Project for Regression Analysis in Public Health
June 21 – July 2, 2004

In this project, you will be exploring the relationship between having a major smoking caused disease and age, gender, smoking status and SES. The data are extracted from the 1987 NMES. The variables of interest are:

mscd = 1 if subject has lung cancer, laryngeal cancer, COPD, CHD, or stroke
0 otherwise

MALE = 1 if subject is male
0 if subject is female

eversmk= 1 if subject is an ever-smoker (at least 100 cigarettes in lifetime)
0 if subject is a never-smoker

LASTAGE = age of subject (ranges from 40 to 94 years of age)

educate= subjects level of education
1 if college grad
2 if some college
3 if hs grad
4 if other

marital= subjects marital status
1 if married
2 if widowed
3 if divorced
4 if separated
5 if never married

POVSTALB= subjects Poverty status
1 if poor
2 if near poor
3 if low income
4 if middle income
5 if high income

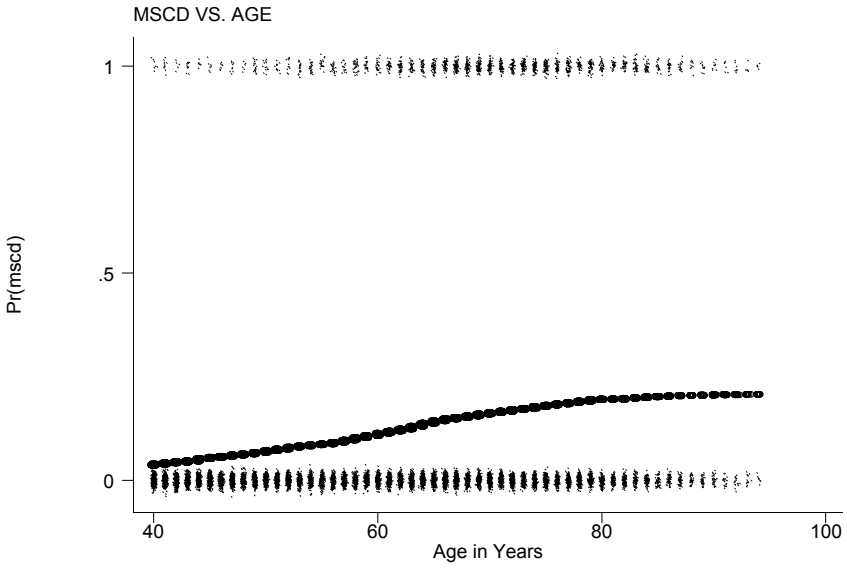
PART I:

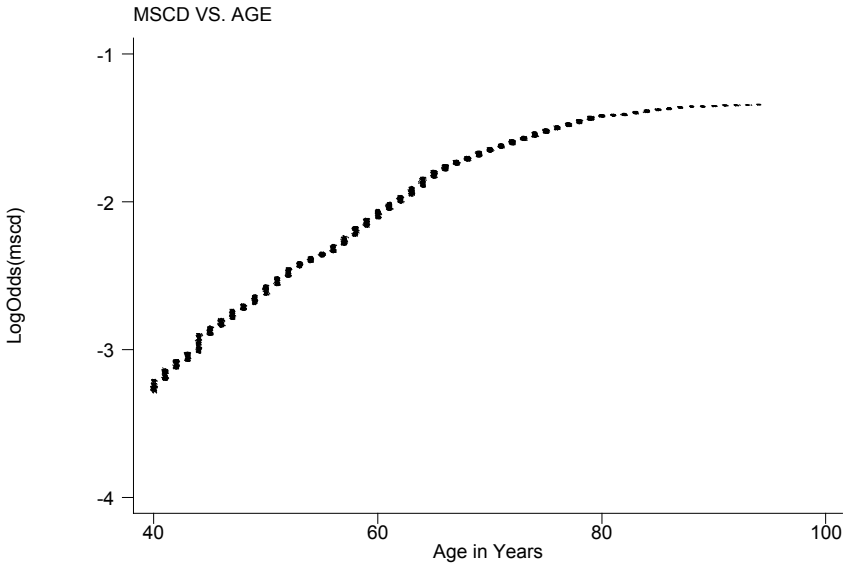
Your task is to build a series of regression models to answer the questions below. Use the plots provided with each question to aid you in your choice of regression model.

