Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Feb 25h 2019

Please note that assignments in this class may be submitted to www.turnitin.com, a web- based anti-plagiarism system, for an evaluation of their originality.

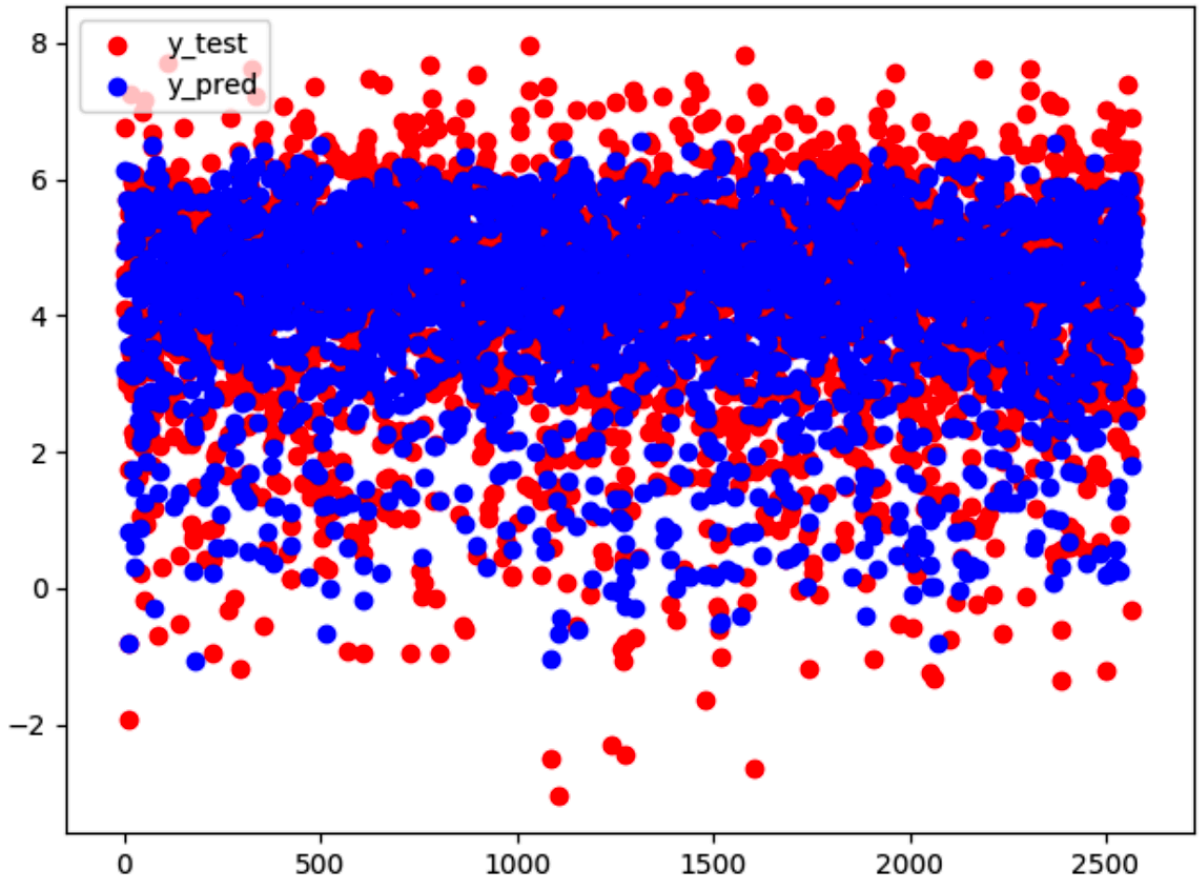
1. Randomly divide data of HW1 (Regression Data.xlsx) to training and validation sets.

```
hw2
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw2.py

The size of training set is:  23171

The size of validation set is:  2575

Process finished with exit code 0
```

In order to get a better model, I use 90% of the data as the training set, and the other 10% of the data as validation set.

