EPIDEMIOLOGY 340.600: STATA PROGRAMMING AND DATA MANAGEMENT

Assignment 1

Due date: 11:59 p.m., Monday, April 19, 2021 via CoursePlus dropbox

Overview: write a .do file which imports data from `assignment1_data_2021.txt` (available on CoursePlus) and performs the data cleaning / exploratory data analysis tasks described below. Your .do file must be called assignment1_*yourname*.do (for example: assignment1_allanmassie.do) and must create a log file called assignment1_*yourname*.log. Your .do file should follow conventions for .do file structure described in class. Make sure your script will run on our machines, even if we are using a different version of Stata. Do not submit your log files as part of the assignment.

This dataset contains simulated data on registrants for the deceased donor kidney waitlist.

Evaluation: Evaluation will be based on the log file produced by your script. ***Your script will be run on a different dataset, in the same format as assignment1_raw.txt but containing different data.*** Partial credit will be awarded if the output is wrong, so have your script print *something* for every question. Make sure the output includes the question number as indicated. Make sure to follow the coding guidelines from class; for example, your script should include comments.

Note – test your code repeatedly to make sure it doesn't crash! Points will be deducted if your code crashes.

Codebook:

Variable	Description	Values
transplants.dta		
fake_id	Patient ID	Numeric
bmi	Body mass index (kg/m^2)	Numeric
dx	Primary diagnosis	String (see dataset)
init_age	Age	Numeric
prev	History of previous transplant	0=No / 1=Yes
peak_pra	Peak PRA	Numeric
female	Sex	0=No / 1=Yes
received_kt	Patient eventually received a kidney transplant?	0=No / 1=Yes

Question 1. Print the following sentence: "`Question 1: There are xxxx records in the dataset`." The "xxxx" should be replaced with the actual number of records in the dataset.

Question 2. Print the following sentence: "`Question 2: The median [IQR] age is XX [XX-XX] among males and XX [XX-XX] among females.`" IQR stands for interquartile range. The "XX" values should be replaced with correct values, rounded to the nearest whole number (for example, "47"). Ignore missing values; that is, calculate median [IQR] only using the non-missing values.

Question 3. Print the following sentence: "`Question 3: XX.X% among males and XX.X% among females have history of previous transplant.`" The "XX.X%" values should be replaced with correct percentage values, with <u>one decimal value</u> to the right of the decimal place (for example, "10.5%").

Question 4. Create a new numeric variable named `htn`, which takes a value of 1 if the variable `dx` has a value of "`4=Hypertensive`", and 0 otherwise. This variable should have value labels, "`Yes`" for 1 and "`No`" for 0. (Hint: what is the data type of this new variable? Is it a numeric or a string?) Run `tab htn` in your assignment do-file so that it would print the following result.

```
        htn |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |       XXXX        XXXX        XXXX
        Yes |       XXXX        XXXX        XXXX
------------+-----------------------------------
      Total |       XXXX        XXXX
```

Question 5. Now you have all the skills to create your first automated Table 1! Write a program called `question5` that prints the following table (including "`Question 5`" in the header). The "XX" values should be replaced with correct values found in the dataset, and should be rounded to the nearest whole number for age and to <u>one decimal place</u> to the right of the decimal point for other variables. Make sure the summary statistics are vertically aligned and justified along the left margin. Run your program and display the table.

```
Question 5                  Males (N=XX)     Females (N=XX)
Age, median (IQR)           XX (XX-XX)       XX (XX-XX)
Previous transplant, %      XX.X%            XX.X%
Cause of ESRD:
Glomerular, %               XX.X%            XX.X%
Diabetes, %                 XX.X%            XX.X%
PKD, %                      XX.X%            XX.X%
Hypertensive, %             XX.X%            XX.X%
Renovascular, %             XX.X%            XX.X%
Congenital, %               XX.X%            XX.X%
Tubulo, %                   XX.X%            XX.X%
Neoplasm, %                 XX.X%            XX.X%
Other, %                    XX.X%            XX.X%
```

Hint 1: Use `quietly {}` and `noisily` to display the lines nicely without interruptions.

Hint 2: init_age has some missing values. Exclude observations with missing init_age only when calculating the median [IQR] of init_age, but not when calculating the proportion with previous transplants.

Hint 3: some techniques taught in lecture 3 may be helpful for completing this table. You are welcome Lecture 3 techniques if you like, but they are not required.

Question 6. Your research group is investigating demographic characteristics associated with receiving a kidney transplant for waitlisted patients. You run a logistic regression using the following command:
`logistic received_kt init_age female`

Print a summary table as shown below, with odds ratios (OR) and 95% confidence intervals (CI). The "XXXX" values should be replaced with the actual values found in the dataset, and should be displayed with **<u>two</u>** decimal places to the right of the decimal point.

```
Question 6
Variable         OR    (95% CI)
Age              X.XX  (X.XX-X.XX)
Female           X.XX  (X.XX-X.XX)
```

Hint: If you like, you may these expressions below after logistic regression to obtain the odds ratio and 95% CI. We will use `init_age` as an example.

Odds ratio	`exp(_b[init_age])`
Lower bound of 95% CI	`exp(_b[init_age]+invnormal(0.025)*_se[init_age])`
Upper bound of 95% CI	`exp(_b[init_age]+invnormal(0.975)*_se[init_age])`

Question 7. The logistic regression you ran in Question 6 may not include all observations in the study dataset. Stata drops observations that have missing values in any of the variables included in the regression.

Using the regression results from Question 6, print the following text: "`Question 7: This regression included XXXX observations whereas the study dataset has XXXX observations in total.`" Replace "XXXX" with correct values.

Hint: use one of the e-class scalars. We talked about e-class scalars in Lecture 2.

Question 8. Print the following text: "`Question 8: I estimate that it took me XXXX hours to complete this assignment.`" For example, if it took you six hours, your .do file will contain the line

```
disp "Question 8: I estimate that it took me 6 hours to complete this assignment."
```

Give an honest answer; this is just for our data collection purposes. Everyone who answers this question will receive full credit for this question. <u>However, this question is worth some points, so don't skip it!</u>