Assignment 2

Fulin Guo

1. Imputing age and gender

(a) First, conduct OLS regression based on observations from SurveyIncome.txt:

$$\begin{aligned} age &= \beta_0 + \beta_1 tot\_inc + \beta_2 wgt + \varepsilon_1 \\ female &= \eta_0 + \eta_1 tot\_inc + \eta_2 wgt + \varepsilon_2 \end{aligned} \qquad (1)$$

Second, we can use the above equations to impute age and gender variable in BestIncome.txt based on labor income, capital income and weight variable in that dataset:

For each individual i:

$$\begin{aligned} age_i &= \beta_0 + \beta_1 (lab\_inc_i + cap\_inc_i) + \beta_2 wgt_i \\ female_i &= \eta_0 + \eta_1 (lab\_inc_i + cap\_inc_i) + \eta_2 wgt_i \end{aligned} \qquad (2)$$

(We use the assumption that for each individual, total income equals to labor income plus capital income).

Lastly, we should ensure that the female variable be either 0 or 1 by making the following transformation:

If $female_i \geq 0.5$, then $female_i = 1$

If $female_i < 0.5$, then $female_i = 0$ (3)

(b) First, we conduct OLS regression for SurveyIncome.txt data, obtaining:

$$\begin{aligned} age &= 44.2097 + 2.52 * 10^{-5} * tot\_inc - 0.0067 wgt + \varepsilon_1 \\ female &= 3.7611 - 5.25 * 10^{-6} * tot\_inc - 0.0195 wgt + \varepsilon_2 \end{aligned} \qquad (4)$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    age   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.6326
Date:                Tue, 16 Oct 2018   Prob (F-statistic):              0.531
Time:                        19:47:00   Log-Likelihood:                -3199.4
No. Observations:                1000   AIC:                             6405.
Df Residuals:                     997   BIC:                             6419.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         44.2097      1.490     29.666      0.000      41.285      47.134
tot_inc      2.52e-05   2.26e-05      1.114      0.266   -1.92e-05    6.96e-05
wgt           -0.0067      0.010     -0.686      0.493      -0.026       0.013
==============================================================================
Omnibus:                        2.460   Durbin-Watson:                   1.921
Prob(Omnibus):                  0.292   Jarque-Bera (JB):                2.322
Skew:                          -0.109   Prob(JB):                        0.313
Kurtosis:                       3.092   Cond. No.                     5.20e+05
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 female   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     2513.
Date:                Tue, 16 Oct 2018   Prob (F-statistic):               0.00
Time:                        19:47:00   Log-Likelihood:                 173.49
No. Observations:                1000   AIC:                            -341.0
Df Residuals:                     997   BIC:                            -326.3
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.7611      0.051     73.600      0.000       3.661       3.861
tot_inc     -5.25e-06   7.76e-07     -6.765      0.000   -6.77e-06   -3.73e-06
wgt           -0.0195      0.000    -58.098      0.000      -0.020      -0.019
==============================================================================
Omnibus:                        0.170   Durbin-Watson:                   1.634
Prob(Omnibus):                  0.918   Jarque-Bera (JB):                0.114
Skew:                          -0.022   Prob(JB):                        0.945
Kurtosis:                       3.029   Cond. No.                     5.20e+05
==============================================================================
```

Second, we use equation (4) to input age and female variable in BestIncome.txt data.

$$\begin{aligned} age_i &= 44.2097 + 2.52*10^{-5}*(lab\_inc_i + cap\_inc_i) - 0.0067wgt_i \\ female_i &= 3.7611 - 5.25*10^{-6}*(lab\_inc_i + cap\_inc_i) - 0.0195wgt_i \end{aligned} \quad (5)$$

Lastly,

If $female_i \geq 0.5$, then $female_i = 1$

If $female_i < 0.5$, then $female_i = 0$ (6)

The first five rows in SurveyIncome.txt data after imputing are showed as follows:

```
   imputed_age  imputed_female
0    44.745897               0
1    45.157777               0
2    44.745701               0
3    44.919024               0
4    44.554687               0
```

(c) Using python, I obtain the descriptive statistics of imputed age and gender variables.

	Imputed age	Imputed female
Mean	44.8940	0.4705
Standard deviation	0.21907	0.4992
Minimum	43.9800	0.0000
Maximum	45.7068	1.0000
Number of observations	10,0000	10,000

Results in python:

