Instructions for Project II. Due in class on Friday, December 10

Preface
This handout provides the background for and instructions that go with your second Project for STAT 460. The project also serves as your final exam, and is worth 25% of your overall grad. It revolves around the analysis of data by using some of the tools you learned in class such as t-test, ANOVA, regression and logistic regression. Although the assignment associated with the project is individual, you are expected and encouraged to work in groups on it, but keep in mind you will each have a slightly different dataset. We begin the handout with the background, then describe the data and how to access it, and finally describe the nature of what is expected of you as a work product.

Please read this document quickly now and then again more carefully later, so that you will understand what you need to do and when. Questions on how to begin thinking about the project and on setting up your own personal dataset can be directed to the TA or to Professor Slavkovic.

This handout is being distributed during the week of November 22 and will be used in conjunction with the last two lab assignments and the last homework. Please begin by reading the background and the description of the data.

I: Background
A small car dealership near State College is evaluating its profits and reconsidering its inventory. With increasing gas prices and consumer awareness of the pollution, the car market has been unstable. For example, in the past years the large SUVs were a big hit as well as more expensive cars, but now the dealership owner thinks that there is a larger demand for the smaller cars and/or mini SUVs and cars under $25,000. But all of this information is a bit overwhelming and he would like to have a summary of (1) what are currently the best indicators for the car price, and (2) what factors or car characteristics would increase his odds of selling lots of cars under $25,000. Ultimately, his goal is to increase his profits. On the other side, a friend of mine is on the market for a new car. She is interested in getting the best deal on an all-wheel drive car under $25,000, but is not sure yet if she wants a SUV or not. Your goal is to try to help them make the most optimal choices given the data described below.

II: Statistical Analysis
This data are obtained from [link to data source]. Please READ the data description at [link to data description]. Here are some pieces of that description:

NAME: 2004 New Car and Truck Data
TYPE: Sample
SIZE: 428 observations, 19 variables

DESCRIPTIVE ABSTRACT:
Specifications are given for 428 new vehicles for the 2004 year. The variables recorded include price,
measurements relating to the size of the vehicle, and fuel efficiency.

SOURCE:

VARIABLES. The variables in the dataset are the following:

   LABEL : Variables
   Name : Vehicle Name
   Sports : Sports Car? (1=yes, 0=no)
   SUV : Sport Utility Vehicle? (1=yes, 0=no)
   Wagon : Wagon? (1=yes, 0=no)
   Minivan : Minivan? (1=yes, 0=no)
   PICKUP : Pickup? (1=yes, 0=no)
   AWD : All-Wheel Drive? (1=yes, 0=no)
   RWD : Rear-Wheel Drive? (1=yes, 0=no)
   SRP : Suggested Retail Price, what the manufacturer thinks the vehicle is worth, including adequate profit for the automaker and the dealer (U.S. Dollars)
   Invoice : Dealer Cost (or "invoice price"), what the dealership pays the manufacturer (U.S. Dollars)
   Engine : Engine Size (liters)
   Cylinders : Number of Cylinders (=1 if rotary engine)
   Horsep : Horsepower
   CMPG : City Miles Per Gallon
   HMPG : Highway Miles Per Gallon
   Weight : Weight (Pounds)
   Wheel : Wheel Base (inches)
   Length : Length (inches)
   Width : Width (inches)

Here are some specific questions you should answer, and these should help you address the major questions of the dealership owner and my friend:

(a) Is there a significant difference in SRP between SUV and non-SUV vehicles?

(b) Given the data in this dataset, we’ll assume that a reasonable measure for a profit for the car dealer is the difference between SRP and INVOICE prices. Create a new variable PROFIT=SRP-INVOICE. Is there a significant difference in profit between SUV and non-SUV vehicles?

(c) Is there significant difference in MPG depending on a car type (e.g. SUV, vs. Sports vs. Minivan..)

(d) The PROFIT ESTIMATION MODEL. This regression model should describe which car characteristics best predict the car price. That is which characteristics will best describe the profit for the car dealership. You should use PROFIT as dependent variable. This could potentially help the dealer decide on his new inventory.

(e) The PROFIT INCREASE MODEL. This logistic regression model should describe which factors or cars characteristics would increase his odds of selling more cars which are priced under $25,000. This model should address the question my friend has as well. Notice, that for this you need to create a new dependent variables that has only two levels (e.g. SRP under $25K and over $25K;
you can also think of how does this translate to the PROFIT variables and create a new dependent categorical variable based on profit, rather than SRP).

You should make sure you use the test data to evaluate your models as well (see the next section).

