More Complicated Modeling Exercise
Yet another example using NMES

When we think about modeling medical expenditures, there are a variety of complications: a) many persons won't have any medical expenditures in a given time period, b) many persons will have quite large medical expenditures in a given time period due to complications/comorbidity/general poor health. Typically, the assumption that medical expenditures are normally distributed is quite poor. There are a variety of modeling strategies to handle this type of data. One of the most common strategies is to build a two-part model:

1) Build a logistic regression model for the probability of having no expenditure
2) Build a linear regression model for the mean expenditure given a positive expenditure.

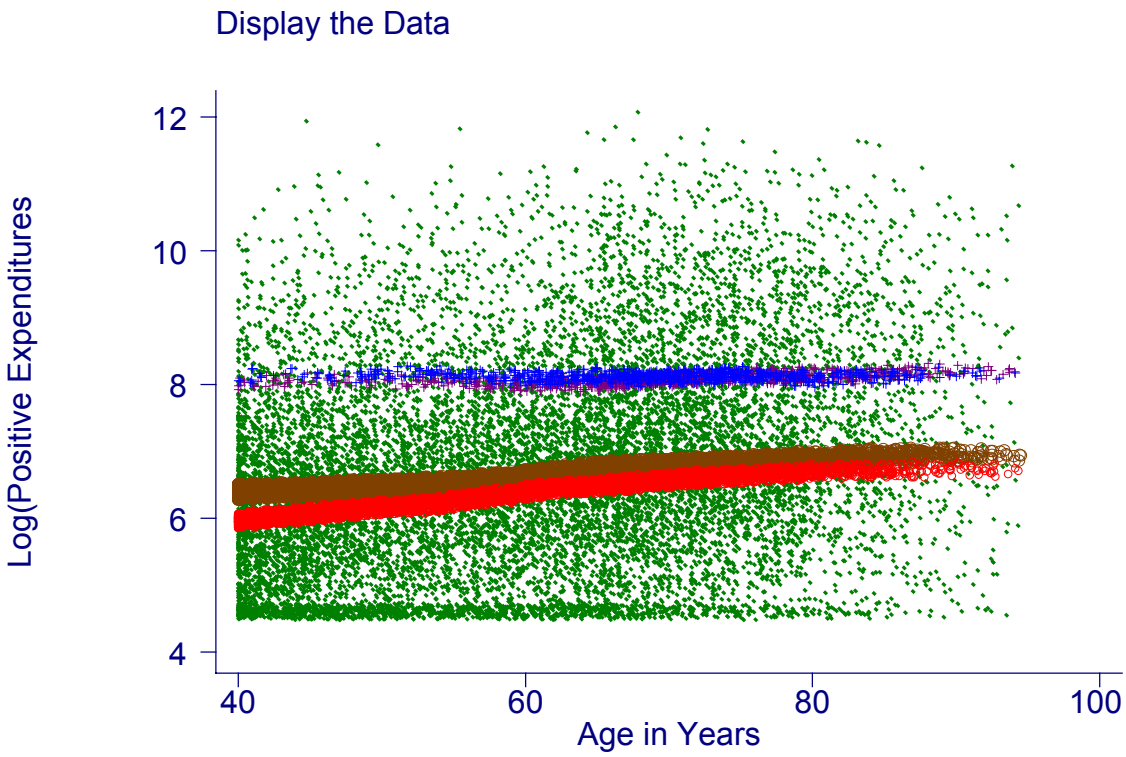
In model 2) the normality assumption is usually still violated, so many use a log-normal model or a model for the log(expenditure).

In this exercise, you would like to model the log(positive expenditures) as a function of age, gender and whether the subject has a major smoking caused disease (mscd, includes lung cancer, COPD, heart disease, and stroke). Again, the data available to us is the 1987 NMES. We will consider the data for subjects aged 40 to 94 as the prevalence of a mscd in ages under 40 is quite low. The figure below displays the data, we have plotted the log(positive expenditures) verses age. The "." are the raw data, the solid masses of points are kernel smoothes of the data (moving average) for each gender and mscd combination. We note that 1) the log(positive expenditures) are greater for those persons with a mscd compared to those without; 2) among those with no mscd, females tend to have larger log(positive expenditures) compared to males.

We would like to build a regression model that will allow for the following features:

1) The log(positive expenditures) are a function of age (possibly non-linear, you decide)
2) The association between log(positive expenditures) and age can be different depending on gender and the presence of a mscd.

Using the figure below and the criterion above, propose a linear regression model for this data.



Model:

Interpretation of the regression coefficients:

Using the model that you specified above, describe the statistical tests that you would perform to answer the following questions:

a) Is there evidence in the data to suggest that persons with a mscd have higher log(positive expenditures) compared to those without a mscd?
b) Is the rate at which log(positive expenditures) change with age different comparing persons with and without a mscd?
c) Is there evidence in the data to suggest that gender is associated with log(positive expenditures)?
d) Is the log(positive expenditures) vs. age relationship different comparing males and females?
e) Is the log(positive expenditures) vs. age relationship different comparing males and females with a mscd?
f) Is the log(positive expenditures) vs. age relationship different comparing males and females without a mscd?