAMINC
2. Family size given by the column named HHSIZE

Note that there are 11 categories of valid income and 2 non-valid income categories in the data. First you need to ***clean the data to get rid off the*** *negative income values, i.e., you would delete the rows where the income is -ve. The code for doing that in R is provided in the cheat sheet.*

Logistics

- **Prepare a 1-to-2-page** (Font size: 10 to 12 point, single spacing) **report in Microsoft Word.**
- Please note that you can work in a group of 2 members for the assignment.
- You can divide the work among yourselves. *But the assignment should clearly identify which portion has been written and analyzed by whom. Without a clear statement about the member contribution the assignment will not be graded.*
- For analysis I highly recommend that you use some programming language and not rely on Excel, I will provide some help on R – but you will need to work on your own if you choose other languages such as Python or Matlab.

Submission

1. First justify why you think the above variables will be useful for explaining household trip generation

2. Develop two models by adding each of the above two variables as if as if the variables are affecting the trip count linearly i.e., the form of the equations developed in the two models would be
   (i) Y = a + b1*WRKCOUNT + b2*HHVEHCNT+ b3 * HHFAMINC
   (ii) Y = a + b1*WRKCOUNT + b2*HHVEHCNT+ b3 * HHFAMINC + b3 * HHFAMINC + b4 * HHSIZE

Or,

   (i) Y = a + b1*WRKCOUNT + Z1*I(HHVEHCNT==1) + .. + Z5*I(HHVEHCNT>=5)+ b3 * HHFAMINC
   (ii) Y = a + b1*WRKCOUNT + Z1*I(HHVEHCNT==1) + .. + Z5*I(HHVEHCNT>=5)+ b3 * HHFAMINC + b3 * HHFAMINC + b4 * HHSIZE

3. Comment on the sign of the estimates obtained in models (i) and (ii). Does the sign support your theoretical basis identified in part 1 or in other words does the sign of the estimate support your intuition?

4. Comment on the t-statistics of the estimated parameter of models (i) and (ii). Can you conclude that the estimates are statistically significantly different from zero?

5. Now let's test the linearity assumption using dummy variable for HHFAMINC. Note that since there are 11 categories of income, you can estimate (11-1) = 10 parameters corresponding to 10 income related dummy variables. *But, in statistics we recommend not to estimate any parameter if the no. of observation in a group is less than 50 (if you want to be conservative) or at least 30.* So, you should first check the frequency distribution of HHFAMINC. The code for doing that is provided in the cheat sheet.

   a. Estimate the parameters to the dummies. Your equation should look something like this: Y = a + b1 * WRKCOUNT + b2*HHVEHCNT + F1*I(HHFAMINC==1) + … +F10 * I(HHFAMINC==1) + b4 * HHSIZE
   Or,
   (iii) Y = a + b1*WRKCOUNT + Z1*I(HHVEHCNT==1) + .. + Z5*I(HHVEHCNT>=5)+ b3 * HHFAMINC + F1*I(HHFAMINC==1) + … +F10 * I(HHFAMINC==1) + b4 * HHSIZE

   Note: you need not use all the 10 dummies in the above equation. You should make a knowledgeable guess about how many dummies you need to include.

b. Comment on the sign and the magnitude of the estimates
c. Can you justify the use of all the dummies you introduced in part (a)? Do you think some dummies should be grouped together?

6. Identify the best model across all the different models you estimated in this assignment based on the adjusted R-square value.