Econ 5253 - Spring 2022	Problem Set 7	Due: Mar. 29 beginning of class

This problem set will give you practice with imputing missing data and automating the process of creating reports of summary tables and model estimates.

As with the previous problem sets, you will submit this problem set by pushing the document to *your* (private) fork of the class repository. You will put this and all other problem sets in the path `/DScourseS22/ProblemSets/PS7/` and name the file `PS7_LastName.*`. Your OSCER home directory and GitHub repository should be perfectly in sync, such that I should be able to find these materials by looking in either place. Your directory should contain at least three files:

- `PS7_LastName.R` (you can also do this in Python or Julia if you prefer)
- `PS7_LastName.tex`
- `PS7_LastName.pdf`

1. Type `git pull origin master` from your OSCER `DScourseS22` folder to make sure your OSCER folder is synchronized with your GitHub repository.
2. Synchronize your fork with the class repository by doing a `git fetch upstream` and then merging the resulting branch.
3. Install the following R packages (if you don't already have them installed):
   - `mice`
   - `modelsummary`

   The first two packages are useful for imputing missing data. modelsummary is useful to researchers by taking summary statistics tables or output from statistical models and converting them to LaTeXtables automatically. (If you'd like to use Python for this assignment, the Python version of `modelsummary` is the `summary_col` function of the library `statsmodels.api`. I don't know of a package like this that exists in Julia.)
4. Using R or Python, load the file `wages.csv` (located in the current folder) in as a data frame. This data set contains information on $\approx 2,250$ women who were working in the US in 1988. The variables should be self-explanatory, except for `tenure`, which refers to how long (in years) each woman has been at her current employer, and `hgc`, which refers to how many years of schooling each woman has completed.
5. Drop observations where either `hgc` or `tenure` are missing.
6. Use `modelsummary` to produce a summary table of this data frame.

If you have never used this package before, consult the online documentation available here. The package will output LaTeXcode in the R console which you can copy and paste into your writeup for this homework. You can also have it write directly to a separate file and then include that in your document.

At what rate are log wages missing? Do you think the `logwage` variable is most likely to be MCAR, MAR, or MNAR?

7. Perform the following imputation methods for missing `logwage` observations. **For each imputation method, estimate the following linear regression model:**

$$logwage_i = \beta_0 + \beta_1 hgc_i + \beta_2 college_i + \beta_3 tenure_i + \beta_4 tenure_i^2 + \beta_5 age_i + \beta_6 married_i + \varepsilon_i$$

Our coefficient of interest is β_1 which can be interpreted as the returns to schooling (where schooling is thought of as an "investment" in "human captial"—we are curious what the Return On Investment is).

- estimate the regression using only complete cases (i.e. do listwise deletion on the log wage variable ... this assumes log wages are Missing Completely At Random)
- perform mean imputation to fill in missing log wages
- impute missing log wages as their predicted values from the complete cases regression above (i.e. this would be consistent with the "Missing at Random" assumption)
- use the `mice` package to perform a multiple imputation regression model (follow the steps here)

Once you have finished all of this, use `modelsummary` to create one regression table which has the estimates of the four regression models. Include this table in your .tex writeup.

The true value of $\hat{\beta}_1 = 0.093$. Comment on the differences of $\hat{\beta}_1$ across the models. What patterns do you see? What can you conclude about the veracity of the various imputation methods? Also discuss what the estimates of $\hat{\beta}_1$ are for the last two methods.

8. Tell me about the progress you've made on your project. What data are you using? What kinds of modeling approaches do you think you're going to take?

9. Compile your .tex file, download the PDF and .tex file, and transfer it to your cloned repository on OSCER using your SFTP client of choice (or via `scp` from your laptop terminal). You may also copy and paste your .tex file from your browser directly into your terminal via `nano` if you prefer, but you will need to use SFTP or `scp` to transer the PDF.[1]

10. You should turn in the following files: .tex, .pdf, and any additional scripts (e.g. .R, .py, or .jl) required to reproduce your work. Make sure that these files each have the correct naming convention (see top of this problem set for directions) and are located in the correct directory (i.e. `~/DScourseS22/ProblemSets/PS7`).

11. Synchronize your local git repository (in your OSCER home directory) with your GitHub fork by using the commands in Problem Set 2 (i.e. `git add`, `git commit -m "message"`, and `git push origin master`). More simply, you may also just go to your fork on GitHub and click the button that says "Fetch upstream." Then make sure to pull any changes to your local copy of the fork. Once you have done this, issue a `git pull` from the location of your other local git repository (e.g. on your personal computer). Verify that the PS7 files appear in the appropriate place in your other local repository.

[1] If you want to try out something different, you can compile your .tex file on OSCER by typing `pdflatex myfile.tex` at the command prompt of the appropriate directory. This will create the PDF directly on OSCER, removing the requirement to use SFTP or `scp` to move the file over.