Lab 1: Applied data science

1) **Which variables have the most explanatory power? Which have the least?**

The next method is to use pearson test to compare both variables:

```
df = pd.read_csv("trafficking_data.csv")
print pearsonr(df["persons prosecuted"],df["Adult victims"])
……and so on

person prosecuted VS adult victim:
(-0.048430976740660318, 0.54565160155960302)
child victim VS adult victim:
(-0.035024116929785513, 0.66219496601816152)
gdp VS adult victim:
(0.028646394597412601, 0.72087188051004913)
life expectancy VS adult victim:
(0.049826116824534659, 0.53412741414372966)
Female primary education VS adult victim:
(-0.10034340943285375, 0.20967825172658733)
policy index VS adult victim:
(0.063669987031620329, 0.42674651732545388)
```

where the return value is (Pearson's correlation coefficient, 2-tailed p-value). Here we understand that almost no correlation between two of them (pearson coefficient varies from -1 to 1, close to zero implies no correlation). P-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. In addition, policy index seems to be dominant factor in predicting Adult victims.

Then we start implementing linear prediction model to try predicting each expected Y values by fixed x variables and compare the R squared value.

```
                          OLS Regression Results
==============================================================================
Dep. Variable:          Adult_victims   R-squared:                       0.106
Model:                            OLS   Adj. R-squared:                  0.089
Method:                 Least Squares   F-statistic:                     6.142
Date:                Wed, 08 Oct 2014   Prob (F-statistic):           0.000567
Time:                        18:24:05   Log-Likelihood:                 -1280.2
No. Observations:                 158   AIC:                             2566.
Df Residuals:                     155   BIC:                             2576.
Df Model:                           3
===============================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
gdp               -3.425e-12   3.06e-11     -0.112      0.911    -6.38e-11   5.7e-11
policy_index        19.3717     31.154      0.622      0.535      -42.169    80.913
life_expectancy      4.5127      8.912      0.506      0.613      -13.092    22.117
females_education   -4.4773     11.804     -0.379      0.705      -27.795    18.840
==============================================================================
Omnibus:                      198.128   Durbin-Watson:                   0.777
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5932.011
Skew:                           5.132   Prob(JB):                         0.00
Kurtosis:                      31.208   Cond. No.                     1.20e+12
==============================================================================
```

Warnings:

[1] The condition number is large, 1.2e+12. This might indicate that there are strong multicollinearity or other numerical problems.

```
Parameters: gdp             -3.425145e-12
policy_index         1.937171e+01
life_expectancy      4.512711e+00
females_education   -4.477272e+00
dtype: float64
```

Warnings:

[1] The condition number is large, 1.2e+12. This might indicate that there are strong multicollinearity or other numerical problems.

```
Parameters: gdp                -2.586435e-12
policy_index          5.257530e+00
life_expectancy      -1.838591e+00
females_education     2.201553e+00
dtype: float64
```

```
                        OLS Regression Results
==============================================================================
Dep. Variable:     persons_prosecuted   R-squared:                       0.034
Model:                            OLS   Adj. R-squared:                  0.015
Method:                 Least Squares   F-statistic:                     1.828
Date:                Wed, 08 Oct 2014   Prob (F-statistic):              0.144
Time:                        18:24:05   Log-Likelihood:                 -1510.8
No. Observations:                 158   AIC:                             3028.
Df Residuals:                     155   BIC:                             3037.
Df Model:                           3
==============================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
gdp                5.426e-12   1.32e-10      0.041      0.967     -2.55e-10   2.65e-10
policy_index        101.7509    134.108      0.759      0.449     -163.163    366.665
life_expectancy     -32.9970     38.363     -0.860      0.391     -108.779     42.785
females_education    36.2056     50.813      0.713      0.477      -64.169    136.581
==============================================================================
Omnibus:                      224.600   Durbin-Watson:                   0.538
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10002.180
Skew:                           6.162   Prob(JB):                         0.00
Kurtosis:                      39.979   Cond. No.                     1.20e+12
==============================================================================

Warnings:
[1] The condition number is large, 1.2e+12. This might indicate that there are
strong multicollinearity or other numerical problems.
Parameters: gdp                5.425977e-12
policy_index          1.017509e+02
life_expectancy      -3.299698e+01
females_education     3.620564e+01
```

```
dtype: float64
```

```
    R^2 of results_victims 0.106242258731
    R^2 of results_prosecuted 0.0341707197668
```

