Assignment 3: Applied data science

**ANSWERS:**

1. Solution are described in the following sections:
2. Reading dta data process is attached in the source code.
3. Summary of statistics :

**rns =**

count 758.000000

mean 0.269129

std 0.443800

min 0.000000

25% 0.000000

50% 0.000000

75% 1.000000

max 1.000000

dtype: float64

**mrt =**

count 758.000000

mean 0.514512

std 0.500119

min 0.000000

25% 0.000000

50

Dimas Rinarso Putro | [drp354@nyu.edu](mailto:drp354@nyu.edu) | Applied Data Science GX5004 | HW 3

1

% 1.000000

75% 1.000000

max 1.000000

dtype: float64

smsa =

count 758.000000

mean 0.704485

std 0.456575

min 0.000000

25% 0.000000

50% 1.000000

75% 1.000000

max 1.000000

dtype: float64

med =

count 758.00000

mean 10.91029

std 2.74112

min 0.00000

25% 9.00000

50% 12.00000

75% 12.00000

max 18.00000

dtype: float64

ziq =

count 758.000000

mean 103.856201

std 13.618666

min 54.000000

25% 95.250000

50% 104.000000

75% 113.750000

max 145.000000

dtype: float64

kww =

count 758.000000

mean 36.573879

std 7.302247

min 12.000000

25% 32.000000

50% 37.000000

75% 41.000000

max 56.000000

dtype: float64

age =

count 758.000000

mean 21.835092

std 2.981756

min 16.000000

25% 20.000000

50% 22.000000

75% 24.000000

max 30.000000

dtype: float64

s =

count 758.000000

mean 13.405013

std 2.231828

min 9.000000

25% 12.000000

50% 12.000000

75% 16.000000

max 18.000000

dtype: float64

expr =

count 758.000000

mean 1.735429

std 2.105542

min 0.000000

25% 0.281500

50% 0.960000

75% 2.440000

max 11.444000

dtype: float64

lw =

count 758.000000

mean 5.686739

std 0.428949

min 4.605000

25% 5.380000

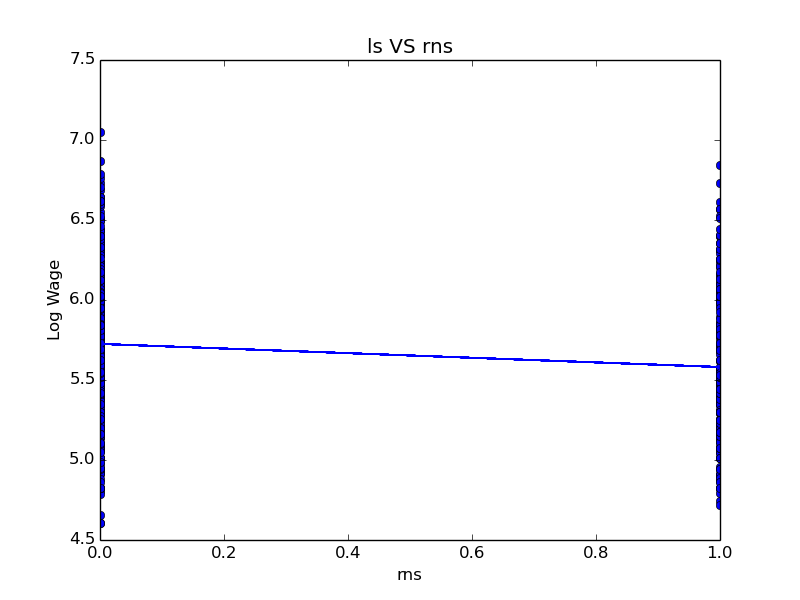
50% 5.684000

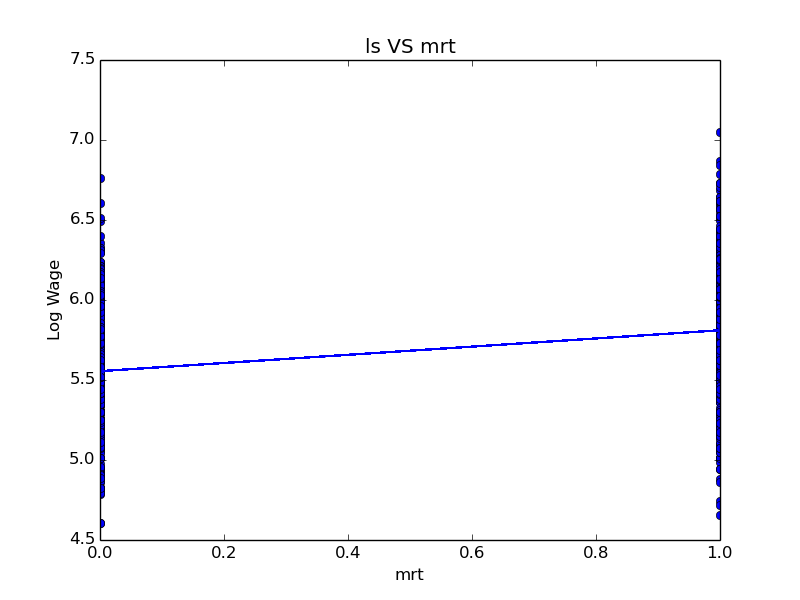
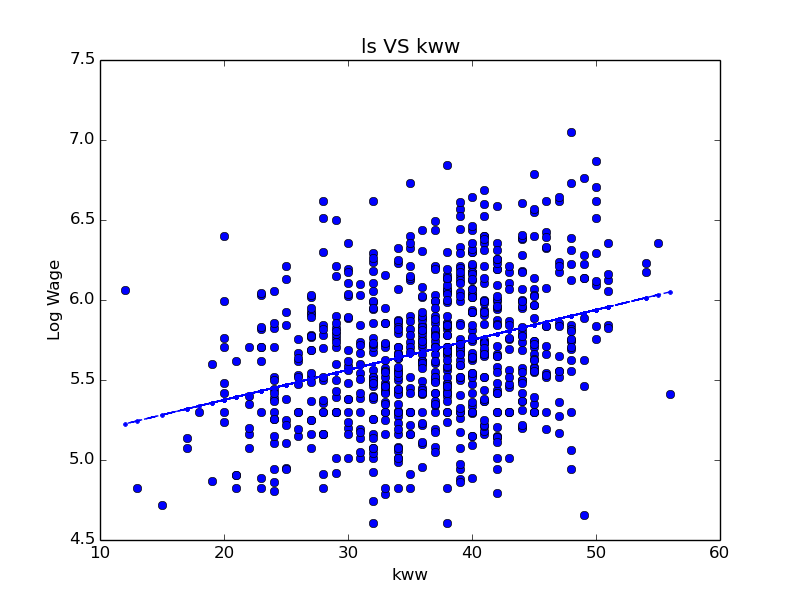
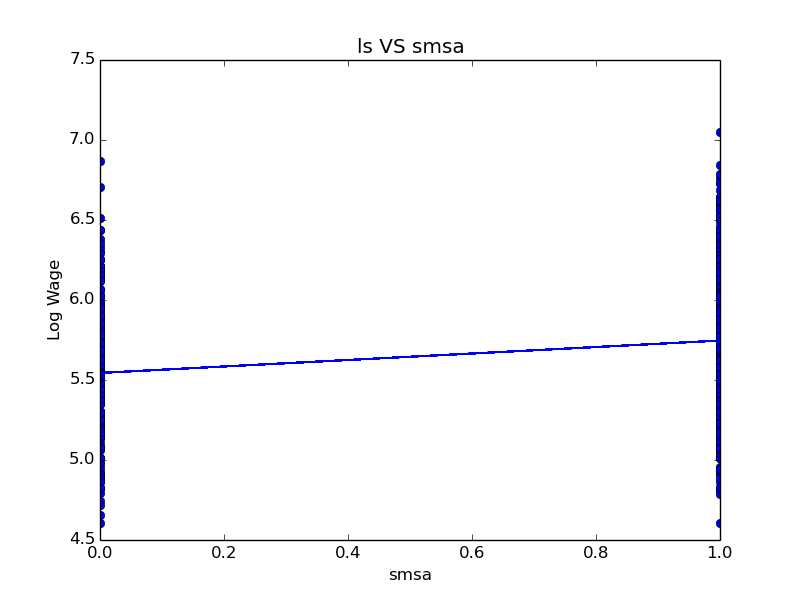
75% 5.991000

