hw2

October 16, 2018

1 Assignment 2

1.0.1 MACS 30000, Dr. Evans

1.0.2 Dongcheng Yang

1.0.3 Oct. 11

```
In [1]: # Import packages
        import numpy as np
        import pandas as pd
        import statsmodels.api as sm
        import matplotlib.pyplot as plt
        # plt.style.use('seaborn')
```

```
C:\ProgramData\Anaconda3\lib\site-packages\statsmodels\compat\pandas.py:56: FutureWarning: The
  from pandas.core import datetools
```

1.0.4 1. imputing age and gender

(a) The procedures of imputing age and gender are as follows:

1) Regress age and female seperately on the variables total income and weight by using SurvIncome dataset
   The regression equations are:

$$age_i = \beta_0 + \beta_1 * totalincome_i + \beta_2 * weight_i + \epsilon_i$$

$$female_i = \beta_0 + \beta_1 * totalincome_i + \beta_2 * weight_i + \epsilon_i$$

2) Compute total income in BestIncome dataset by adding up variables "lab_inc" and "cap_inc"

3) Use the created variable total income and weight from BestIncome dataset and the two linear regressions above to impute age and gender

4) For the computed gender variables with value more than 0.5, let it equal to 1 as a representative of female

```
In [2]: # Define Data
        SI = pd.read_csv('SurvIncome.txt', index_col=0,header=None).reset_index()
        SI.columns=["tot_inc","wgti","age","female"]
        print(SI.head())
        BI = pd.read_csv('BestIncome.txt', index_col=0,header=None).reset_index()
        BI.columns=["lab_inc","cap_inc","hgt","wgt"]
        print(BI.head())
```

```
        tot_inc        wgti        age  female
0  63642.513655  134.998269  46.610021     1.0
1  49177.380692  134.392957  48.791349     1.0
2  67833.339128  126.482992  48.429894     1.0
3  62962.266217  128.038121  41.543926     1.0
4  58716.952597  126.211980  41.201245     1.0
        lab_inc       cap_inc        hgt         wgt
0  52655.605507   9279.509829  64.568138  152.920634
1  70586.979225   9451.016902  65.727648  159.534414
2  53738.008339   8078.132315  66.268796  152.502405
3  55128.180903  12692.670403  62.910559  149.218189
4  44482.794867   9812.975746  68.678295  152.726358
```

(b) Here is where I'll use my proposed method from part (a) to impute variables.

```
In [3]: # regression of age on tot_inc and wgti
        outcome  ='age'
        features = ['tot_inc', 'wgti']
        X, y = SI[features], SI[outcome]
        X_vars = sm.add_constant(X, prepend=False)
        m = sm.OLS(y, X_vars)
        res = m.fit()
        print(res.summary())

        #Getting Age With a Custom Formula
        def get_age(row):
            tot_inc = row[0]
            wgt=row[1]
            age = 44.2097+(tot_inc* 2.52e-05)+(wgt*(-0.0067))
            return age

        # Impute Variable Age
        BI['tot_inc']=BI['lab_inc']+BI['cap_inc']
        BI['imputed_age'] = BI[['tot_inc', 'wgt']].apply(get_age, axis=1)
        BI.head()
```

```python
# regression of gender on tot_inc and wgti
outcome  ='female'
features = ['tot_inc', 'wgti']
X, y = SI[features], SI[outcome]
X_vars = sm.add_constant(X, prepend=False)
m = sm.OLS(y, X_vars)
res = m.fit()
print(res.summary())

#Getting Gender With a Custom Formula
def get_gender(row):
    tot_inc = row[0]
    wgt=row[1]
    female = 3.7611+(tot_inc* (-5.25e-06))+(wgt*(-0.0195))
    return female

# Impute Variable Female
BI['imputed_female'] = BI[['tot_inc', 'wgt']].apply(get_gender, axis=1)

# Change imputed_female to int
BI['imputed_female'] = np.where(BI['imputed_female']>=0.5, 1, 0)
BI.head()
```

```
                            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:                        18:32:08   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]
------------------------------------------------------------------------------
tot_inc      2.52e-05   2.26e-05      1.114      0.266   -1.92e-05    6.96e-05
wgti          -0.0067      0.010     -0.686      0.493      -0.026       0.013
const         44.2097      1.490     29.666      0.000      41.285      47.134
==============================================================================
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
==============================================================================

Warnings:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.2e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
                            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:                        18:32:08   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]
------------------------------------------------------------------------------
tot_inc     -5.25e-06   7.76e-07     -6.765      0.000   -6.77e-06   -3.73e-06
wgti          -0.0195      0.000    -58.098      0.000      -0.020      -0.019
const          3.7611      0.051     73.600      0.000       3.661       3.861
==============================================================================
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
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.2e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
Out[3]:         lab_inc       cap_inc        hgt         wgt       tot_inc  \
        0  52655.605507   9279.509829  64.568138  152.920634  61935.115336
        1  70586.979225   9451.016902  65.727648  159.534414  80037.996127
        2  53738.008339   8078.132315  66.268796  152.502405  61816.140654
        3  55128.180903  12692.670403  62.910559  149.218189  67820.851305
        4  44482.794867   9812.975746  68.678295  152.726358  54295.770612

           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) Here is where I'll report the descriptive statistics for my new imputed variables.

