Statistical Reasoning in Public Health 2009
Biostatistics 612, Homework #2

1. Suppose it is the year 1985 and you are doing research on the differences in wages earned by men and women in the U.S. workforce. You gain access to a data set that contains information on a random sample of 534 U.S. workers surveyed in 1985. The data set contains information about the hourly wage (in U.S. dollars) and the sex of each of the workers surveyed, as well as information about each worker's age, union membership, and type of occupation (collapsed into 8 different categories). You decide to use linear regression to estimate unadjusted differences in the mean hourly wage for female workers as compared to males, as well as adjusted gender-wage differences, adjusted for various other worker characteristics. Below find the estimated coefficient for sex, along with its standard error, from 4 different linear regression models, all which include sex as a predictor.

Linear Regression of Wages ($/hr) on Sex, and Other Predictors

MODEL	Predictors (*x*s) in Model	Estimated Regression Coefficient (Slope) for Sex (1 = Female, 0 = Male)	Standard Error of Slope for Sex
A	Sex	-2.1	0.44
B	Sex, Age	-2.2	0.43
C	Sex, Age, Union Membership	-2	0.43
D	Sex, Age, Union Membership, Job Type	-1.9	0.43

   a. What is the estimated unadjusted mean difference in hourly wages for females as compared to males? Give a 95% confidence interval for this difference. Write a sentence interpreting both the unadjusted mean difference and the corresponding confidence interval. *(3 points)*
   b. What is the estimated adjusted mean difference in hourly wages for females as compared to males, adjusting for age, union membership, and job type? Give a 95% confidence interval for this difference. Write a sentence interpreting both the adjusted mean difference and the corresponding confidence interval. *(3 points)*
   c. Comment on any disparities in the estimated mean difference in hourly wages between males and females in the four models whose results are listed above. Does it appear from these results that the wage/gender relationship in confounded by other worker characteristics such as worker age, membership in a union, and job type? Why or why not? *(3 points)*
   d. Use the results from Model D to estimate the mean difference in hourly wages for females, age 42, who are union members with manufacturing jobs, as compared to

42-year male union members with manufacturing jobs. (note that you have already done this in a previous portion of the problem – I am just trying to "drill into you" how to interpret multiple linear regression coefficients.) *(1 point)*

e. Does the given information allow you to assess whether the relationship between hourly wages and sex is modified by age? If not, what additional results would you need to see? *(2 points)*

2. Carotid artery intima-media thickness (IMT) is a measure of thickening in the arterial wall. Higher values are associated with the development of atherosclerosis (thickening and hardening of the arterial walls resulting in restricted blood flow). A study published in 2003 in the Journal of the American Medical Association[1] concerns potential risk factors associated with increased IMT. This study was a population based large study in Finland involving over 2,000 subjects. One of the analyses utilized multiple linear regression to assess the relationship between average IMT and other subject characteristics for persons in the age range 29-39 years. The results are presented in the following table taken directly from the article:

Table 3. Multivariable Model of the Relationships Between Current Risk Variables and Common Carotid Artery Intima-Media Thickness in Adults Aged 29 Through 39 Years (N = 2229)*

Risk Variable	Regression Coefficient†	SE	*P* Value
Male, sex	0.009	0.004	.02
Age	0.026	0.002	<.001
LDL-C	0.004	0.002	.06
Body mass index	0.011	0.002	<.001
Systolic blood pressure	0.010	0.002	<.001
Smoking (no/yes)	0.011	0.004	.004

Abbreviation: LDL-C, low-density lipoprotein cholesterol.
*Diastolic blood pressure was also a significant correlate of intima-media thickness (*P*<.001) when entered into the model instead of systolic blood pressure.
†Expressed in millimeters for a 5-unit change in age (year) and a 1-SD change in other continuous variables and for the presence or absence of smoking.

The above results estimate a model of the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 age + \hat{\beta}_3 LDL + \hat{\beta}_4 BMI + \hat{\beta}_5 SBP + \hat{\beta}_6 smoke$$

Where $\hat{y}$ is estimated mean IMT (in mm), and the predictors are in units described in the table footnotes.

a) According to the table footnotes, what unit did the authors use for age in the multiple linear regression? *(1 point)*

[1] Raitakari O et al. Cardiovascular Risk Factors in Childhood and Carotid Artery Intima-Media Thickness in Adulthood: The Cardiovascular Risk in Young Finns Study. (2003) Journal of the American Medical Association, Vol 290 No 17. 2277-2283.

b) What is the estimated mean difference in IMT for two groups of persons who differ by one-year in age, adjusted for the other predictors in the model? *(1 point)*
c) Estimated a 95% CI for the quantity estimated in part b. *(1 point)*
d) Interpret the slope of sex in words. *(1 point)*
e) Give a 95% CI for the slope of sex. *(1 point)*
f) What additional information would you need to assess whether the relationship between IMT and sex is confounded by at least some of the additional predictors from the given multiple linear regression? *(1 point)*
g) What additional information would you need to assess whether sex modifies the relationship between IMT and smoking, after adjusting for age, LDL, BMI and SBP? *(1 point)*
h) Given the above results, can you ascertain whether the linear relationship between IMT and the six predictors in the regression is strong? Why or why not? *(1 point)*
i) Suppose above results are used to compare average IMT between 39 year olds to 29 year old after adjusting for the other 5 predictors in the model– what would be the estimated average difference in IMT? Compute a 95% confidence interval for this difference. *(2 points)*
j) Would it be appropriate to use the above results to estimate the adjusted average difference in IMT levels for 80 year olds compared 70 year olds? Why or why not? *(1 point)*