From the result above we could see that R squared for adult victim is larger than results prosecuted. Therefore, we will use Adult victims as the value to predict in this assignment.

2) ***Remove some the outlier countries, how does this effect your model?*** -----

To get a better image on how removing outlier countries, we can observe the distribution of influential plot (H leverage VS studentized residuals). Here we remove Brazil who has unevenly distributed residual value, bigger than the standardized value of 3:
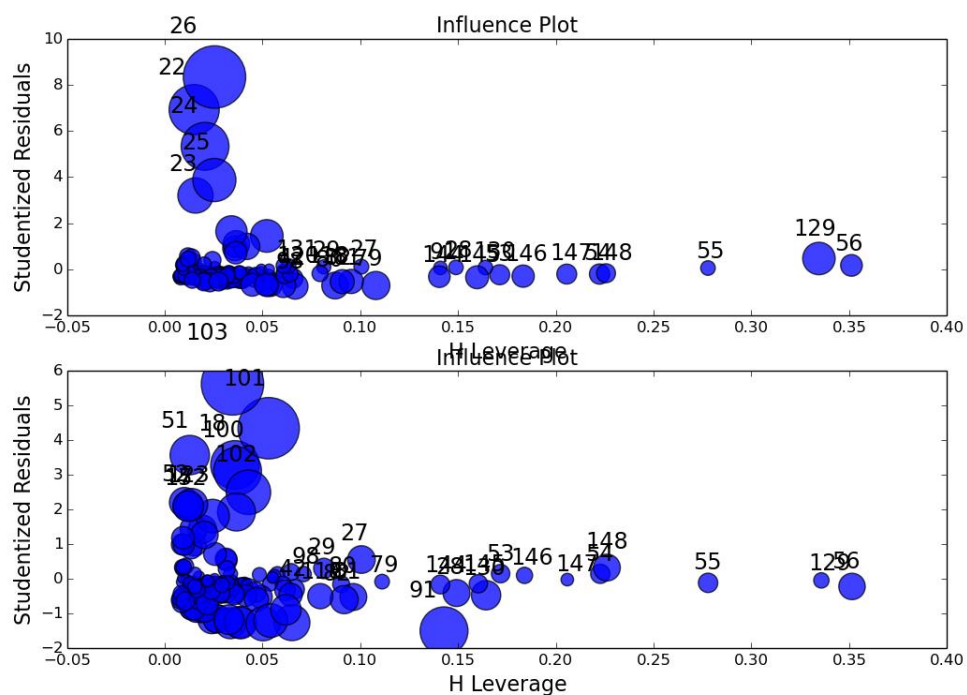
More detail summary of the coefficient can be seen here:

```
                        OLS Regression Results
=============================================================================
Dep. Variable:        Adult_victims    R-squared:                  0.024
Model:                         OLS    Adj. R-squared:            -0.002
```

```
Method:              Least Squares   F-statistic:                     0.9258

Date:             Wed, 08 Oct 2014   Prob (F-statistic):              0.451

Time:                     18:24:12   Log-Likelihood:                 -1278.7

No. Observations:              158   AIC:                             2567.

Df Residuals:                  153   BIC:                             2583.

Df Model:                        4

=================================================================================
                        coef     std err           t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
const              6367.3284    4201.836       1.515      0.132   -1933.778   1.47e+04

gdp                4.989e-12    3.12e-11       0.160      0.873    -5.66e-11   6.65e-11

policy_index         16.6306      31.606       0.526      0.600     -45.810     79.071

females_education  -130.7897      84.940      -1.540      0.126    -298.596     37.017

life_expectancy       1.3003       9.139       0.142      0.887     -16.755     19.356

persons_prosecuted   -0.0182       0.019      -0.948      0.344      -0.056      0.020

child_victims        -0.6412       0.768      -0.835      0.405      -2.159      0.876

=================================================================================
Omnibus:                   191.304   Durbin-Watson:                   0.794

Prob(Omnibus):               0.000   Jarque-Bera (JB):             5184.289

Skew:                        4.886   Prob(JB):                         0.00

Kurtosis:                   29.306   Cond. No.                     1.61e+14

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


Warnings:

[1] The condition number is large, 1.61e+14. This might indicate that there are

strong multicollinearity or other numerical problems.

Outlier:  [22, 23, 24, 25, 26]
   country   year  persons_prosecuted  Adult_victims  child_victims  \
22  Brazil  2003                  52           5223              0

23  Brazil  2004                 130           2887              0

24  Brazil  2005                 128           4348              0

25  Brazil  2006                 117           3417              0

26  Brazil  2007                 200           5975              0


           gdp  policy_index  females_education  life_expectancy
22  5.524693e+11            10           47.71990               71
```

```
23  6.637603e+11            11            47.70676            71

24  8.821857e+11            12            47.59256            71

25  1.088917e+12            11            47.20861            71

26  1.366824e+12            11            47.20861            71
```