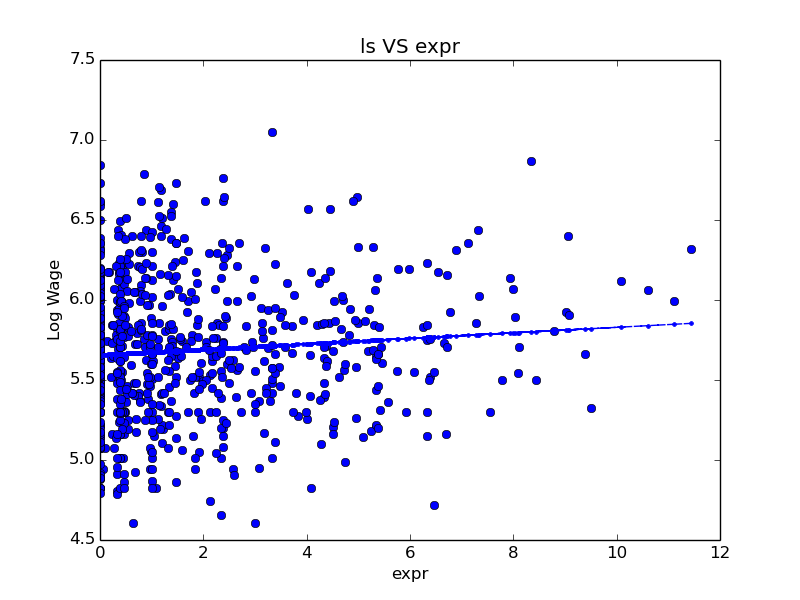
max 7.051000

dtype: float64

1. Scatter plot of Log Wages against (RNS ,MRT, SMSA,KWW,EXPR):





1. The bivariate square models for model point c:

**RNS**

OLS Regression Results

==============================================================================

Dep. Variable: lw R-squared: 0.022

Model: OLS Adj. R-squared: 0.021

Method: Least Squares F-statistic: 17.30

Date: Thu, 25 Sep 2014 Prob (F-statistic): 3.56e-05

Time: 14:47:11 Log-Likelihood: -424.90

No. Observations: 758 AIC: 853.8

Df Residuals: 756 BIC: 863.1

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 5.7256 0.018 317.540 0.000 5.690 5.761

rns -0.1446 0.035 -4.159 0.000 -0.213 -0.076

==============================================================================

Omnibus: 7.316 Durbin-Watson: 1.734

Prob(Omnibus): 0.026 Jarque-Bera (JB): 7.370

Skew: 0.223 Prob(JB): 0.0251

Kurtosis: 2.817 Cond. No. 2.45

==============================================================================

**MRT**

OLS Regression Results

==============================================================================

Dep. Variable: lw R-squared: 0.089

Model: OLS Adj. R-squared: 0.088

Method: Least Squares F-statistic: 73.70

Date: Thu, 25 Sep 2014 Prob (F-statistic): 5.14e-17

Time: 14:47:12 Log-Likelihood: -398.21

No. Observations: 758 AIC: 800.4

Df Residuals: 756 BIC: 809.7

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 5.5552 0.021 260.095 0.000 5.513 5.597