For each question you should

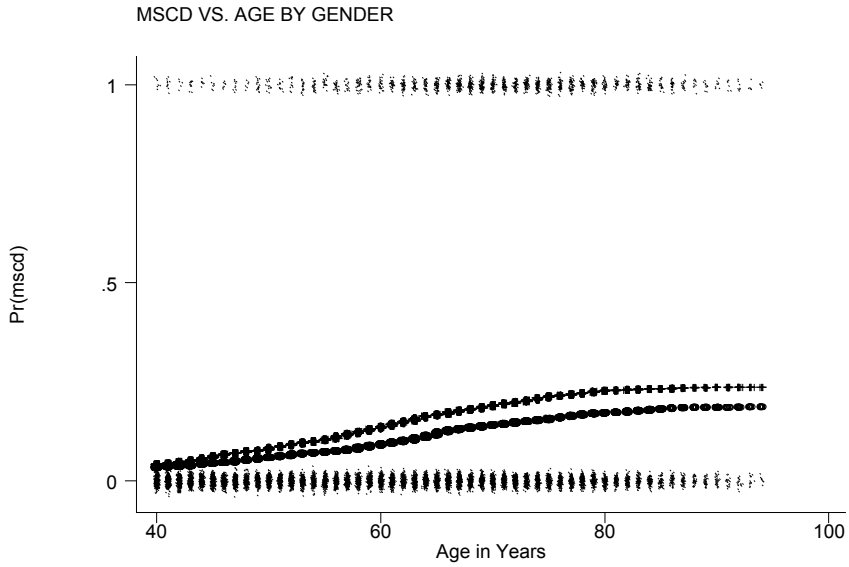
1. specify the regression model
2. interpret the regression coefficients
3. describe the statistical test that you would perform to answer the question

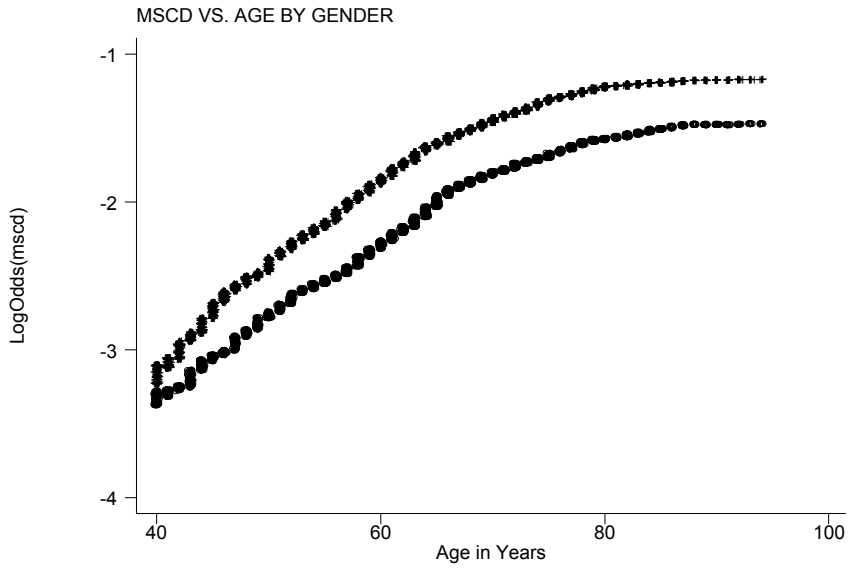
Question 1: Based on the figures below, what is the association between having a mscd and age? Specify a regression model for the relationship you observe in the data (note the non-linearity). Using this model, specify how you would test if the mscd/age relationship is linear.