                                OLS Regression Results

==

Dep. Variable:	Adult_victims	R-squared:	0.153
Model:	OLS	Adj. R-squared:	0.130
Method:	Least Squares	F-statistic:	6.683
Date:	Wed, 08 Oct 2014	Prob (F-statistic):	5.68e-05
Time:	18:24:13	Log-Likelihood:	-1037.2
No. Observations:	153	AIC:	2084.
Df Residuals:	148	BIC:	2100.
Df Model:	4		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	-3407.2866	1153.837	-2.953	0.004	-5687.410	-1127.163
gdp	8.754e-12	8.38e-12	1.044	0.298	-7.81e-12	2.53e-11
policy_index	-0.3600	8.523	-0.042	0.966	-17.203	16.483
females_education	82.6792	23.400	3.533	0.001	36.438	128.920
life_expectancy	-7.1481	2.466	-2.899	0.004	-12.021	-2.275
persons_prosecuted	-0.0029	0.005	-0.555	0.580	-0.013	0.007
child_victims	0.3980	0.208	1.913	0.058	-0.013	0.809

==

Omnibus:	91.217	Durbin-Watson:	0.738
Prob(Omnibus):	0.000	Jarque-Bera (JB):	417.753
Skew:	2.289	Prob(JB):	1.93e-91
Kurtosis:	9.676	Cond. No.	1.64e+14

==

Warnings:

[1] The condition number is large, 1.64e+14. This might indicate that there are

strong multicollinearity or other numerical problems.

From the summary above we could see that R squared increased from 0.024 to 0.153, showing that the model works better without outliers.

3) *Log-scale each of the variables, how does this change your model?   Does it improve the models predictive power?   How can you tell?*

After scales were changed to logarithmic, we could observer that R squared increased significantly:

```
                        OLS Regression Results
==============================================================================
Dep. Variable:          Adult_victims   R-squared:                       0.153
Model:                            OLS   Adj. R-squared:                  0.130
Method:                 Least Squares   F-statistic:                     6.683
Date:                Wed, 08 Oct 2014   Prob (F-statistic):           5.68e-05
Time:                        19:50:51   Log-Likelihood:                 -1037.2
No. Observations:                 153   AIC:                             2084.
Df Residuals:                     148   BIC:                             2100.
Df Model:                           4
==================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept          -3407.2866   1153.837     -2.953      0.004   -5687.410  -1127.163
persons_prosecuted    -0.0029      0.005     -0.555      0.580      -0.013      0.007
child_victims          0.3980      0.208      1.913      0.058      -0.013      0.809
gdp                 8.754e-12   8.38e-12      1.044      0.298   -7.81e-12   2.53e-11
policy_index          -0.3600      8.523     -0.042      0.966     -17.203     16.483
females_education     82.6792     23.400      3.533      0.001      36.438    128.920
life_expectancy       -7.1481      2.466     -2.899      0.004     -12.021     -2.275
==============================================================================
Omnibus:                       91.217   Durbin-Watson:                   0.738
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              417.753
Skew:                           2.289   Prob(JB):                     1.93e-91
Kurtosis:                       9.676   Cond. No.                          nan
==============================================================================
```