mrt 0.2556 0.030 8.585 0.000 0.197 0.314

==============================================================================

Omnibus: 3.526 Durbin-Watson: 1.667

Prob(Omnibus): 0.172 Jarque-Bera (JB): 3.192

Skew: 0.092 Prob(JB): 0.203

Kurtosis: 2.741 Cond. No. 2.65

==============================================================================

**SMSA**

OLS Regression Results

==============================================================================

Dep. Variable: lw R-squared: 0.046

Model: OLS Adj. R-squared: 0.045

Method: Least Squares F-statistic: 36.85

Date: Thu, 25 Sep 2014 Prob (F-statistic): 2.02e-09

Time: 14:47:13 Log-Likelihood: -415.43

No. Observations: 758 AIC: 834.9

Df Residuals: 756 BIC: 844.1

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 5.5441 0.028 197.967 0.000 5.489 5.599

smsa 0.2025 0.033 6.070 0.000 0.137 0.268

==============================================================================

Omnibus: 5.469 Durbin-Watson: 1.766

Prob(Omnibus): 0.065 Jarque-Bera (JB): 5.357

Skew: 0.174 Prob(JB): 0.0687

Kurtosis: 2.781 Cond. No. 3.45

==============================================================================

**KWW**

OLS Regression Results

==============================================================================

Dep. Variable: lw R-squared: 0.102

Model: OLS Adj. R-squared: 0.101

Method: Least Squares F-statistic: 85.91

Date: Thu, 25 Sep 2014 Prob (F-statistic): 1.93e-19

Time: 14:47:14 Log-Likelihood: -392.68

No. Observations: 758 AIC: 789.4

Df Residuals: 756 BIC: 798.6

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 5.0005 0.076 66.228 0.000 4.852 5.149

kww 0.0188 0.002 9.269 0.000 0.015 0.023

==============================================================================

Omnibus: 4.135 Durbin-Watson: 1.731

Prob(Omnibus): 0.126 Jarque-Bera (JB): 3.864

Skew: 0.123 Prob(JB): 0.145

Kurtosis: 2.752 Cond. No. 191.

==============================================================================

**EXPR**

OLS Regression Results

==============================================================================

Dep. Variable: lw R-squared: 0.007

Model: OLS Adj. R-squared: 0.006

Method: Least Squares F-statistic: 5.452

Date: Fri, 26 Sep 2014 Prob (F-statistic): 0.0198

Time: 18:10:22 Log-Likelihood: -430.75

No. Observations: 758 AIC: 865.5

Df Residuals: 756 BIC: 874.8

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 5.6568 0.020 280.924 0.000 5.617 5.696

expr 0.0172 0.007 2.335 0.020 0.003 0.032

==============================================================================

Omnibus: 7.070 Durbin-Watson: 1.673

Prob(Omnibus): 0.029 Jarque-Bera (JB): 6.678

Skew: 0.187 Prob(JB): 0.0355

Kurtosis: 2.733 Cond. No. 3.74

==============================================================================

**Comments:**

* RNS

Negative coefficient shows that people from southern states tend to have lower wage than non-southern resident.

* MRT

Positive coefficient shows that married people tend to have higher wage, possibly because the demand of financial stability.

* SMSA

Positive coefficient shows that people who live in urban areas have a tendency to have higher wage.

* KWW

The data shows strong positive correlation between the result of "Knowledge of the World of Work" test score with high wage.

* EXPR

The graph shows strong correlation between experience in years to higher wage.

1. Bivariate least squares model relating log wages to schooling and its 95 confidence interval:

OLS Regression Results

==============================================================================

Dep. Variable: lw R-squared: 0.253

Model: OLS Adj. R-squared: 0.252

Method: Least Squares F-statistic: 255.7

Date: Thu, 25 Sep 2014 Prob (F-statistic): 8.52e-50

Time: 14:47:14 Log-Likelihood: -323.05

No. Observations: 758 AIC: 650.1

Df Residuals: 756 BIC: 659.4

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 4.3915 0.082 53.481 0.000 4.230 4.553

s 0.0966 0.006 15.991 0.000 **0.085 0.108**

==============================================================================

Omnibus: 1.749 Durbin-Watson: 1.733

Prob(Omnibus): 0.417 Jarque-Bera (JB): 1.697

Skew: 0.021 Prob(JB): 0.428

Kurtosis: 3.228 Cond. No. 83.2

==============================================================================

**Note:** using python statsmodels we can automatically calculate the range of 95% confidence level, which is **0.085** (lower boundaries) and **0.108** (upper boundaries).

