

## Assignment 2

**Due: Start of class, Friday, October 23rd**

---

### **Instructions:**

Group Work: I encourage all of you to work on the problem set in groups, but each member of the group must write up his or her own solutions and perform his or her own analysis. Problem sets that are verbatim copies will receive a zero.

To submit your solutions, include your answers to these questions and your log file (or other supporting documentation) and submit it to the Assignment on Canvas.

### **Assignment:**

This assignment examines consumer vehicle ownership using publicly available data from the National Household Transportation Survey, conducted every 4-5 years by the Department of Energy.

The file ps2\_data.dta contains information about the vehicles owned by ~137,000 people.

<u>Variable Name</u>	<u>Description</u>
vehtype	Vehicle Type
epatmpg	EPA MPG rating
hybrid	Is the car a hybrid / alternative vehicle?
domestic	Is the car a domestic brand?
automaker	Brand
vehage	Vehicle Age
fueltype	Fuel Type (1 = diesel, 4 = gasoline)
gscost	Gas Price (\$/gallon)
rural	Rural household location
children_under_16	# of children under 16 in household
adults	# of adults in household
hh_income	Household Income (in thousands of dollars)
ba_grad	College graduate
hhstate	State

1. Open a log file to keep a record of your analysis. Click “File → Log → Begin” and give it a name that will be easy to remember.
2. Open the dataset and summarize at the data. How many observations are in the dataset? What is the average, age and fuel economy in the data? What fraction of automobiles in the data are hybrid vehicles and what fraction are domestic brands?
3. Let’s learn a bit more about the vehicles in the data. What are the market shares of different vehicle types? (Hint: In STATA, “**tab vehtype**” will count the observations in the data, grouping them by vehicle type)
4. We might be interested in understanding how the fuel economy of the car a person drives relates to their characteristics. Run a regression where fuel economy is a continuous dependent variable and the explanatory variables are: the number of children under 16, the age of the vehicle, whether someone lives in a rural area, gasoline price, household income, and whether someone has a bachelor’s degree. Interpret the coefficients of the regression.

5. Now, let's run a similar regression where we include a dummy variable for each vehicle type (i.e., each value of `vehtype`). Interpret the coefficients on vehicle-type dummy variables.
6. Comparing your regression results from #3 and #4, look at the coefficient on `children_under_16`. Has it changed? If so, what can we learn from how the coefficient changed?
7. In class 10, we discussed (or will discuss) how technological innovation may lead firms to develop more fuel efficient vehicles. Let's graph the relationship between the age of vehicle and the fuel economy. Since we might expect a different relationship for different classes of vehicles, let's graph each one separately. (Hint: we can get stata to create a separate scatter plots by including the "by" option in our command. E.g., a scatter plot of `rents` on `commuting distance`, by number of bedrooms would read **"graph twoway scatter rents distance, by(bedrooms)"**)

What do the scatter plots tell us?

8. Now, let's try to understand something about who purchases different classes of vehicles. We'll do this with a multinomial logit from class. We want to explain what vehicle types people purchase and will use the same explanatory variables as those above in question 4. (Hint: See the notes for class 8 for a simple example of the syntax (or in stata, type "help mlogit")). Interpret the results from the regression. In particular, what can we conclude about (a) households with children and (b) locations with high gas prices. Are there any other coefficients that you find particularly interesting?
9. What would happen to vehicle market shares if the U.S. government increased gasoline taxes (and retail prices) by \$0.20 per gallon?  
  
To do this, we need the predicted probabilities of purchasing the different types of vehicles. To get the predicted probabilities before and after (using CARS as an example):
  - (1) type **"predict car\_prob if e(sample), outcome(1)"** (This gives us probabilities for the "base outcome" (CARS) and names the variable "car\_prob")
  - (2) type **"replace gscost=gscost+0.20"** (This increases gasoline costs by 20 cents)
  - (3) type **"predict car\_prob\_withtax if e(sample), outcome(1)"** (This gives us probabilities for cars at the new, higher gasoline prices)
  - (4) Once you've done this, not just for cars, but for other vehicle types, summarize the predicted probabilities by vehicle type.
10. Finally, take a look at the data. Form a hypothesis of a relationship you think might exist in the data. For example, one might hypothesize that high income consumers are less sensitive to gasoline prices when buying a vehicle...

Then, use the tools you've learned (regress Y X) or (mlogit Y X) to test the hypothesis.

Make sure to: (1) clearly state the question you're asking, (2) show the regression results, and (3) interpret the results of your regression.

11. Don't forget to close your log file! Go to "File → Log → Close."