Project for Regression Analysis in Public Health Solution
June 16 – June 27, 2003

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.

Solution to Question 1:
The model could be specified as follows:

$$\text{Logit}(\Pr(\text{mscd}=1)) = b0 + b1(\text{age}-65) + b2(\text{age}-65)^{+}$$

where $(\text{age}-65)^{+} = (\text{age}-65)$ if age > 65 and 0 otherwise.
NOTE: you could have made a more complicated non-linear function for age, but the spline with one knot at 65 years of age is fine.

The regression coefficients are interpreted as follows:

b0 = log odds of mscd for 65 year olds

b1 = log OR of mscd comparing persons who differ in age by one year and whom are ≤ 65 years of age

b1+b2 = log OR of mscd comparing persons who differ in age by one year and whom are over 65 years of age

b2 = difference in the log OR of mscd per year of age comparing persons ≤ 65 to those over 65.

TEST: To test to see if the mscd/age association is linear, we have to test if b2 = 0. This can be done using either a Wald test, or a likelihood ratio test.

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.

Solution to Question 2:
The model could be specified as follows:

$$\text{Logit}(\Pr(\text{mscd}=1)) = b0 + b1(\text{age}-65) + b2(\text{age}-65)^{+} + b3\text{Male} + b4\text{Male}*(\text{age}-65) + b5\text{Male}*(\text{age}-65)^{+}$$

where $(\text{age}-65)^{+} = (\text{age}-65)$ if age > 65 and 0 otherwise

The model for females is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = b0 + b1(\text{age}-65) + b2(\text{age}-65)^{+}$$

The model for males is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = (b0 + b3) + (b1+b4)(\text{age}-65) + (b2+b5)(\text{age}-65)^{+}$$

The regression coefficients are interpreted as follows:

b0 = log odds of mscd for 65 year old females

b1 = log OR of mscd comparing females who differ in age by one year and whom are ≤ 65 years of age

b1+b2 = log OR of mscd comparing females who differ in age by one year and whom are over 65 years of age

b2 = difference in the log OR of mscd per year of age comparing females ≤ 65 to those over 65.

b0+b3 = log odds of mscd for 65 year old males

b3 = log OR of mscd comparing 65 year old males to 65 year old females

b1+b4 = log OR of mscd comparing males who differ in age by one year and whom are ≤ 65 years of age

b4 = difference in log OR of mscd per year of age comparing males to females whom are ≤ 65 years of age

b1+b4+b2+b5 = log OR of mscd comparing males who differ in age by one year and whom are over 65 years of age

b4+b5 = difference in log OR of mscd per year of age comparing males to females whom are over 65 years of age

b5 = This is tricky! Difference of the differences in log OR of mscd per year of age for males and females comparing over 65 years of age to ≤ 65 years of age

TEST: To test to see if the mscd/age association is different for males and females, we have to test if b4 = 0 and b5 = 0. This can be done using either a Wald test or a likelihood ratio test.

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)?

Solution to Question 3:
The model could be specified as follows:

$$\text{Logit}(\Pr(\text{mscd}=1)) = b0 + b1(\text{age}-65) + b2(\text{age}-65)^{+} + b3\text{Male} + b4\text{Male}*(\text{age}-65) + b5\text{Male}*(\text{age}-65)^{+} + b6\text{eversmk} + b7*\text{eversmk}*\text{Male} + b8\text{eversmk}*(\text{age}-65) + b9\text{eversmk}(\text{age}-65)^{+} + b10\text{eversmk}*\text{Male}*(\text{age}-65) + b11\text{eversmk}*\text{Male}*(\text{age}-65)^{+}$$

where $(\text{age}-65)^{+} = (\text{age}-65)$ if age > 65 and 0 otherwise

The model for females is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = b0 + b1(\text{age}-65) + b2(\text{age}-65)^{+} + b6\text{eversmk} + b8\text{eversmk}*(\text{age}-65) + b9\text{eversmk}(\text{age}-65)^{+}$$