1. Multivariate least squares model relating log wages to the variables in b:

OLS Regression Results

==============================================================================

Dep. Variable: lw R-squared: 0.433

Model: OLS Adj. R-squared: 0.426

Method: Least Squares F-statistic: 63.38

Date: Thu, 25 Sep 2014 Prob (F-statistic): 4.21e-86a

Time: 15:47:24 Log-Likelihood: -218.67

No. Observations: 758 AIC: 457.3

Df Residuals: 748 BIC: 503.6

Df Model: 9

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 3.4149 0.123 27.838 0.000 3.174 3.656

rns -0.0877 0.027 -3.203 0.001 -0.142 -0.034

mrt 0.1007 0.027 3.716 0.000 0.047 0.154

smsa 0.1368 0.027 5.144 0.000 0.085 0.189

med 0.0059 0.005 1.258 0.209 -0.003 0.015

iq 0.0042 0.001 3.998 0.000 0.002 0.006

kww -0.0023 0.002 -1.174 0.241 -0.006 0.002

age 0.0497 0.006 8.342 0.000 0.038 0.061

s 0.0479 0.008 6.159 0.000 **0.033 0.063**

expr 0.0022 0.007 0.310 0.757 -0.012 0.016

==============================================================================

Omnibus: 12.122 Durbin-Watson: 1.798

Prob(Omnibus): 0.002 Jarque-Bera (JB): 18.748

Skew: -0.104 Prob(JB): 8.49e-05

Kurtosis: 3.742 Cond. No. 1.19e+03

==============================================================================

Warnings:

[1] The condition number is large, 1.19e+03. This might indicate that there are

strong multicollinearity or other numerical problems.

**Note:** using python statsmodels we can automatically calculate the range of 95% confidence level, which is **0.033** (lower boundaries) and **0.063** (upper boundaries).

1. Generate a variable that is age raised to the power of two (i.e., is age squared), then re-estimate f. including age-squared:

Results

==============================================================================

Dep. Variable: lw R-squared: 0.438

Model: OLS Adj. R-squared: 0.430

Method: Least Squares F-statistic: 58.21

Date: Thu, 25 Sep 2014 Prob (F-statistic): 1.06e-86

Time: 16:14:55 Log-Likelihood: -215.09

No. Observations: 758 AIC: 452.2

Df Residuals: 747 BIC: 503.1

Df Model: 10

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 4.8628 0.557 8.728 0.000 3.769 5.957

rns -0.0847 0.027 -3.103 0.002 -0.138 -0.031

mrt 0.1118 0.027 4.095 0.000 0.058 0.165

smsa 0.1400 0.027 5.281 0.000 0.088 0.192

med 0.0056 0.005 1.207 0.228 -0.004 0.015

iq 0.0041 0.001 3.880 0.000 0.002 0.006

kww -0.0020 0.002 -1.037 0.300 -0.006 0.002

age -0.0838 0.050 -1.660 0.097 -0.183 0.015

s 0.0511 0.008 6.519 0.000 0.036 0.066

expr 0.0037 0.007 0.515 0.606 -0.010 0.018

a 0.0029 0.001 2.664 0.008 0.001 0.005

==============================================================================

Omnibus: 14.225 Durbin-Watson: 1.795

Prob(Omnibus): 0.001 Jarque-Bera (JB): 22.965

Skew: -0.121 Prob(JB): 1.03e-05

Kurtosis: 3.818 Cond. No. 2.46e+04

==============================================================================

Warnings:

[1] The condition number is large, 2.46e+04. This might indicate that there are

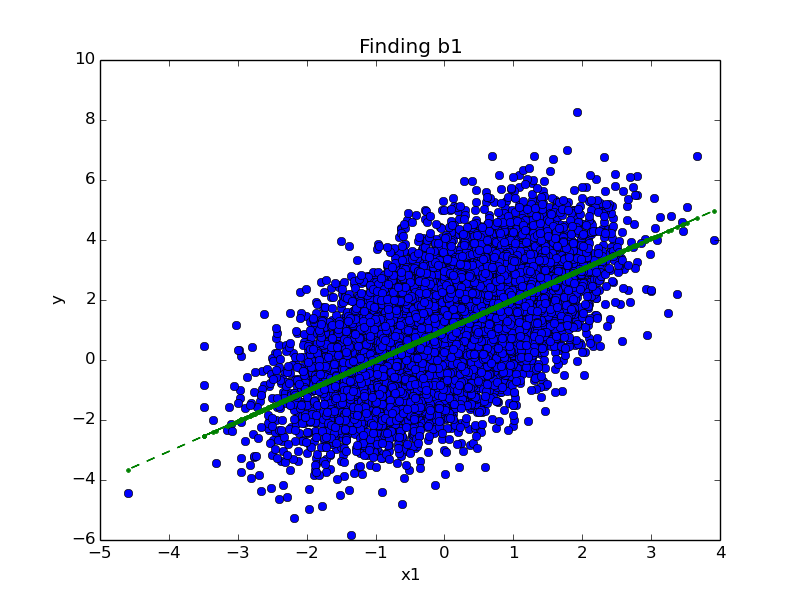
strong multicollinearity or other numerical problems.

**Note:** by including age square into the equation, the coefficient of original age variable is getting smaller while some other variables’ coefficients also changed. This shows that in multivariate regression adding new variables affect the whole coefficient, mostly one with high correlation with the one to be included (in this one age and age square).

1. The difference of estimates of the returns to schooling in e. and f. results shows that multivariable might have correlation to each other. Moreover, it can also be assumed that the covariance of the schooling and the unknown error might not be zero.

**ANSWERS:**

1. Solution are described in the following sections:
   1. simulation of DGP assuming 10,000 observations and estimate the least squares value for based on equations in problem 2a:



OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.342

Model: OLS Adj. R-squared: 0.342

Method: Least Squares F-statistic: 5202.

Date: Thu, 25 Sep 2014 Prob (F-statistic): 0.00

Time: 17:09:09 Log-Likelihood: -17503.

No. Observations: 10000 AIC: 3.501e+04

Df Residuals: 9998 BIC: 3.502e+04

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 1.0101 0.014 72.509 0.000 0.983 1.037

**x1 1.0143 0.014** 72.122 0.000 0.987 1.042

==============================================================================

Omnibus: 1.133 Durbin-Watson: 2.017

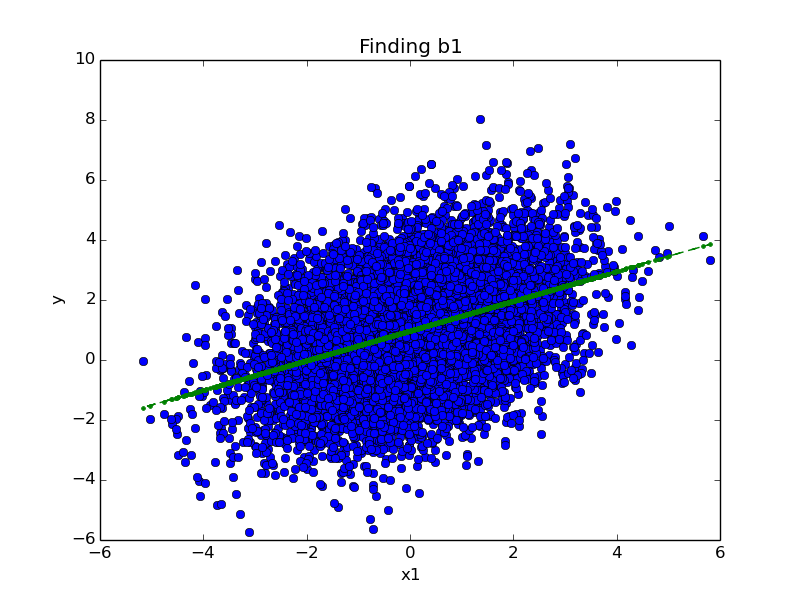
Prob(Omnibus): 0.567 Jarque-Bera (JB): 1.109

Skew: -0.002 Prob(JB): 0.574

Kurtosis: 3.051 Cond. No. 1.02

==============================================================================

* 1. simulation of DGP assuming 10,000 observations and estimate the least squares value for based on equations in problem 2b:



OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.160

Model: OLS Adj. R-squared: 0.160

Method: Least Squares F-statistic: 1906.