2. Develop an MLR model by using variable selection method.

Following is the plot of the original regression model which using all the variables



In this model, the R-squared value is about 0.68.

```
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw2.py

The R-squared value of the model is  0.680384709562976

Process finished with exit code 0
```

Then we use variable selection method (Forward selection)
Firstly, compute the R-squared value of y with each Xi. Following are the results:

X1: 0.517
X2: 0.011
X3: 0.087
X4: 0.088
X5: 0.014
X6: 0.005
X7: 0.007
X8: 0.000
X9: 0.001
X10: 0.002

And in these variables, the X8's p-value is greater than our cutoff-value, so we will not select the X8 variable.

Then we try the multiple regression for the variables except X8 without interaction.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.673
Model:	OLS	Adj. R-squared:	0.673
Method:	Least Squares	F-statistic:	5895.
Date:	Mon, 25 Feb 2019	Prob (F-statistic):	0.00
Time:	17:00:37	Log-Likelihood:	-34665.
No. Observations:	25746	AIC:	6.935e+04
Df Residuals:	25736	BIC:	6.943e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.8535	0.117	24.399	0.000	2.624	3.083
x1	-0.4375	0.002	-201.777	0.000	-0.442	-0.433
x4	-0.0514	0.003	-18.564	0.000	-0.057	-0.046
x3	0.0377	0.003	13.279	0.000	0.032	0.043
x5	-0.1201	0.014	-8.362	0.000	-0.148	-0.092
x2	0.4283	0.017	24.958	0.000	0.395	0.462
x7	-0.0473	0.002	-21.169	0.000	-0.052	-0.043
x6	-0.6672	0.035	-19.087	0.000	-0.736	-0.599
x10	0.0969	0.008	12.779	0.000	0.082	0.112
x9	-0.0005	3.3e-05	-15.492	0.000	-0.001	-0.000

Omnibus:	218.725	Durbin-Watson:	0.762
Prob(Omnibus):	0.000	Jarque-Bera (JB):	239.120
Skew:	-0.192	Prob(JB):	1.19e-52
Kurtosis:	3.275	Cond. No.	8.94e+03

We can find in this way, the R-squared value is about 0.673.

And then try the multiple regression for the variables except X8 with interaction.

```
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw2.py
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.797
Model:                            OLS   Adj. R-squared:                  0.797
Method:                 Least Squares   F-statistic:                     4814.
Date:                Mon, 25 Feb 2019   Prob (F-statistic):               0.00
Time:                        17:00:37   Log-Likelihood:                -28532.
No. Observations:               25746   AIC:                         5.711e+04
Df Residuals:                   25724   BIC:                         5.729e+04
Df Model:                          21
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.1363      0.161     25.669      0.000       3.820       4.452
x1            -1.8677      0.032    -58.581      0.000      -1.930      -1.805
x4            -0.4725      0.028    -16.693      0.000      -0.528      -0.417
x3             0.4049      0.029     14.190      0.000       0.349       0.461
x1:x3          0.0004   4.37e-05      8.357      0.000       0.000       0.000
x3:x4          0.0001    2.4e-06     48.160      0.000       0.000       0.000
x1:x5         -0.0175      0.004     -4.193      0.000      -0.026      -0.009
x2             0.3049      0.024     12.693      0.000       0.258       0.352
x1:x2          0.2148      0.005     46.073      0.000       0.206       0.224
x3:x2         -0.0595      0.004    -14.370      0.000      -0.068      -0.051
x4:x2          0.0641      0.004     15.676      0.000       0.056       0.072
x5:x2          0.0190      0.005      4.093      0.000       0.010       0.028
x7            -0.3644      0.031    -11.765      0.000      -0.425      -0.304
x7:x1         -0.0105      0.001    -18.209      0.000      -0.012      -0.009
x4:x7          0.0006    7.1e-05      8.620      0.000       0.000       0.001
x5:x7          0.0836      0.004     19.150      0.000       0.075       0.092
x2:x7          0.0532      0.005     10.331      0.000       0.043       0.063
x6            -0.3323      0.029    -11.654      0.000      -0.388      -0.276
x10            0.1349      0.008     17.061      0.000       0.119       0.150
x5:x10        -0.1387      0.012    -11.548      0.000      -0.162      -0.115
x9            -0.0005   3.11e-05    -15.668      0.000      -0.001      -0.000
x1:x9         -0.0001   9.46e-06    -12.076      0.000      -0.000   -9.57e-05
==============================================================================
Omnibus:                      157.986   Durbin-Watson:                   1.042
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              180.374
Skew:                          -0.142   Prob(JB):                     6.80e-40
Kurtosis:                       3.295   Cond. No.                     3.37e+05
==============================================================================
```