```
[5 rows x 6 columns]
count    10000.000000
mean        44.894036
std          0.219066
min         43.980016
25%         44.747065
50%         44.890281
75%         45.042239
max         45.706849
Name: imputed_age, dtype: float64
count    10000.000000
mean         0.470500
std          0.499154
min          0.000000
25%          0.000000
50%          0.000000
75%          1.000000
max          1.000000
Name: imputed_female, dtype: float64
```

(d) I obtain the correlation matrix for the six variables in the BestIncome.txt data by coding in python.

The correlation matrix for the six variables:

	lab_inc	cap_inc	hgt	wgt	imputed_age	imputed_female
lab_inc	1.000000	0.005325	0.002790	0.004507	0.924329	-0.164857
cap_inc	0.005325	1.000000	0.021572	0.006299	0.234234	-0.046594
hgt	0.002790	0.021572	1.000000	0.172103	-0.044927	-0.134172
wgt	0.004507	0.006299	0.172103	1.000000	-0.299395	-0.778537
imputed_age	0.924329	0.234234	-0.044927	-0.299395	1.000000	0.074288
imputed_female	-0.164857	-0.046594	-0.134172	-0.778537	0.074288	1.000000

2. Stationarity and data drift

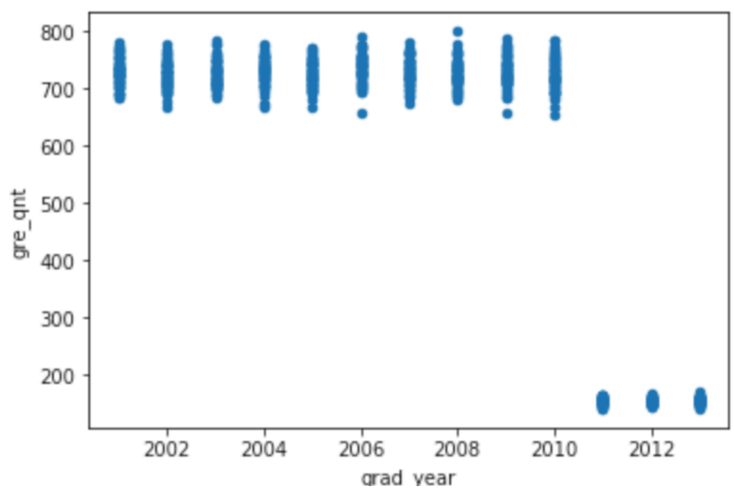
(a) The dependent variable is salary_p4, and the explanatory variable is gre_gnt. Coding in python, I obtain the following outcomes.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              salary_p4   R-squared:                       0.263
Model:                            OLS   Adj. R-squared:                  0.262
Method:                 Least Squares   F-statistic:                     356.3
Date:                Tue, 16 Oct 2018   Prob (F-statistic):           3.43e-68
Time:                        19:57:42   Log-Likelihood:                -10673.
No. Observations:                1000   AIC:                         2.135e+04
Df Residuals:                     998   BIC:                         2.136e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8.954e+04    878.764    101.895      0.000    8.78e+04    9.13e+04
gre_qnt      -25.7632      1.365    -18.875      0.000     -28.442     -23.085
==============================================================================
Omnibus:                        9.118   Durbin-Watson:                   1.424
Prob(Omnibus):                  0.010   Jarque-Bera (JB):                9.100
Skew:                           0.230   Prob(JB):                       0.0106
Kurtosis:                       3.077   Cond. No.                     1.71e+03
==============================================================================
```

The estimated β_0 is 8.95*10^4, the standard errors of β_0 is 878.764.

The estimated β_1 is -25.7632, the standard errors of β_1 is 1.365. The p-value is 0.000.

(b) The scatter of GRE and graduation year is:



The problem is that the gre_gnt variable has an obvious relationship with grad_year. That is, the gre_gnt variable is between 600 and 800 when grad_year is smaller than 2011 and less than 200 when grad_year is larger than or equal to 2011. This is a potential problem when testing the hypothesis since the correlation between gre_qnt and salary_p4 showed in part(a) might just because the two variables both have a correlation with the third variable grad_year and there is no causal correlation between gre_qnt and salary_p4. This is a biased estimation.

To solve this problem, I remove the influence of grad_year on gre_qnt by conducting the OLS regression:

$$gre\_qnt = \alpha_0 + \alpha_1 after\_2011 + \varepsilon$$

$after\_2011 = 1$ if grad_year is larger or equal to 2011. Otherwise, $after\_2011 = 0$