Date: Thu, 25 Sep 2014 Prob (F-statistic): 0.00

Time: 17:15:39 Log-Likelihood: -18896.

No. Observations: 10000 AIC: 3.780e+04

Df Residuals: 9998 BIC: 3.781e+04

Df Model: 1

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 0.9779 0.016 61.073 0.000 0.947 1.009

**x1 0.4957 0.011** 43.663 0.000 0.473 0.518

==============================================================================

Omnibus: 0.328 Durbin-Watson: 1.976

Prob(Omnibus): 0.849 Jarque-Bera (JB): 0.359

Skew: -0.005 Prob(JB): 0.836

Kurtosis: 2.972 Cond. No. 1.41

==============================================================================

* 1. **Comments:**

To find all variables that are relevant in the real world is impossible, and what’s more it may also lead to add unnecessary process as follows:

* + - * If we limit to the data being used in this problem 2, (data are random variables and independent), the overuse of many more statistical model might result in these problems:
        + Model might show that predictors have correlation to each other (although it should be totally independent)
        + Covariance to new unknown errors might not be zero.
      * Every time you add more predictor, by default the value of R-squared increases. Too many predictors and higher order polynomials may lead to model new random noise in the data, as known as *overfitting*, where it could produce higher R values and make it harder to make predictions.

**ANSWERS:**

1. Solution are described in the following sections:
2. Reading dta data process is attached in the source code.
3. Treating the years 70-78 of the NLSW data as a training set and estimate the model presented in class both as a linear and a logit:

* Get the data year 70-78 and fit the linear and logit model

...

df\_sliced2 **=** df**[**df**[**'year'**]** **>=**70**]**

df\_sliced **=** df\_sliced2**[**df**[**'year'**]** **<=**78**]**

x**=** df\_sliced**[[**'year'**,**'age'**,**'grade'**,**'south'**,**'black'**,**'smsa'**]]**

y **=** df\_sliced**.**union

X **=** sm**.**add\_constant**(**x**)**

#Least square regression

model\_linear **=** sm**.**OLS**(**y**,** X**)**

results\_linear **=** model\_linear**.**fit**()**

**print(**results\_linear**.**summary**())**

#Logit square regression

model\_logit **=** sm**.**Logit**(**y**,** X**)**

results\_logit **=** model\_logit**.**fit**()**

**print(**results\_logit**.**summary**())**

OLS Regression Results

==============================================================================

Dep. Variable: union R-squared: 0.045

Model: OLS Adj. R-squared: 0.045

Method: Least Squares F-statistic: 204.8

Date: Thu, 25 Sep 2014 Prob (F-statistic): 2.98e-256

Time: 18:16:49 Log-Likelihood: -13562.

No. Observations: 26200 AIC: 2.714e+04

Df Residuals: 26193 BIC: 2.720e+04

Df Model: 6

==============================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const 0.1714 0.052 3.277 0.001 0.069 0.274

year -0.0029 0.001 -3.183 0.001 -0.005 -0.001

age 0.0044 0.001 5.240 0.000 0.003 0.006

grade 0.0121 0.001 11.283 0.000 0.010 0.014

south -0.1421 0.005 -26.236 0.000 -0.153 -0.131

black 0.1442 0.006 24.148 0.000 0.132 0.156

smsa 0.0159 0.006 2.781 0.005 0.005 0.027

==============================================================================

Omnibus: 4332.120 Durbin-Watson: 1.987

Prob(Omnibus): 0.000 Jarque-Bera (JB): 6880.188

Skew: 1.252 Prob(JB): 0.00

Kurtosis: 2.815 Cond. No. 1.90e+03

==============================================================================

Warnings:

[1] The condition number is large, 1.9e+03. This might indicate that there are

strong multicollinearity or other numerical problems.