We can find after we add some interactions, the R-squared value is raised to 0.797.

Then we try another variable selection method (Backward selection).
Firstly, we add all of the variables and some interaction in regression model:

```
mod013 = smf.ols(formula = 'y ~ x1 + x1:x2 + x1:x3 + x1:x4 + x1:x5 + x1:x6 + x1:x7 + x1:x8 + x1:x9 + x1:x10 + '
                             'x2 + x2:x3 + x2:x4 + x2:x5 + x2:x6 + x2:x7 + x2:x8 + x2:x9 + x2:x10 + '
                             'x3 + x3:x4 + x3:x5 + x3:x6 + x3:x7 + x3:x8 + x3:x9 + x3:x10 + '
                             'x4 + x4:x5 + x4:x6 + x4:x7 + x4:x8 + x4:x9 + x4:x10 + '
                             'x5 + x5:x6 + x5:x7 + x5:x8 + x5:x9 + x5:x10 + '
                             'x6 + x6:x7 + x6:x8 + x6:x9 + x6:x10 + '
                             'x7 + x7:x8 + x7:x9 + x7:x10 + '
                             'x8 + x8:x9 + x8:x10 + '
                             'x9 + x9:x10 + '
                             'x10', data = data).fit()
sum013= mod013.summary()
print(sum013)
```

```
hw2 ×
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw2.py
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.802
Model:                            OLS   Adj. R-squared:                  0.801
Method:                 Least Squares   F-statistic:                     1958.
Date:                Mon, 25 Feb 2019   Prob (F-statistic):               0.00
Time:                        17:05:58   Log-Likelihood:                -28248.
No. Observations:               25746   AIC:                         5.660e+04
Df Residuals:                   25692   BIC:                         5.704e+04
Df Model:                          53
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.2903      1.081      7.671      0.000       6.172      10.409
x1            -1.8145      0.040    -45.489      0.000      -1.893      -1.736
x1:x2          0.2019      0.005     38.713      0.000       0.192       0.212
x1:x3          0.0023      0.001      2.710      0.007       0.001       0.004
x1:x4         -0.0019      0.001     -2.285      0.022      -0.004      -0.000
x1:x5         -0.0231      0.006     -4.060      0.000      -0.034      -0.012
x1:x6         -0.0630      0.011     -5.971      0.000      -0.084      -0.042
x1:x7         -0.0126      0.001    -16.506      0.000      -0.014      -0.011
x1:x8         -0.0002      0.000     -1.355      0.175      -0.000    7.76e-05
x1:x9       -9.19e-05   9.65e-06     -9.525      0.000      -0.000    -7.3e-05
x1:x10         0.0164      0.002      7.370      0.000       0.012       0.021
x2            -0.3758      0.162     -2.320      0.020      -0.693      -0.058
x2:x3         -0.0896      0.010     -9.122      0.000      -0.109      -0.070
x2:x4          0.0943      0.010      9.781      0.000       0.075       0.113
```