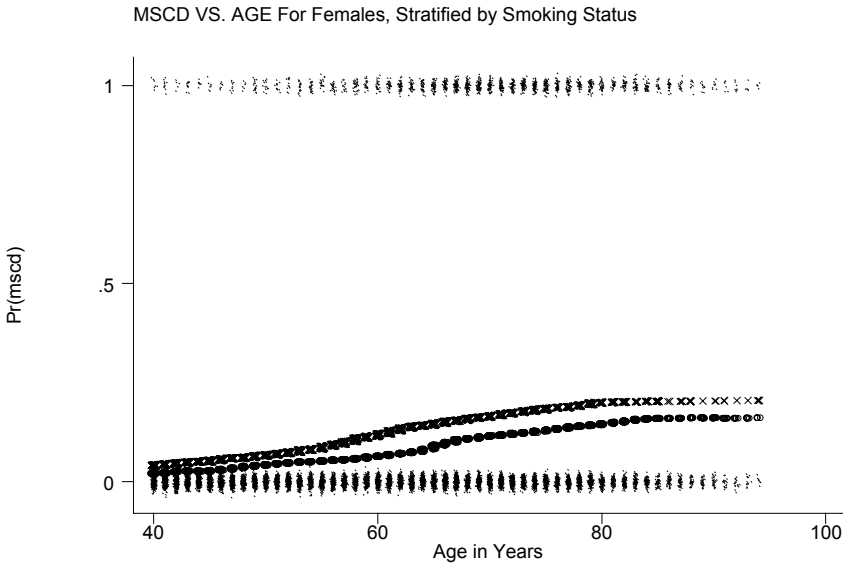


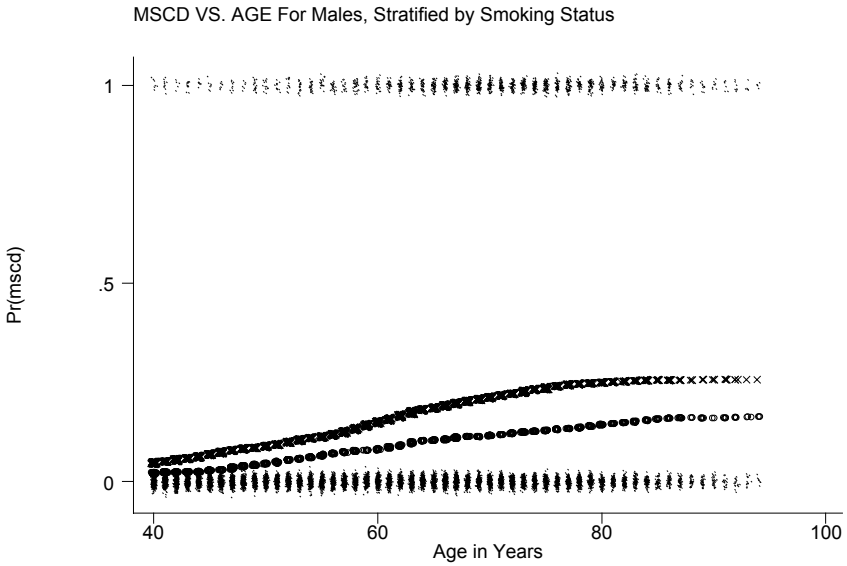
Question 2: Is there evidence in the data to suggest that the mscd/age relationship is different for men and women? Use the figures below and propose a regression model to address this question. Specify the statistical test that you would perform to answer the question.

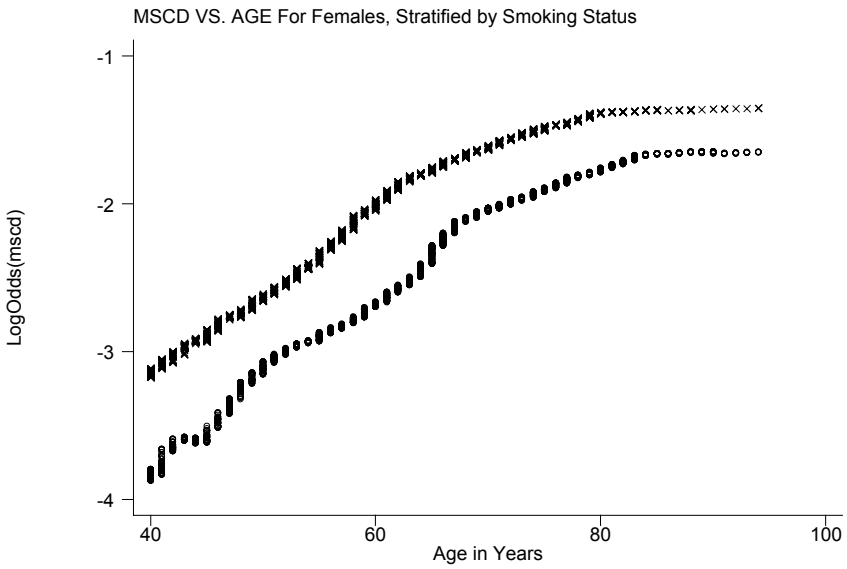


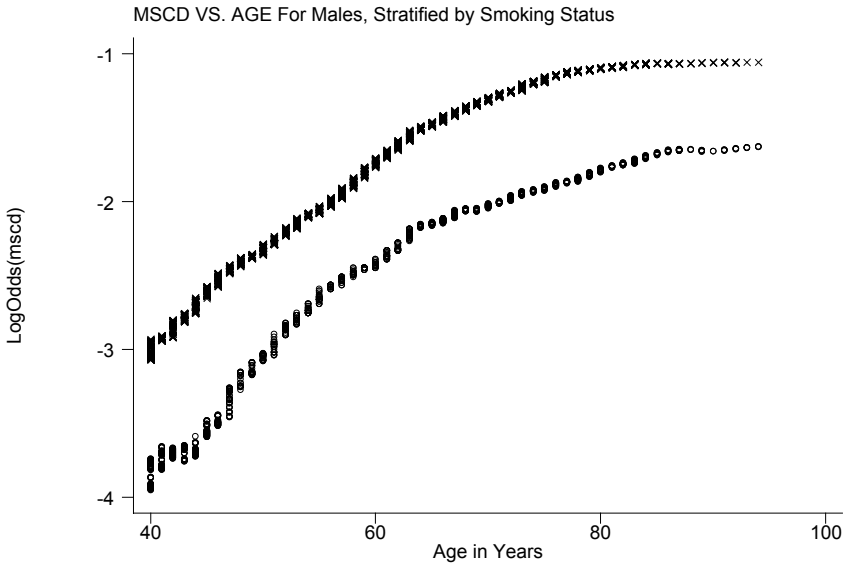


Question 3: Is there evidence in the data, that the mscd/age association is a) different for male smokers and male non-smokers and b) different for female smokers and female non-smokers? Using the figures below, specify an appropriate regression model to answer this question. What statistical tests would you perform to answer the questions a) and b)?









PART II:

Suppose now that you wish to use your model from question 3 to predict having a mscd as a function of age, gender and smoking status. Describe how would you assess whether you could improve the predictive power of your regression model by adding the available SES variables to the model?