```
In [4]: #Get imputed_age Descriptive Stats
        print(BI['imputed_age'].describe())
        #Get imputed_female Descriptive Stats
        print(BI['imputed_female'].describe())
```

```
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) Correlation matrix for the now six variables

```
In [5]: # Correction Matrix
        del BI['tot_inc']
        corr = BI.corr()
        corr.style.background_gradient()
```

```
Out[5]: <pandas.io.formats.style.Styler at 0x21ec43628d0>
```

1.0.5 2. Stationarity and data drift

(a) Estimate by OLS and report coefficients

```
In [6]: # Define Data
        II = pd.read_csv('IncomeIntel.txt', index_col=0,header=None).reset_index()
        II.columns=["grad_year","gre_qnt","salary_p4"]
        print(II.head())

        # regression of salary_p4 on gre_qnt
        outcome  ='salary_p4'
        features = 'gre_qnt'
        X, y = II[features], II[outcome]
        X_vars = sm.add_constant(X, prepend=False)
        m = sm.OLS(y, X_vars)
```

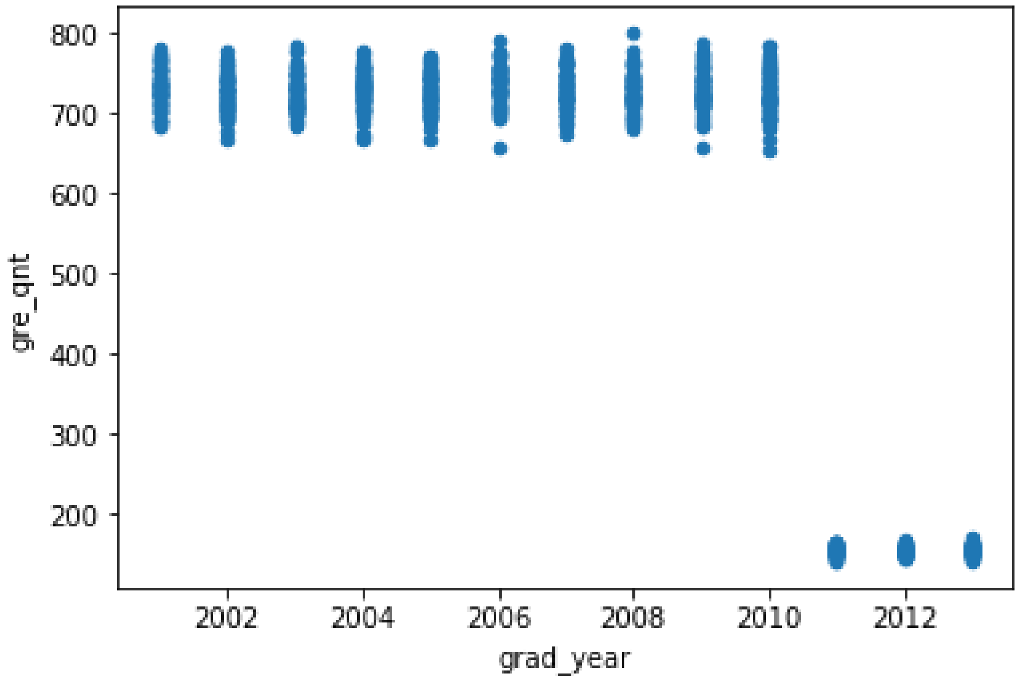
```
res = m.fit()
print(res.summary())
```

```
   grad_year     gre_qnt     salary_p4
0     2001.0  739.737072  67400.475185
1     2001.0  721.811673  67600.584142
2     2001.0  736.277908  58704.880589
3     2001.0  770.498485  64707.290345
4     2001.0  735.002861  51737.324165
                            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:                        18:32:25   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]
------------------------------------------------------------------------------
gre_qnt      -25.7632      1.365    -18.875      0.000     -28.442     -23.085
const       8.954e+04    878.764    101.895      0.000    8.78e+04    9.13e+04
==============================================================================
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
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(b) Create a scatterplot of GRE score and graduation year.