3. Total lung capacity (TLC) is a key indicator of pulmonary function. TLC is important in lung transplantation because it is important for the donor's lungs to be similar to that of the recipient. We have data on pre-transplant TLC (liters) of 32 recipients of heart lung transplants, obtained by whole body plethysmography and their age (years, ranging from 11-52), sex (1=female, 0 =male), and height (cm, range 138-189) The data is also on a file and more details on how to access the data are on the course web page (see homework section of the web page). The necessary Stata commands for completing each part of this exercise appear at the end of this document. Also included on the course website is the Stata output you will get if you use the commands listed at the end of this document – so you may do this assignment without using Stata directly.
IMPORTANT: Please do not include any Stata output in your responses! If you wish to include graphics in your document, this is fine – however, it is also fine just to describe what you "see" in a graph where asked.

   a. Graph the relationship between TLC and age in a scatterplot. Comment on the nature of the relationship between TLC and age. *(1 point)*
   b. Perform a simple linear regression of TLC on age. Are the results consistent with what you saw in the scatterplots? Interpret the estimated coefficient (slope) of age in a sentence. Report a 95% confidence interval for the (true) coefficient of age for this population. *(3 points)*
   c. Graph the relationship between TLC and height in a scatterplot. Comment on the nature of the relationship between TLC and height. *(1 point)*
   d. Perform a simple linear regression of TLC on height. Are the results consistent with what you saw in the scatterplots? Interpret the estimated coefficient (slope) of height in a sentence. Report a 95% confidence interval for the (true) coefficient of height for this population. *(3 points)*

e. Graph the relationship between TLC and patient's sex in a scatterplot. Is this a useful exploratory approach for assessing gender differences in TLC? Can you suggest other ways of exploring the relationship between a continuous outcome and a binary predictor? *(1 point)*
f. Perform a simple linear regression of TLC on sex. Interpret the estimated coefficient (slope) of sex in a sentence. Report a 95% confidence interval for the (true) coefficient of sex for this population. *(3 points)*
g. Now perform a multiple linear regression of TLC on height, age, and sex together. Interpret the slope estimates for height, age, and sex in words. *(3 points)*
h. Which predictors are statistically significantly associated with TLC (α=.05) in the multiple linear regression model? *(1 point)*
i. Compare the unadjusted relationship between TLC and sex, to the height and age adjusted association between TLC and sex. Is there any suggestion of confounding? Why/why not? *(1 point)*
j. What is the R^2 value for the multiple regression model you fit in (g)? What is the the interpretation of this value? *(1 point)*
k. Using the regression model results from part (g), estimate the mean TLC level for:
    a. 42 year old males, 170 cm tall *(1 point)*
    b. 35 year old females, 145 cm tall *(1 point)*
l. Using the regression model results from part (g), estimate the mean difference in TLC between 50 year old females 150 cm tall, and 40 year old males 160 cm tall, *(1 point)*

Parts (j) and (k) are *extra credit!*

m. Create a scatterplot of TLC versus height separately for male and for females. Does the relationship between TLC and height appear similar for both sexes? (up to 2 point extra credit)
n. Run a regression of TLC on height, sex, and an interaction between sex and height. Based on this result: *(up to 3 points extra credit)*
    a. What is the estimated mean difference in TLC for two groups of men who differ by 1 cm in height?
    b. What is the estimated difference between two groups of women who differ by 1 cm in height.
    c. Is there a statistically significant interaction between sex and height?

Sample Quiz Questions: Choose the correct answer from the following multiple choice question. **Include a sentence or two justifying your answer choice.** ***(1 point for each correct answer, 1 point for correct justification)***

The objective of a study is to understand the factors that are associated with systolic blood pressure in infants. Systolic blood pressure, weight (ounces) and age (days) are measured in 100 infants. A multiple linear regression is performed to predict blood pressure (mm Hg) from age and weight. The following results are presented in a journal article. (Questions 3-5 refer to these results)

Multiple Linear Regression Analysis of the Predictors of Systolic Blood Pressure in Infants

	coefficients ($\hat{\beta}'s$)	SE of $\hat{\beta}'s$
Intercept	**50.**	**4.0**
Birth Weight	**0.10**	**0.3**
Age (days)	**4.0**	**0.60**

4. How much higher would you expect the blood pressure to be of an infant who weighed 120 ounces compared to an infant who weighed 90 ounces if both infants were of exactly the same age?

   a. 0.1 mm Hg
   b 1.0 mm Hg
   c. 2.0 mm Hg
   d. 3.0 mm Hg
   e. 4.0 mm Hg

5. Which of the following is a 95% confidence for the difference in SBP between two infants of the same weight who differ by 2 days in age (older compared to younger)?

   a. 2.8 mmHg to 5.2 mmHg
   b. 5.6 mmHg to 10.4 mmHg
   c. 6.8 mmHg to 9.6 mmHg
   d. –0.5 mmHg to 0.7 mmHg

6. Suppose the R^2 from the above regression model is .57, which means that roughly 57% of the variability in the infant's blood pressure measurements is explained by infant's age and weight. What would happen to this R^2 value, if weight had been recorded as kilograms instead of ounces?

   a. R^2 would go increase.
   b. R^2 would decrease
   c. R^2 would equal .57.
   d. Not enough information to determine.

Appendix : Stata Commands for Problem 3

If you are using STATA, the commands are as follows:

a. twoway (scatter tlc age)
b. regress tlc age
c. twoway (scatter tlc height)
d. regress tlc height
e. twoway (scatter tlc sex)
f. regress tlc sex
g. regress tlc age height sex
h – l . no command necessary
m. twoway (scatter tlc height) if sex == 0
twoway (scatter tlc height) if sex == 1
n. This is up to you to figure out if you want the extra credit. You can find information on how to do this in the notes.

note that the "==" in commands *j* and *i* is actually two, adjacent equals (=) signs.