III: Downloading Your Data

There are two tab-delimited datasets at http://www.stat.psu.edu/ sesa/stat460/project.html, SLA_cars.txt and 04carsTEST.txt. Begin here and copy down these two files.

SLA_cars.txt is a subset of the dataset described in the previous section and it has 378 observations (not 428) and 19 variables. Load the dataset into SAS. Before you do anything with the data, try to interpret the values and match them with the information above to make sure that you have variables that make sense. You should also verify that the missing values are denoted with *.

This is IMPORTANT! Next, you will want to create your own personal dataset with 300 observations and 19 variables, and with proper labels to help you with your analyses. You are to pick a random sample of 300 observation from 378 in SLA_cars.txt. In SAS you do the following:

- Choose Data/Random Sample
- In the number of rows put 300
- Hit OK

Now you will get a new worksheet. Make sure you have 300 observations and 19 variables. You want to save this new dataset:

- File /Save As
- Save in: Desktop (or some other location)
- Filename: use first three letters of your last name, underscore, word cars with extension .dat (to be saved as tab-delimited data); e.g. I would save it as SLA_cars.dat

All your analysis will be done on this dataset, and you will TURN IN (email me and the TA) this dataset when you submit TWO hard copies of your project report! (See next section for further instructions).

You are not required to do analyses in SAS. If you do not use SAS, however, it would be wise in your written report to indicate the program or package that you did use.

04carsTEST.txt is a subset with 50 observations and 19 variables from the original dataset of 428 observations. Everyone will have this same test dataset. You will use this dataset to test the prediction power of your models that you developed based on your own personal dataset of 300 observations. You are not do anything with this dataset, until you have decided on the best models based on the analysis of the personal dataset. Then, think of these 50 observations as if someone has given you a NEW data, and you want to check how well does your model perform on this new data. We’ll talk more about this in the upcoming labs and the class.
IV: Lab Assignments and Other Milestones

Either before the lab on Monday, Nov.29 or at the beginning you will need to create your personal dataset. You will be asked to work initially in groups in the lab, and you are free to discuss all of your work with the others. You will find it useful to show your plots to other students, TA, faculty, etc., and share ideas.

The last homework assignment will involve using your datasets. As a part of this homework, you will email me AND the TA the dataset **only** with the email subject: stat460 [name of your data set]; e.g. stat460 SLA_cars.dat.

On Friday, Dec. 10 you will submit in class TWO hard copies of your project report. If you intend to turn it in on Monday, please make sure to let me know at least by Wed. Dec. 8 that you intend to turn it on Monday rather than Friday.
V: Project Written Report – Due in class, Friday December 10.

Your Project II Assignment is to prepare your own (not a group) memorandum (of at most five double-spaced pages) addressed to the car dealership owner and my friend describing and evaluating the specific questions you answered and two models you created. Emphasis should be on what you found, not how you found it. Because the owner does not have a statistical background (although he is very intelligent), your memo should be concise and easy to read with no statistical jargon. This is not an invitation to be sloppy; it is simply a directive to interpret the results in a nontechnical yet precise manner. Make sure you include all appropriate caveats in straightforward terms.

You should prepare a one-page cover to your report with a brief title, your name and Section, and an executive summary of your findings.

You should back up your memo with technical appendices (10 pages at most); a suggested (but not mandatory) format is:

1. A table summarizing your final results.

2. Why you did what you did it. Document anything unusual (outliers, transformations, etc.). Be specific – Don’t just say “I transformed this variable,” but rather say “I transformed this variable because...”

3. Residual checks of the selected models. Evaluate the plausibility of your assumptions.

4. Present a summary of the two or three best models (at each stage) and discuss your final choice. Include graphs, charts, or whatever but all computer output should be labeled, highlighted, and neatly included. Note, it is not necessary (or recommended) for you to hand in all the computer output that you generated.

Most of your grade will depend on the quality of analysis, not on whether you found the “correct” model. Presentation, clarity and conciseness will be important factors in your grade. Points will be allocated for the executive summary, the 5-page report, AND the 10-page technical backup.

The final decisions of what to include in the report and how to interpret the plots are to be your own. Aside from the time spent in the labs working on the project, we expect that you should be able to complete your data analysis in under 10 hours, but it may require additional time to prepare the writeup. Get started on your writeup early, even though we will only start our class discussion of model selection on Wednesday December 1.

The last homework assignment due on Wednesday December 8 will also focus on the project and thus you will be able to devote much of your time and energy associated with the rest of the class on this project.

Note: The actual analyses and the writeups are to be yours and the equations and we expect that the related conclusions may well differ from person to person since the data sets also differ. In fact we plan to replicate at least some of your work with your data set, and if you report results using someone else’s you will receive zero credit on the assignment.