```
In [7]: # Make scatterplot of GRE score and graduation year
        # simple scatterplot using matplotlib
        II.plot(x='grad_year', y='gre_qnt', kind='scatter')
        plt.show()
```



From the scatterplot, we can see that at year 2011, the GRE underwent a substantial revision to its scoring system, which can be viewed as a drift. The solution to this problem is to create a one-to-one mapping that could transfer the score under current system to the previous. For score x,y seperately in current and previous system, I assume that the correlation can be roughly calculated as:

$$\frac{x-130}{170-130} = \frac{y-200}{800-200}$$

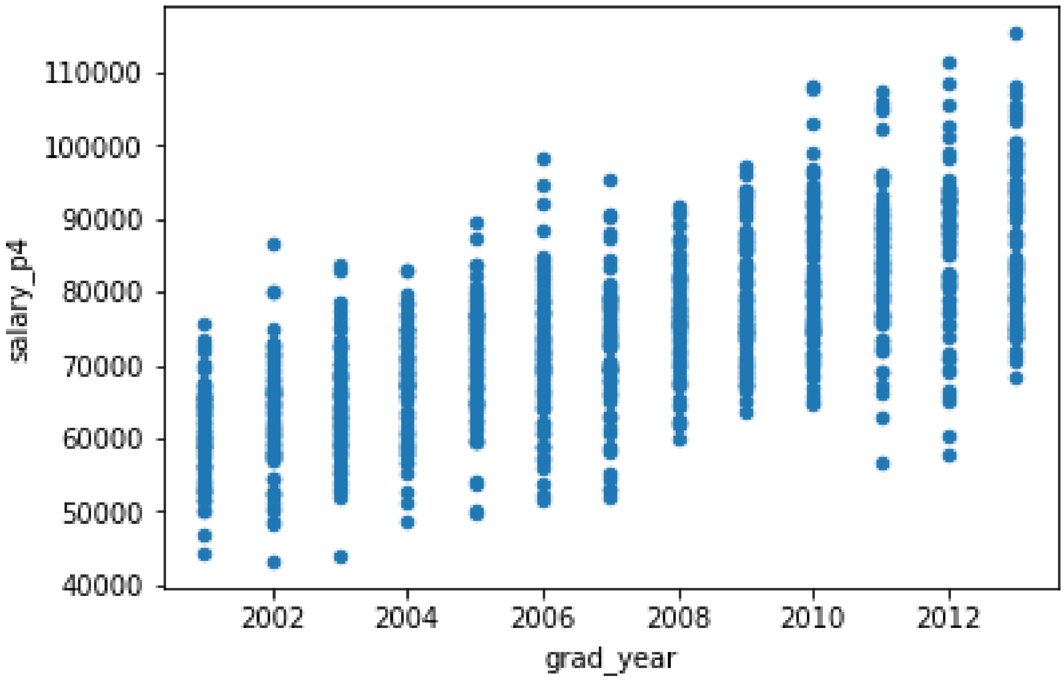
In [8]:
```
# Change the score after 2011 according to the mapping above
II['new_gre_qnt'] = np.where(II['grad_year']>=2011, (II['gre_qnt']-130)*15+200, II['gre
II.head()
```

Out[8]:

	grad_year	gre_qnt	salary_p4	new_gre_qnt
0	2001.0	739.737072	67400.475185	739.737072
1	2001.0	721.811673	67600.584142	721.811673
2	2001.0	736.277908	58704.880589	736.277908
3	2001.0	770.498485	64707.290345	770.498485
4	2001.0	735.002861	51737.324165	735.002861

(c) Create a scatterplot of income and graduation year

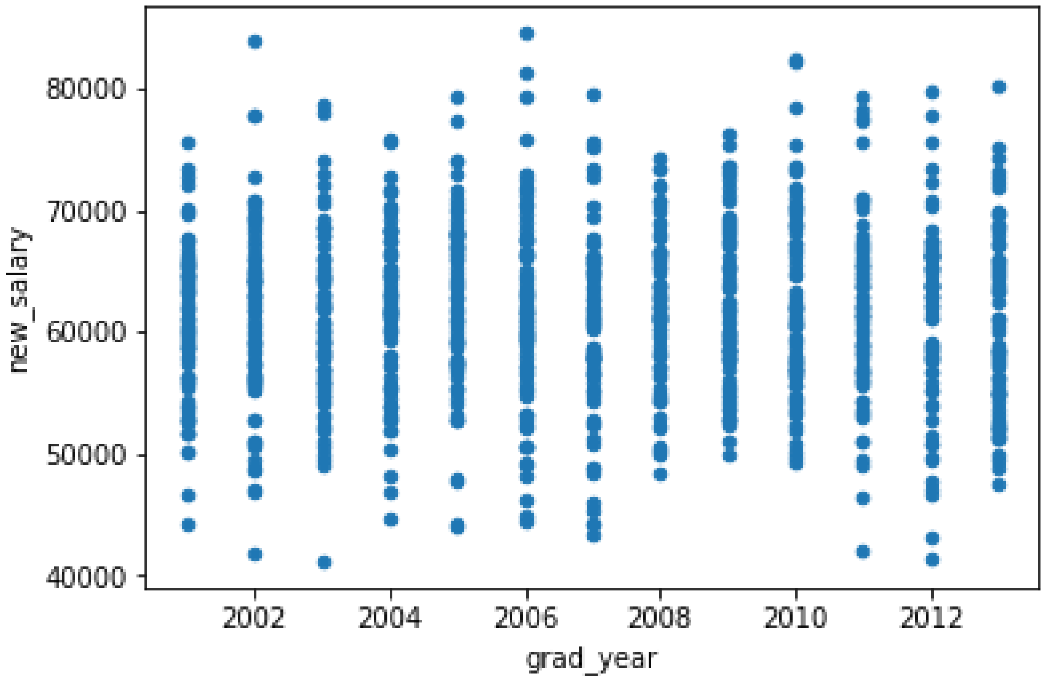
In [9]:
```
# Make scatterplot of income and graduation year
# simple scatterplot using matplotlib
II.plot(x='grad_year', y='salary_p4', kind='scatter')
plt.show()
```



The obvious problem here is that there is an increasing trend of the variable salary_p4. The procedures to deal with the problem are as follows:
1) Treat the first year of the data grad_year=2001 equal to the base year.
2) Calculate the average growth rate in salary by calculating the mean salary each year and calculating the average growth rate in salaries across all 13 years.
3) Divide each salary according to the corresponding year and get a new salary variable.