The result of this regression is:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                gre_qnt   R-squared:                       0.993
Model:                            OLS   Adj. R-squared:                  0.993
Method:                 Least Squares   F-statistic:                 1.363e+05
Date:                Tue, 16 Oct 2018   Prob (F-statistic):               0.00
Time:                        20:17:43   Log-Likelihood:                -4446.7
No. Observations:                1000   AIC:                             8897.
Df Residuals:                     998   BIC:                             8907.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
aft_2011    -573.5272      1.553   -369.222      0.000    -576.575    -570.479
const        728.4214      0.745    977.806      0.000     726.960     729.883
==============================================================================
Omnibus:                       10.161   Durbin-Watson:                   1.952
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               15.242
Skew:                          -0.003   Prob(JB):                     0.000490
Kurtosis:                       3.605   Cond. No.                         2.53
==============================================================================
```

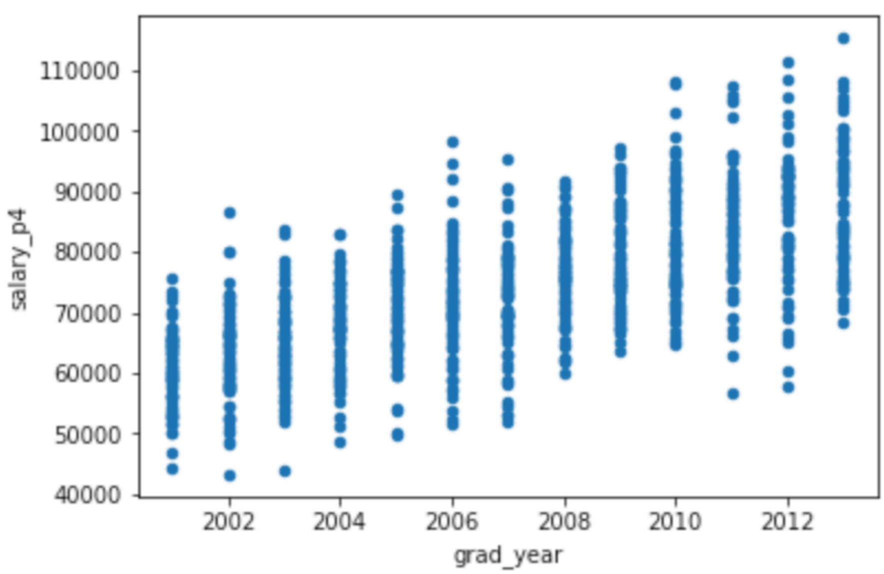
Then I update the gre_qnt:
If the grad_year is larger or equal to 2011, updated_gre = gre_qnt (do not change).
If the grad_year is smaller than 2011, updated_gre = gre_qnt – 573.5272

(In this method, I actually make all gre_qnt has the scoring scale between 130 and 170. However, another way to eliminate the problem is to get the residual of the above regression, which also eliminates the impact of grad_year on gre_qnt)

Here is the first five rows of the updated_gre variable.

	updated_gre
0	166.211872
1	148.286473
2	162.752708
3	196.973285
4	161.477661

(c) The scatter of income four years after graduation and graduation year is:



The problem is that the salary_p4 variable has an obvious relationship with grad_year variable. That is, the salary_p4 variable has an increasing trend. This is a potential problem when testing the hypothesis since the correlation between gre_qnt and salary_p4 showed in part(a) might just because the two variables both have a correlation with the third variable grad_year and there is no causal correlation between gre_qnt and salary_p4, which results in a biased estimation.

To solve this problem, I remove the influence of grad_year on salary_p4 by estimating the following OLS regression:

$$\ln(salary\_p4) = \delta_0 + \delta_1 grad\_year + \varepsilon$$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.490
Model:	OLS	Adj. R-squared:	0.489
Method:	Least Squares	F-statistic:	957.7
Date:	Tue, 16 Oct 2018	Prob (F-statistic):	5.73e-148
Time:	20:49:59	Log-Likelihood:	720.33
No. Observations:	1000	AIC:	-1437.
Df Residuals:	998	BIC:	-1427.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
grad_year	0.0309	0.001	30.947	0.000	0.029	0.033
const	-50.7169	2.001	-25.349	0.000	-54.643	-46.791

Omnibus:	16.351	Durbin-Watson:	2.019
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16.700
Skew:	-0.304	Prob(JB):	0.000236
Kurtosis:	3.177	Cond. No.	1.08e+06

Then I update the salary_p4:

$$updated\_sal = salary\_p4 - \exp(\hat{\delta}_0 + \hat{\delta}_1 grad\_year)$$

Using this method, I eliminate the influence of graduation year on salary_p4.

Here are the first five rows of updated_sal:

updated_sal
296.403632
496.512589
-8399.190964
-2396.781208
-15366.747388

(d) Because I have eliminated the influence of grad_year on both salary_p4 and gre_qnt, using updated_sal and updated_gre variable to test the hypothesis will reflect the net correlation between salary and GRE score (without the influence of grad_year). I conduct the OLS regression in python and obtain the following outcomes:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            updated_sal   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.3956
Date:                Tue, 16 Oct 2018   Prob (F-statistic):              0.529
Time:                        20:58:58   Log-Likelihood:                -10495.
No. Observations:                1000   AIC:                         2.099e+04
Df Residuals:                     998   BIC:                         2.100e+04
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const       -5819.1738   2094.231     -2.779      0.006   -9928.776   -1709.572
updated_gre    -8.4296     13.402     -0.629      0.529     -34.728      17.869
==============================================================================
Omnibus:                        1.261   Durbin-Watson:                   2.030
Prob(Omnibus):                  0.532   Jarque-Bera (JB):                1.164
Skew:                          -0.006   Prob(JB):                        0.559
Kurtosis:                       3.167   Cond. No.                     1.18e+03
==============================================================================
```