We can find the original Adj. R-squared value is 0.801, but there are some meaningless variables in this model, then we remove variables based on their p-value from greatest to smallest which are greater than our cutoff value. Following is the result:

```
mod014 = smf.ols(formula = 'y ~ x1 + x1:x2 + x1:x3 + x1:x4 + x1:x5 + x1:x6 + x1:x7 + x1:x9 + x1:x10 + '
                           'x2 + x2:x3 + x2:x4 + x2:x5 + x2:x7 + x2:x8 + x2:x10 + '
                           'x3 + x3:x4 + x3:x5 + x3:x7 + x3:x8 + x3:x9 + '
                           'x4 + x4:x5 + x4:x7 + x4:x8 + x4:x9 + x4:x10 + '
                           'x5 + x5:x7 + x5:x9 + x5:x10 + '
                           'x6 + x6:x10 + '
                           'x7 + '
                           'x8 + x8:x9 + '
                           'x9 ', data = data).fit()
sum014 = mod014.summary()
print(sum014)
```

hw2

```
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw2.py
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.801
Method:                 Least Squares   F-statistic:                     2803.
Date:                Mon, 25 Feb 2019   Prob (F-statistic):               0.00
Time:                        17:23:47   Log-Likelihood:                -28264.
No. Observations:               25746   AIC:                         5.660e+04
Df Residuals:                   25708   BIC:                         5.691e+04
Df Model:                          37
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.9287      0.717     11.065      0.000       6.524       9.333
x1            -1.8424      0.034    -54.005      0.000      -1.909      -1.776
x1:x2          0.2039      0.005     40.710      0.000       0.194       0.214
x1:x3          0.0026      0.001      3.132      0.002       0.001       0.004
x1:x4         -0.0022      0.001     -2.706      0.007      -0.004      -0.001
x1:x5         -0.0179      0.004     -4.269      0.000      -0.026      -0.010
x1:x6         -0.0593      0.010     -5.816      0.000      -0.079      -0.039
x1:x7         -0.0120      0.001    -18.462      0.000      -0.013      -0.011
x1:x9      -9.284e-05   9.63e-06     -9.642      0.000      -0.000    -7.4e-05
```

After we remove some meaningless variables, the Adj. R-squared value is still 0.801.

3. Using at least one nonlinear term to improve the MLR model.

My nonlinear term is to replace X3 to Ln(X3), and I use Backward variable selection method:

```
data1 = data.copy()
data1['x3'] = np.log(data['x3'])

mod013 = smf.ols(formula = 'y ~ x1 + x1:x2 + x1:x3 + x1:x4 + x1:x5 + x1:x6 + x1:x7 + x1:x8 + x1:x9 + x1:x10 + '
                           'x2 + x2:x3 + x2:x5 + x2:x6 + x2:x7 + x2:x8 + x2:x9 + x2:x10 + '
                           'x3 + x3:x4 + x3:x5 + x3:x8 + x3:x9 + '
                           'x4 + x4:x6 + x4:x7 + x4:x8 + x4:x9 + '
                           'x5 + x5:x7 + x5:x8 + x5:x10 + '
                           'x6 + x6:x7 + x6:x10 + '
                           'x7:x8 + x7:x10 + '
                           'x8 + '
                           'x10', data = data1).fit()
sum013 = mod013.summary()
print(sum013)
```

```
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw2.py
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.810
Model:                            OLS   Adj. R-squared:                  0.810
Method:                 Least Squares   F-statistic:                     2886.
Date:                Mon, 25 Feb 2019   Prob (F-statistic):               0.00
Time:                        17:34:37   Log-Likelihood:                -27681.
No. Observations:               25746   AIC:                         5.544e+04
Df Residuals:                   25707   BIC:                         5.576e+04
Df Model:                          38
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     13.0947      0.878     14.910      0.000      11.373      14.816
x1            -1.7563      0.040    -43.888      0.000      -1.835      -1.678
x1:x2          0.2084      0.005     42.377      0.000       0.199       0.218
x1:x3         -0.0317      0.005     -6.908      0.000      -0.041      -0.023
x1:x4          0.0009   9.4e-05      9.936      0.000       0.001       0.001
x1:x5         -0.0203      0.006     -3.667      0.000      -0.031      -0.009
x1:x6         -0.0563      0.010     -5.438      0.000      -0.077      -0.036
x1:x7         -0.0113      0.001    -15.431      0.000      -0.013      -0.010
x1:x8         -0.0004      0.000     -2.913      0.004      -0.001      -0.000
x1:x9      -9.625e-05   9.45e-06    -10.187      0.000      -0.000   -7.77e-05
x1:x10         0.0150      0.002      6.875      0.000       0.011       0.019
x2            -0.8170      0.132     -6.201      0.000      -1.075      -0.559
x2:x3          0.2163      0.016     13.211      0.000       0.184       0.248
x2:x5          0.4073      0.069      5.940      0.000       0.273       0.542
x2:x6          0.6050      0.124      4.882      0.000       0.362       0.848
x2:x7          0.0045      0.001      3.108      0.002       0.002       0.007
x2:x8          0.0071      0.001      5.498      0.000       0.005       0.010
x2:x9          0.0001   2.32e-05      5.937      0.000    9.24e-05       0.000
x2:x10        -0.0836      0.023     -3.689      0.000      -0.128      -0.039
x3            -2.1412      0.124    -17.313      0.000      -2.384      -1.899
```