The model for female non-smokers is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = b0 + b1(\text{age}-65) + b2(\text{age}-65)^{+}$$

The model for female smokers is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = (b0+b6) + (b1+b8)(\text{age}-65) + (b2+b9)(\text{age}-65)^{+}$$

The model for males is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = (b0+b3) + (b1+b4)(\text{age}-65) + (b2+b5)(\text{age}-65)^{+} + (b6+b7)\text{eversmk} + (b8+b10)\text{eversmk}*(\text{age}-65) + (b9+b11)\text{eversmk}(\text{age}-65)^{+}$$

The model for male non-smokers is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = (b0+b3) + (b1+b4)(\text{age}-65) + (b2+b5)(\text{age}-65)^{+}$$

The model for male smokers is:

$$\text{Logit}(\Pr(\text{mscd}=1)) = (b0+b3+b6+b7) + (b1+b4+b8+b10)(\text{age}-65) + (b2+b5+b9+b11)(\text{age}-65)^{+}$$

The regression coefficients are interpreted as follows (It is sufficient here to make interpret the coefficients that you are interested in the most: comparing female smokers and non-smokers and comparing male smokers and non-smokers)

b0 = log odds of mscd for 65 year old female non-smokers

b1 = log OR of mscd comparing female non-smokers who differ in age by one year and whom are ≤ 65 years of age

b1+b2 = log OR of mscd comparing female non-smokers who differ in age by one year and whom are over 65 years of age

b0+b6 = log odds of mscd for 65 year old female smokers

b6 = log OR of mscd for 65 year old females comparing smokers to non-smokers

b1+b8 = log OR of mscd comparing female smokers who differ in age by one year and whom are ≤ 65 years of age

b8 = difference in the log OR of mscd per year of age comparing female smokers to female non-smokers whom are ≤ 65 years of age

b1+b2+b8+b9 = log OR of mscd comparing female smokers who differ in age by one year and whom are over 65 years of age

b8+b9 = difference in the log OR of mscd per year of age comparing female smokers to female non-smokers whom are over 65 years of age

b0+b3 = log odds of mscd for 65 year old male non-smokers

b1+b4 = log OR of mscd comparing male non-smokers who differ in age by one year and whom are ≤ 65 years of age

b1+b4+b2+b5 = log OR of mscd comparing male non-smokers who differ in age by one year and whom are 65 years of age

b2+b5 = difference in log OR of mscd per year of age comparing male non-smokers whom are ≤ 65 years of age to over 65 years of age

b0+b3+b6+b7 = log odds of mscd for 65 year old male smokers

b6+b7 = log OR comparing 65 year old male smokers to non-smokers

b1+b4+b8+b10 = log OR of mscd comparing male smokers who differ in age by one year and whom are ≤ 65 years of age

b8+b10 = difference in log OR of mscd per year of age comparing male smokers to non-smokers whom are ≤ 65 years of age

b1+b4+b8+b10+b2+b5+b9+b11 = log OR of mscd comparing male smokers who differ in age by one year and whom are over 65 years of age

b8+b9+b10+b11 = difference in log OR of mscd per year of age comparing male smokers to non-smokers whom are over 65 years of age

TEST: To test to see if there is evidence in the data, that the mscd/age association is
a) different for male smokers and male non-smokers: test b8=0, b9=0, b10=0, b11=0
b) different for female smokers and female non-smokers: test b8=0, b9=0
Both tests can be performed using either the Wald or likelihood ratio test.

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?

Solution: If we are talking about the "predictive power" of the regression model, we would need to consider one of the techniques that we discussed to use our regression model as a predictive tool.

First, we would fit the model from question 3 and an extended model that also includes categorical variables for each of the available SES variables.

We would then construct ROC curves for each of the two models. The model with higher "predictive power" has the largest area under the ROC curve.