Problem Set 5
(Due on Wednesday, Nov 11th, 11:59pm)

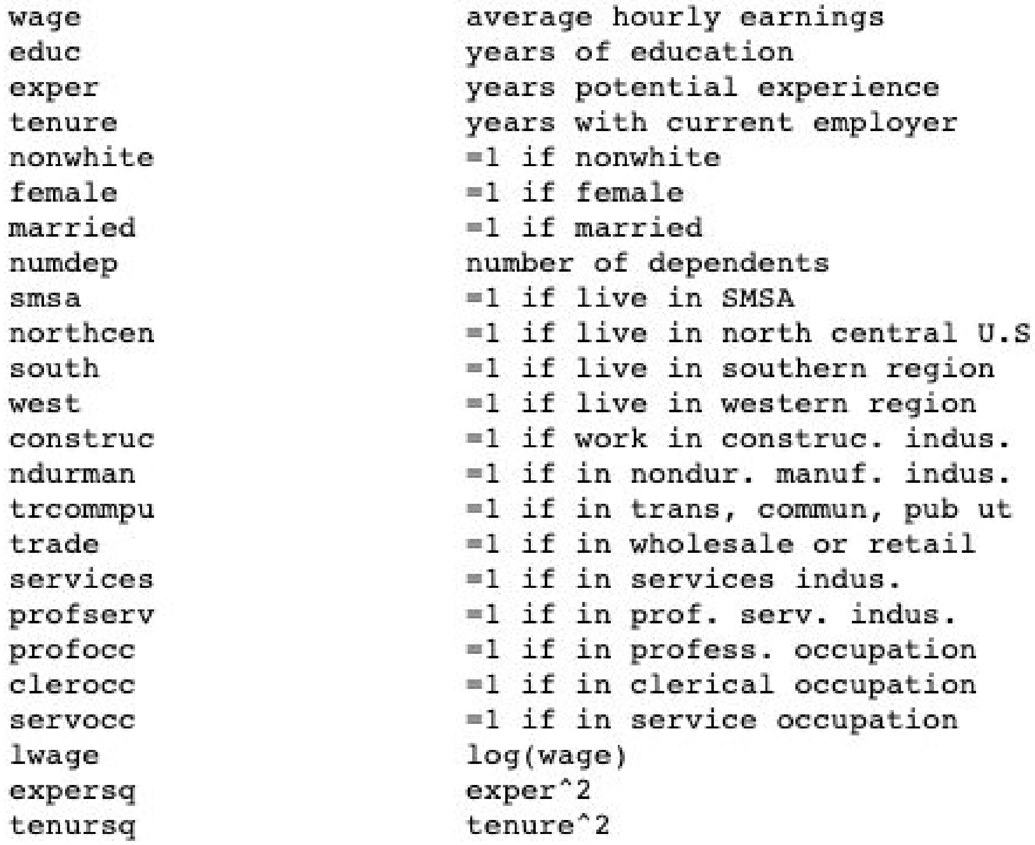
ECON 406: Data Science Computing for Economics

Expectations and deliverables:

- A pdf : including your code, graphs, and your explanation to each question. Your code should be attached at the end of the file. Make sure you upload a pdf file instead of a doc file. Points may be deducted if it's not a pdf file.
- A python file name: **regression.py**
  - If Canvas changes your file name. It's fine. Don't worry about it.
  - The formatting will be checked with pylint, and points will be deducted if the formatting is wrong.If you think there is some problem with the pylint error, please contact us.
- No zip file is needed. No git repository is needed.
- This homework will be focusing on practice OLS Regression and Logistic Regression.
- For problems that require you to run some type of regression, use statsmodels (you may choose whether to use sm or smf). Besides that, you are free to use any packages you feel are needed.
- When loading files in your code, assume they are in the same directory as your code. For example, if you are loading a .csv file called my_data.csv, your pandas call would be "pd.read_csv('my_data.csv')", not something like, "pd.read_csv('~/Documents/UW/SickEconClass/EvenSickerAssignment/my_data.csv')"

Exercise 1 : Wage

You will be given the dataset “**wage.csv”** to better understand the impact of different variables on expected wage rates**.** Please finish the following exercise with this dataset. It’s a cross-sectional dataset on wages. In doing this exercise, you should make one function, called “first_exercise”, that generates all your output. Note: this function should not take any arguments.



wage	average hourly earnings
educ	years of education
exper	years potential experience
tenure	years with current employer
nonwhite	=1 if nonwhite
female	=1 if female
married	=1 if married
numdep	number of dependents
smsa	=1 if live in SMSA
northcen	=1 if live in north central U.S
south	=1 if live in southern region
west	=1 if live in western region
construc	=1 if work in construc. indus.
ndurman	=1 if in nondur. manuf. indus.
trcommpu	=1 if in trans, commun, pub ut
trade	=1 if in wholesale or retail
services	=1 if in services indus.
profserv	=1 if in prof. serv. indus.
profocc	=1 if in profess. occupation
clerocc	=1 if in clerical occupation
servocc	=1 if in service occupation
lwage	log(wage)
expersq	exper^2
tenursq	tenure^2

1. Prep your data: this should include loading the data and making sure it’s ready for analysis (deal with missing variables, generate any transformed variables if needed, etc.).

2. Data visualization: make at least one plot with one or more variables in the dataset. These can be any plots you think are appropriate to help this exercise. One possibility includes using a scatterplot to see whether any obvious relationship exists between wage and years of education. You could do similar plots of education, experience, or tenure against logged wages.

3. Based on the visualization exercise and what you know about the variables, do you think OLS or Logistic Regression is more suitable for understanding what factors explain variability in wages？And why?