Warnings:

[1] The smallest eigenvalue is -6.49e+08. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

```
                          OLS Regression Results
==============================================================================
Dep. Variable:        Adult_victims   R-squared:                     0.454
Model:                          OLS   Adj. R-squared:                0.432
Method:               Least Squares   F-statistic:                   20.27
Date:              Wed, 08 Oct 2014   Prob (F-statistic):         3.63e-17
Time:                      19:50:51   Log-Likelihood:              -1003.6
No. Observations:               153   AIC:                           2021.
Df Residuals:                   146   BIC:                           2042.
Df Model:                         6
======================================================================================
                                 coef    std err          t      P>|t|      [95.0%
Conf. Int.]
--------------------------------------------------------------------------------------
Intercept                   -1.001e+04   3555.325     -2.815      0.006     -1.7e+04
-2981.212
np.log1p(persons_prosecuted)   29.3962      7.357      3.995      0.000       14.855
43.937
np.log1p(child_victims)        24.5572      7.602      3.231      0.002        9.534
39.581
np.log1p(gdp)                 -32.4480      3.509     -9.246      0.000      -39.384
-25.512
np.log1p(policy_index)         82.4559     53.638      1.537      0.126      -23.552
188.464
np.log1p(females_education)  2615.0982    911.383      2.869      0.005      813.891
4416.306
np.log1p(life_expectancy)      97.1963    137.693      0.706      0.481     -174.932
369.325
==============================================================================
Omnibus:                       68.088   Durbin-Watson:                 1.049
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            198.134
```

```
Skew:                        1.820    Prob(JB):                    9.46e-44

Kurtosis:                    7.222    Cond. No.                    6.52e+03

================================================================================
```

Warnings:

[1] The condition number is large, 6.52e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Therefore, we may conclude logarithmic conversion did improve the predictive power of this particular regression model.

4) ***Can you think of any other modeling techniques (from class) that could be used instead of linear regression? Try using one of these and explain your results, with diagrams and if possible, a visualization as well as descriptive statistics.***

One of the other model that was discussed in the class (not lab) were polinomial. In this assignment linear, 2nd order and 3rd order of polynomial regressions were tested. It can be done by implementing formula in statsmodels OLS feature:

```python
# 3-rd order polynomial
poly_3 = smf.ols(formula='Adult_victims~ 1 + persons_prosecuted +
child_victims+    gdp+    policy_index+    females_education    +
life_expectancy+I(persons_prosecuted ** 2.0) + I(child_victims ** 2.0)+
I(gdp ** 2.0) + I(policy_index ** 2.0)+I(females_education ** 2.0) +
I(life_expectancy ** 2.0)+I(persons_prosecuted ** 3.0) + I(child_victims
** 3.0)+ I(gdp ** 3.0) + I(policy_index ** 3.0)+I(females_education ** 3.0)
+ I(life_expectancy ** 3.0)', data=df).fit()
print poly_3.summary()
plt.plot(x, poly_3.predict(X), 'go', label='Poly n=3 $R^2$=%.2f' %
poly_3.rsquared,
        alpha=0.9)
```