By change X3 to a nonlinear term, the R-square value is raised to 0.810.

4. I ignore some variables which have really small coefficient, following is the result:

```
mod014 = smf.ols(formula = 'y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x1:x2 + x2:x6', data = data1).fit()
sum014= mod014.summary()
print(sum014)
```

```
hw2 ×
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw2.py
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.794
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     8254.
Date:                Mon, 25 Feb 2019   Prob (F-statistic):               0.00
Time:                        21:03:15   Log-Likelihood:                -28745.
No. Observations:               25746   AIC:                         5.752e+04
Df Residuals:                   25733   BIC:                         5.762e+04
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.3159      0.117     45.480      0.000       5.087       5.545
x1            -2.3132      0.020   -117.751      0.000      -2.352      -2.275
x2             0.2924      0.015     20.001      0.000       0.264       0.321
x3            -0.8625      0.013    -67.509      0.000      -0.888      -0.837
x4             0.0008      0.000      2.903      0.004       0.000       0.001
x5             0.4376      0.016     28.060      0.000       0.407       0.468
x6            -3.5620      0.331    -10.757      0.000      -4.211      -2.913
x7             0.0375      0.002     18.376      0.000       0.033       0.041
x8             0.0106      0.000     31.139      0.000       0.010       0.011
x9            -0.0007   2.65e-05    -28.166      0.000      -0.001      -0.001
x10            0.0869      0.006     14.279      0.000       0.075       0.099
x1:x2          0.2700      0.003     95.848      0.000       0.265       0.276
x2:x6          0.5383      0.052     10.301      0.000       0.436       0.641
==============================================================================
Omnibus:                      179.897   Durbin-Watson:                   0.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              225.824
Skew:                          -0.125   Prob(JB):                     9.18e-50
Kurtosis:                       3.385   Cond. No.                     3.18e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In my final model, the R-squared value is 0.794.

My function is:

$$y = 5.3159 - 2.3132 * x1 + 0.2924 * x2 - 0.8625 * \ln(x3) + 0.0008 * x4 + 0.4376 * x5 - 3.562 * x6 + 0.0375 * x7 + 0.0106 * x8 - 0.0007 * x9 + 0.0869 * x10 + 0.27 * x1 * x2 + 0.5383 * x2 * x6$$

And the final regression model's formula is also in the Excel file.