Optimization terminated successfully.

Current function value: 0.506056

Iterations 6

Logit Regression Results

==============================================================================

Dep. Variable: union No. Observations: 26200

Model: Logit Df Residuals: 26193

Method: MLE Df Model: 6

Date: Thu, 25 Sep 2014 Pseudo R-squ.: 0.04368

Time: 18:16:49 Log-Likelihood: -13259.

converged: True LL-Null: -13864.

LLR p-value: 1.894e-258

==============================================================================

coef std err z P>|z| [95.0% Conf. Int.]

------------------------------------------------------------------------------

const -1.5927 0.318 -5.016 0.000 -2.215 -0.970

year -0.0180 0.006 -3.270 0.001 -0.029 -0.007

age 0.0269 0.005 5.293 0.000 0.017 0.037

grade 0.0744 0.007 11.265 0.000 0.061 0.087

south -0.9013 0.035 -25.440 0.000 -0.971 -0.832

black 0.8535 0.036 23.994 0.000 0.784 0.923

smsa 0.0890 0.036 2.445 0.014 0.018 0.160

==============================================================================

1. Treat the years 80-88 of the NLSW data as a set of attributes on individuals that you would like to classify as union/non-union, using a threshold of 0.25 and both the linear and the logit classifiers estimated in b.:

The output of the calculation:

* Using the coefficient from year 70-78 to predict year 80-88:

...

df\_sample2 **=** df**[**df**[**'year'**]** **>=**80**]**

df\_sample **=** df\_sample2**[**df**[**'year'**]** **<=**88**]**

y\_hat\_linear **=** **[]**

y\_hat\_logit **=** **[]**

var**=** df\_sample**[[**'year'**,**'age'**,**'grade'**,**'south'**,**'black'**,**'smsa'**]]**

X\_pred **=** sm**.**add\_constant**(**var**)**

#linear

y\_hat\_linear **=** results\_linear**.**predict**(**X\_pred**)**

#logit

y\_hat\_logit **=** results\_logit**.**predict**(**X\_pred**)**

...

Output:

LINEAR: Estimated number of people which is Union : 4308

LOGIT : Estimated number of people which is Union : 3749

1. For both models, summarize the accuracy of your support vector machine (with a threshold of 0.2) in a table by comparing your union prediction to what was actually observed. It might look something like the table below.

* Now comparing prediction based on data year 70-78 with the actual linear and logit model from actual observation (year 80-88):

...

df\_sample2 **=** df**[**df**[**'year'**]** **>=**80**]**

df\_sample **=** df\_sample2**[**df**[**'year'**]** **<=**88**]**

y\_act **=** df\_sample**.**union

var**=** df\_sample**[[**'year'**,**'age'**,**'grade'**,**'south'**,**'black'**,**'smsa'**]]**

X\_pred **=** sm**.**add\_constant**(**var**)**

#linear regression

model\_linear\_act **=** sm**.**OLS**(**y\_act**,** X\_pred**)**

results\_linear\_act **=** model\_linear\_act**.**fit**()**

#Logit square regression

model\_logit\_act **=** sm**.**Logit**(**y\_act**,** X\_pred**)**

results\_logit\_act **=** model\_logit\_act**.**fit**()**

#Actual value from the data

##############################

union\_count\_real **=** 0

**for** people **in** df\_sample**.**union**:**

**if** float**(**people**)** **==** 1**:** union\_count\_real **+=** 1

**print** 'REAL DATA : Actual number of people which is Union : %d' **%** union\_count\_real

...

Because there might be misunderstanding to this questions, the middle column stating the number of union student predicted using coefficient in 80-88 were added

|  |  |  |  |
| --- | --- | --- | --- |
| SVM | Number of Union Members (Predicted using 70-78 training data) | Number of Union Members (predicted using 80-88 actual data) | Number of Union Members (Actual) |
| Linear | 9217 | 9968 | 3323 |
| Logit | 8230 | 9421 |