Problem Set 1
Health Economic

Bill Evans
Spring 2013

The first problem set is an exercise in Stata to help those rusty in their programming skills and for those who need to learn some about Stata programming capacity.

If you need help with Stata, everything in this problem set can be answered with my introduction to Stata which can be found here http://www3.nd.edu/~wevans1/econ30331/Introduction%20to%20STATA.pdf .

The data for this problem set examines gradient in socio-economic status and health as measured by mortality.

The National Health Interview Survey (NHIS) is an annual survey conducted by the Department of Health and Human Services to measure the state of health in the country. The National Death Index is a master database of everyone that has died in the US including their Social Security number. Using detailed identifiable information about respondents to the NHIS, starting in the 1980s, respondents to the NHIS were matched to the National Death Index and the file then included a variable that identified when and if a person died within a followup period. The data also includes the cause of death. This merged data is called the NHIS Multiple Cause of Death data (MCOD). I have constructed a data set that has respondents aged 25-64 from the NHIS/MCOD data. Below is a table that describes the variables in the data set.

Codebook, nhis_mcod_data.dta

Variable	Description
diedin5	Dummy variable, =1 if the respondent died within 5 years of the survey, =0 otherwise
Male	Dummy variable, =1 if the respondent is male, =0 otherwise
Age	Age in years
Married	Dummy variable, =1 if the respondent is married, =0 otherwise
Race	Categorical variable for race and ethnicity. =1 if respondent is white, non-Hispanic, =2 if black, non-Hispanic, =3 if other race, non-Hispanic, =4 if Hispanic
Educ	Categorical variable for educational level. =1 if respondent has less than a high school degree, =2 if a high school degree, =3 if some college, =4 with a bachelor's degree or more
Incomeg	Categorical variable for family income. =1 if family income is ≤$10K, =2 if >$10K and ≤$20K, =3 if >$20K and ≤$30K, =4 if >$30K and ≤$40K, =5 if >$40K and ≤$50K, =6 if >$50K.
Bmi	Body mass index, weight in kg/ height in cm squared
Srhealth	Categorical variable for self-reported health status. =1 if excellent health, =2 if very good, =3 if good, =4 if fair =5 if poor.

1. Using the "tab" procedure, what fraction of the sample died in the 5 years after the survey?

2. Using the tab procedure, calculate the fraction of "diedin5" by education and income groups. Mortality rates should be declining with more education and income.

```
tab educ diedin5, row column
```

3. Redo problem 2 but compare mortality rates across racial and ethnic groups. Compare the diedin5 rate for those who are white, non-Hispanic and Hispanic. Does this result make sense?

4. Sort the data by education and income groups. Then, get the means of "diedin5" by educ and incomeg

```
sort educ incomeg
by educ incomeg: sum diedin5
```

Using these results, fill in the following table of "means for diedin5 by education and income"

Means of diedin5 by Education and Income

		Income variable					
		1	2	3	4	5	6
Education variable	1						
	2						
	3						
	4						

5. Construct dummy variables for race groups 2-4, income groups 2-6 and education groups 2-4. Next, run a regression of diedin5 on age, male, married, plus the race, income and education variables.
    a. Interpret the coefficient on age. As a person ages 10 years, what happens to five year mortality rates?
    b. Interpret the coefficient on married. Does this coefficient make sense?
    c. Look at the coefficients on the income dummy variables. Interpret the coefficient on the dummy for income group 6.
    d. Suppose one were interested in the impact of moving from income group 3 to income group 6. Using the results from the regression, calculate this value.
    e. Constrcuct dummy variables for the following four variables:
        i. Underweight is BMI<=19
        ii. Overweight is BMI>25 and BMI<=30
        iii. Obese is MI>30 and BMI<=35
        iv. Severlyobese is BMI>35.

    Add these four variables to the regression in part d. Interpret the coefficient on overweight. Do you feel silly for not eating more? Yes you do. Why is the coefficient on underweight such a large positive number? At the 95% confidence level, can you reject or not reject the null hypothesis that the coefficient on severlyoese is equal to zero?