The estimated β_0 is -5819.7318, the standard errors of β_0 is 2094.231.

The estimated β_1 is -8.4296, the standard errors of β_1 is 13.402.

Compare the result of (d) and (a), firstly, I find that the estimated constant β_0 is smaller in (d) than (a). This is just because the updated_sal is smaller than the origin variable salary_p4 because in part(c), we remove the influence of grad_year on salary_p4 by $updated\_sal = salary\_p4 - \exp(\hat{\delta}_0 + \hat{\delta}_1 grad\_year)$, where $\exp(\hat{\delta}_0 + \hat{\delta}_1 grad\_year)$ is the fitted value of salary_p4.

The important change is β_1, which is larger in (d) than it is in part (a). The p-value in part (a) is 0.000, but in part (d), the p=value is 0.529. This means we cannot reject the hypothesis that $\beta_1 = 0$ in part (d), but I need to reject this hypothesis in part(a). This big difference is because in part (a), we do not remove the influence of grad_year on gre_qnt and salary_p4. The salary_p4 variable increases as the grad_year increases and gre_qnt is smaller when the grad_year is larger than 2011 than when the grad_year is smaller than 2011. Because the positive correlation between salary_p4 and grad_year as well as the negative correlation (although not linear) between gre_qnt and grad_year, the salary_p4 and gre_qnt will have a negative correlation due to the third variable grad_year, which means in part(a), I underestimate the coefficient β_1. In part(d), the result shows there is no significant negative correlation between salary and GRE after removing the influence of graduation year. In conclusion, there is no evidence that "higher intelligence is associated with higher income".

3. Assessment of Kossinets and Watts (2009)

The research question in the paper is “To what extent can observed homophily be attributed to individual preference (choice homophily) and structural constraints (induced homophily) respectively?”

To answer this question, the authors utilize a network data which records interactions between students, faculty, and staff as well as individual features and structural organizations. There are three different data sources: The first is the logs of e-mail interactions between individuals in a U.S. university. The second is the data of individual attributes, including status, gender, age, etc. The third data source is the record of course registration. There are 30396 individuals in the data, including undergraduate students (21%), graduate and professional students (27%), faculty (13%), and staff (13.4%) in the university (There were 43,553 individuals who used university e-mail to both send and receive messages during the academic year. However, the authors only include 30,396 individuals among them who exchanged messages with others that are active in both fall and spring semester). The authors only include email that were sent to singer recipient other than the sender, which has 7,156,162 messages, accounting for 82% of all email. The time period is one fall semester and one spring semester, 270 days in total (Though the full data set spans two calendar years, the authors only analyze one calendar years). The description and definition of all variable are in appendix A of the paper.

In the footnote 23 on the page 423 of the paper, the authors indicate the method of treating missing variables when calculating the aggregate measure of pairwise similarity, which is using the population mean for this similarity-scale component. The authors said that the missing values are nonrandom, like nonstudents having more missing data than students. The problem is that the group that has more missing data might also has higher/lower similarity (i.e. There is a correlation between the proportion of missing data and similarity). For example, if faculty have more missing values of age than others, suppose that either faculty i or j has a missing value age, and according to the authors’ suggestions, we would assign age match (i, j) = 0.175 for faculty i and j because 17.5% of all pairs are of the same age in the university. The problem is that the differences in age between faculty might be larger than that between students, indicating that 0.175 overestimates the similarity between faculty i and j. Similarity is a very important variable in the paper. If at the same time, faculty are more or less likely to form new ties with others, there would be a biased estimation of the correlation between similarity and relationship forming.

There are some weakness of the match of data source and theoretical construct. For instance, some email contact might not represent interpersonal relationship, like the email sent from an administrative staff to all students in a department. The authors address this weakness by including only messages that were sent to a single recipient other than the sender (eliminating multi-recipient e-mail). The authors also eliminate the simultaneous messages from the same sender that differed in size by less than 100 bytes. Therefore, after eliminating the email logs which do not represent interpersonal communication, the match between theoretical construct and data improves.