```
In [10]: avg_inc_by_year = II['salary_p4'].groupby(II['grad_year']).mean().values
         avg_growth_rate = ((avg_inc_by_year[1:] - avg_inc_by_year[:-1]) / avg_inc_by_year[:-1
         II['new_salary'] = II['salary_p4']/((1 + avg_growth_rate) ** (II['grad_year'] - 2001)
         print(II.head())
         II.plot(x='grad_year', y='new_salary', kind='scatter')
         plt.show()
```

```
   grad_year     gre_qnt     salary_p4  new_gre_qnt    new_salary
0     2001.0  739.737072  67400.475185   739.737072  67400.475185
1     2001.0  721.811673  67600.584142   721.811673  67600.584142
2     2001.0  736.277908  58704.880589   736.277908  58704.880589
3     2001.0  770.498485  64707.290345   770.498485  64707.290345
4     2001.0  735.002861  51737.324165   735.002861  51737.324165
```



(d) Re-estimate coefficients with updated variables.

```
In [11]: # Code to re-estimate, output of new coefficients
         outcome  ='new_salary'
         features = 'new_gre_qnt'
         X, y = II[features], II[outcome]
         X_vars = sm.add_constant(X, prepend=False)
         m = sm.OLS(y, X_vars)
         res = m.fit()
         print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             new_salary   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                   0.05257
Date:                Tue, 16 Oct 2018   Prob (F-statistic):              0.819
Time:                        18:32:38   Log-Likelihood:                -10291.
No. Observations:                1000   AIC:                         2.059e+04
Df Residuals:                     998   BIC:                         2.060e+04
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
```

```
-------------------------------------------------------------------------------
new_gre_qnt    -0.6645      2.898     -0.229      0.819     -6.352       5.023
const        6.188e+04   2020.482     30.626      0.000    5.79e+04    6.58e+04
==============================================================================
Omnibus:                        0.757   Durbin-Watson:                   2.026
Prob(Omnibus):                  0.685   Jarque-Bera (JB):                0.668
Skew:                           0.059   Prob(JB):                        0.716
Kurtosis:                       3.048   Cond. No.                     6.24e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.24e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The estimated coefficient of the interest variable "gre_qnt" is now insignificant. This is because I put gre score in the same system, and delete the growth trend of salary data, which changed both variables' sample variance. With the unchanged regression results, we may conclude that there is a significant negative correlation between income and GRE quant scores, which violates the author's hypothesis. In this regression with transformed "gre_qnt" variable, we can see that this effect is insignificant, though the coefficient is still negative.

1.0.6 3. Assessment of Kossinets and Watts.

See attached PDF.