

TUFTS UNIVERSITY
Medical School

Prof. M. Bianconi
Course CMBA0264-01; Statistics with Applications; Summer 2009
E-Mail: marcelo.bianconi@tufts.edu

Assignment IV

HANDED OUT July 23, 2009
HANDED IN July 28, 2009 by 6:00pm (via e-mail if possible)

Data sets and other materials posted at the course web page:
<http://www.tufts.edu/~mbiancon/CMBA0264-2009.html>

1. Data on the P1 worksheet of data for PS4. Data are a sample of 2,763 post-menopausal women with existing coronary heart disease (CHD). This is the Heart and Estrogen/Progestin Study (HERS), a clinical trial of hormone therapy for prevention of recurrent heart attack and death (Hulley et al, 1998):

variable name	type	format	label	variable label
age	byte	%9.0g		age in years
drinkany	byte	%9.0g	noyes	any current alcohol consumption
exercise	byte	%9.0g	noyes	exercise at least 3 times per week
diabetes	byte	%9.0g	noyes	diabetes
BMI	float	%9.0g		BMI (kg/m ²)
glucose	int	%9.0g		fasting glucose (mg/dl)

- (i) Use the command *summarize (summ)* to examine the basic statistical properties of the data. Select a sub-sample of individuals without diabetes: *drop if diabetes ==1*, the subsample should have 2,032 observations.
- (ii) Use the command *histogram* to graph histograms of the four variables and write a one line comment.
- (iii) Use the *sctest* command to test the equality of variances of average fasting glucose levels among women without diabetes according to exercise: *sctest glucose, by(exercise)*. Interpret the results.
- (iv) Based on your results in (iii), use the *ttest* command to test average fasting glucose levels among women without diabetes according to exercise: *ttest glucose, by(exercise)*. Interpret the results.
- (v) Use the twoway scatter command to graph *glucose* on *exercise*: *twoway (scatter glucose exercise)*. Write a one line comment.
- (vi) Use the command *regression (reg)* to run a linear regression of *glucose* on *exercise* with the robust qualifier. What are the results?
- (vii) Use the command *predict* to obtain $E[\text{glucose} \mid \text{exercise}]$ (*predict gluhat*). Then, use the command *twoway scatter* to graph the actual and predicted values.
- (viii) Are the variables body mass index, age and drink potential candidates for omitted variables? Speculate on the potential signs of the biases, if any.
- (ix) Use the command *regression (reg)* to run a linear regression of *glucose* on *exercise* and *bmi* with the robust qualifier. What are the results?
- (x) Use the command *pwcorr* with the *, sig* qualifier to estimate a correlation matrix of the independent variables. Speculate whether there is high correlation among the independent variables.
- (xi) Compare your answers here to your answers to the model of (vi) above. Speculate whether or not you think there is an omitted variable bias in model (iv).

- (xii) Use the command *regression (reg)* to run a linear regression of *glucose* on *exercise*, *bmi*, *drink*, *age* and $age2=age*age$ with the robust qualifier. What are the results?
- (xiii) Use the command *pwcorr* with the *, sig* qualifier to estimate a correlation matrix of the independent variables. Speculate whether there is high correlation among the independent variables.
- (xiv) Use the command *postgr3* to plot the effect of *age* on *glucose*: *postgr3 age, asis(age age2)*. Interpret.
- (xv) Use the command *postgr3* to plot the effect of *bmi* on *glucose* by *exercise*: *postgr3 bmi, by(exercise)*. Interpret.
- (xvi) Use the command *postgr3* to plot the effect of *drinkany* on *glucose* by *exercise*: *postgr3 drinkany, by(exercise)*. Interpret.
- (xvii) Briefly evaluate your analysis of the determinants of fastening glucose in this sample.

2. Data on the P2 worksheet of data for PS4. These are cross-country observations on life expectancy at birth from the United Nations Development Programme (UNDP). Please, provide the UNDP with a regression analysis of the determinants of life expectancy based on the data available.