4. Propose a data generating process for wages. Write out your proposed model, use any specification you think is valid. Note that wages (or some function of wages) should be on the left hand side. A combination of additional variables, coefficients, and an error term should be on the right hand side. And don't forget the subscripts!

5. Estimate the model you proposed in problem 1.4 above. Show the regression table.

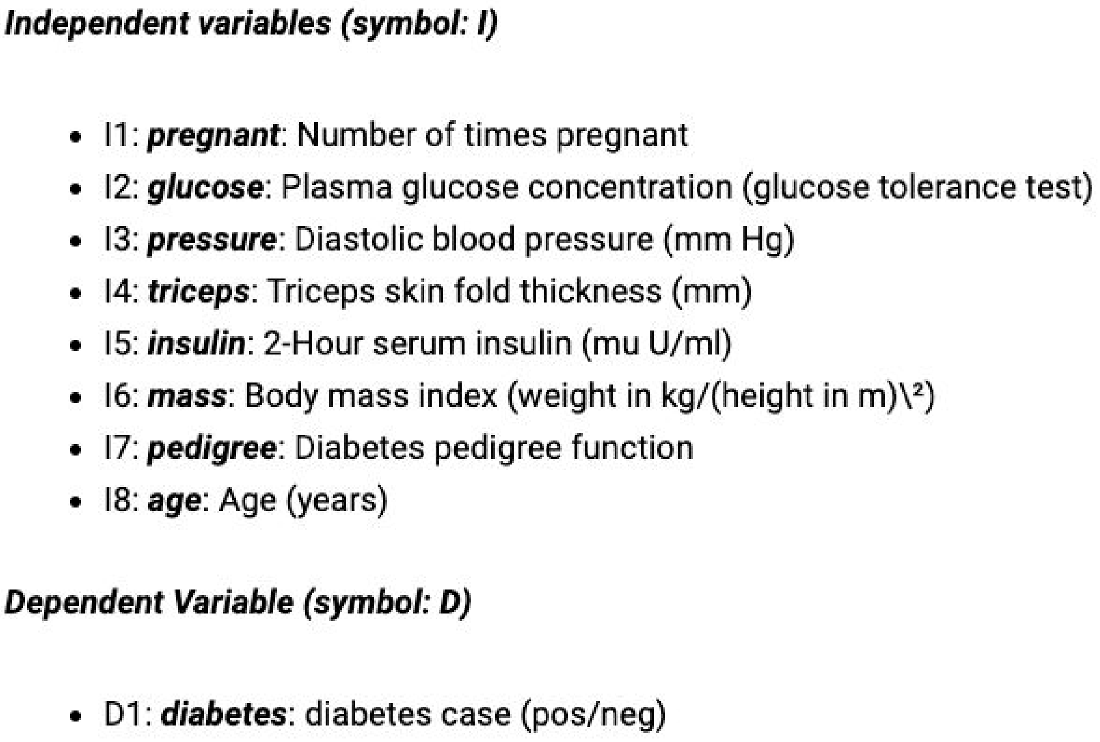
6. Interpret each coefficient. And are they statistically significant at $\alpha = 0.05$?

7. How much of the variation of wages for different individuals can be explained by the variable you choose for your model? (How well your model will do at prediction?)

8. Given your estimated model, give an example of a type of person that could expect to have hourly wages of $150 (example, how much education, experience, tenure, etc.).

Exercise 2: Diabetes

You will be provided with a dataset "diabetes.csv".The objective of the data set is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements (independent variables) included in the data set. In doing this exercise, you should make one function, called "second_exercise", that generates all your output. Note: this function should not take any arguments.



Independent variables (symbol: I)

- I1: ***pregnant***: Number of times pregnant
- I2: ***glucose***: Plasma glucose concentration (glucose tolerance test)
- I3: ***pressure***: Diastolic blood pressure (mm Hg)
- I4: ***triceps***: Triceps skin fold thickness (mm)
- I5: ***insulin***: 2-Hour serum insulin (mu U/ml)
- I6: ***mass***: Body mass index (weight in kg/(height in m)$\backslash^{2}$)
- I7: ***pedigree***: Diabetes pedigree function
- I8: ***age***: Age (years)

Dependent Variable (symbol: D)

- D1: ***diabetes***: diabetes case (pos/neg)

1. Prep your data: this should include loading the data and making sure it's ready for analysis (deal with missing variables, generate any transformed variables if needed, etc.). As a part of this, you will need to convert the dependent variable into something that can be used by statsmodels [Hint: convert string into integers by mapping neg: 0 and pos: 1 using the .map( ) method]

2. Data visualization: make at least one plot with one or more variables in the dataset. These can be any plots you think are appropriate to help this exercise. For example, you could visualize how the probability of having diabetes changes with the pedigree label. In this case, "pedigree" would be plotted on x-axis and "diabetes" on the y-axis. Consider comparing an LPM plot to a Logistic regression plot.

3. Based on your plots and your understanding of the data, do you think OLS or Logistic Regression is more suitable to analyze this problem? And why?

4. Propose a data generating process for diabetes. Write out your proposed model, use any specification you think is valid. If you decide on OLS, use the regular regression model. If you decide on Logistic Regression, use the sigmoid function.

5. Estimate the model you proposed in 2.4 and show the regression table.

6. Interpret and compare each coefficient.

7. Consider a patient who has the median value of each of your independent variables (in the 50th percentile). What is this patient’s probability of getting diabetes? How much more/less likely is it for a patient who is in the 75th percentile of each of your independent variables? What about for a patient in the 25th percentile?