The statistic summary:

```
                      OLS Regression Results

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

Dep. Variable:        Adult_victims    R-squared:                   0.454

Model:                          OLS    Adj. R-squared:              0.432

Method:               Least Squares    F-statistic:                 20.27
```

```
Date:                Wed, 08 Oct 2014   Prob (F-statistic):         3.63e-17
Time:                        22:03:28   Log-Likelihood:              -1003.6
No. Observations:                 153   AIC:                           2021.
Df Residuals:                     146   BIC:                           2042.
Df Model:                           6
==============================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const             -1.001e+04   3555.325     -2.815      0.006    -1.7e+04  -2981.212
gdp                 -32.4480      3.509     -9.246      0.000     -39.384    -25.512
policy_index         82.4559     53.638      1.537      0.126     -23.552    188.464
females_education  2615.0982    911.383      2.869      0.005     813.891   4416.306
life_expectancy      97.1963    137.693      0.706      0.481    -174.932    369.325
persons_prosecuted   29.3962      7.357      3.995      0.000      14.855     43.937
child_victims        24.5572      7.602      3.231      0.002       9.534     39.581
==============================================================================
Omnibus:                       68.088   Durbin-Watson:                  1.049
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             198.134
Skew:                           1.820   Prob(JB):                     9.46e-44
Kurtosis:                       7.222   Cond. No.                     6.52e+03
==============================================================================
```

Warnings:

[1] The condition number is large, 6.52e+03. This might indicate that there are strong multicollinearity or other numerical problems.

```
                         OLS Regression Results
==============================================================================
Dep. Variable:         Adult_victims   R-squared:                      0.673
Model:                           OLS   Adj. R-squared:                 0.629
Method:                Least Squares   F-statistic:                    15.29
Date:                Wed, 08 Oct 2014   Prob (F-statistic):          2.45e-24
Time:                       22:03:28   Log-Likelihood:                -964.53
No. Observations:                153   AIC:                            1967.
Df Residuals:                    134   BIC:                            2025.
Df Model:                         18
==============================================================================
```

```
========
                                coef    std err          t    P>|t|     [95.0%
Conf. Int.]
--------------------------------------------------------------------------------
--------
Intercept              -5.723e+07   1.04e+08    -0.551    0.582    -2.62e+08
1.48e+08
persons_prosecuted       -133.4874     33.380    -3.999    0.000    -199.508
-67.467
child_victims             -15.8600     59.138    -0.268    0.789    -132.825
101.105
gdp                      -107.1898     47.483    -2.257    0.026    -201.103
-13.276
policy_index             -637.5839    504.583    -1.264    0.209    -1635.560
360.393
females_education          4.34e+07      8e+07     0.542    0.588     -1.15e+08
2.02e+08
life_expectancy          4.193e+05     3.9e+05     1.075    0.284     -3.52e+05
1.19e+06
I(persons_prosecuted ** 2.0)    50.3547     10.252     4.912    0.000      30.079
70.631
I(child_victims ** 2.0)     8.6766     27.626     0.314    0.754     -45.962
63.315
I(gdp ** 2.0)               3.8014      3.751     1.013    0.313      -3.618
11.221
I(policy_index ** 2.0)     601.7357    406.774     1.479    0.141    -202.792
1406.263
I(females_education ** 2.0)  -1.108e+07   2.06e+07    -0.538    0.591     -5.18e+07
2.97e+07
I(life_expectancy ** 2.0)   -9.996e+04     9.4e+04    -1.064    0.289     -2.86e+05
8.59e+04
I(persons_prosecuted ** 3.0)    -3.8735      0.737    -5.257    0.000      -5.331
-2.416
I(child_victims ** 3.0)     -0.1597      3.171    -0.050    0.960      -6.431
6.112
I(gdp ** 3.0)               -0.0361      0.075    -0.480    0.632      -0.185
```

```
0.113

I(policy_index ** 3.0)         -131.4920     86.698     -1.517     0.132     -302.965
39.981

I(females_education ** 3.0)    9.435e+05    1.77e+06     0.534     0.594     -2.55e+06
4.44e+06

I(life_expectancy ** 3.0)      7933.6415    7540.870     1.052     0.295     -6980.886
2.28e+04

==============================================================================
Omnibus:                        69.548    Durbin-Watson:                   1.413
Prob(Omnibus):                   0.000    Jarque-Bera (JB):              283.826
Skew:                            1.678    Prob(JB):                     2.33e-62
Kurtosis:                        8.766    Cond. No.                     8.61e+10
==============================================================================


Warnings:
[1] The smallest eigenvalue is 4.91e-12. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```
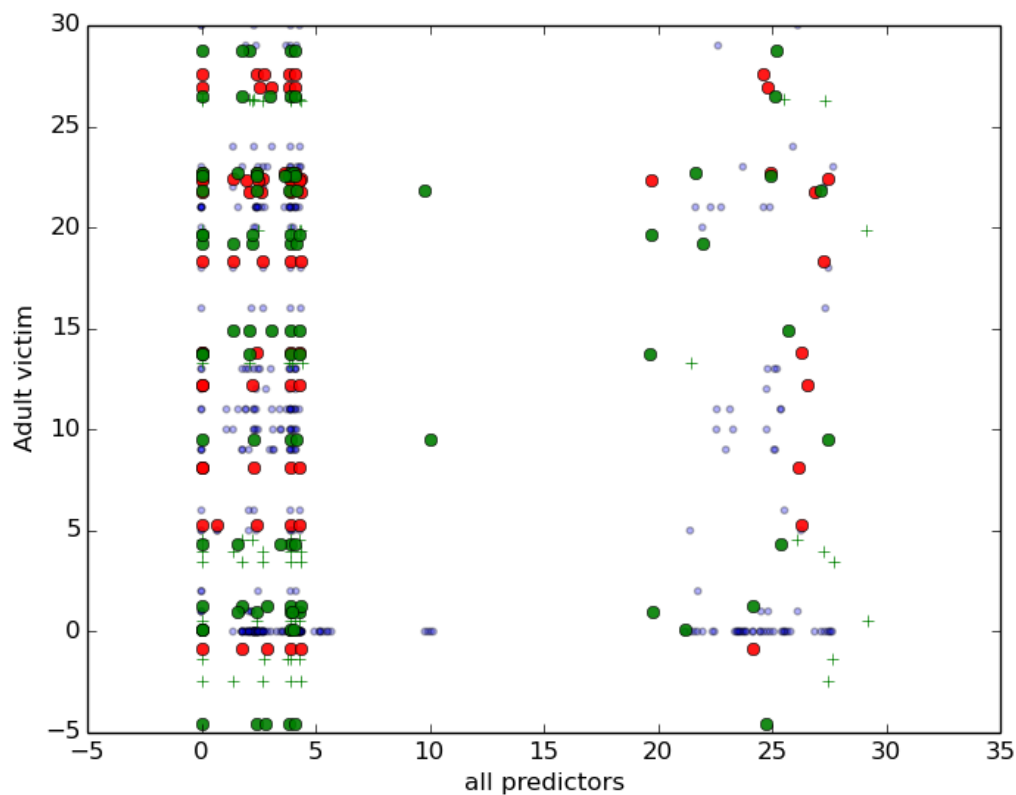
Therefore, we may conclude the best regression model is the one with highest R squared value, which is 3rd order of polynomial. In the following graphs, red is the 2nd order, green dots are the 3rd order polynomial, and plus sign shows regular linear regression.

5) *Think about how this model might be improved by adding more data.   Then add this data to the model and test your hypothesis.   What did you find.   Provide descriptive statistics and visualizations as well as a few paragraphs explaining how you chose what data you did and why.*

Testing model achieved from the test training datasets and apply the same model to new unemployment rate data sets:

```
                        OLS Regression Results
==============================================================================
Dep. Variable:         Adult_victims   R-squared:                    0.673
Model:                           OLS   Adj. R-squared:               0.629
Method:                Least Squares   F-statistic:                  15.29
```

```
Date:              Wed, 08 Oct 2014  Prob (F-statistic):         2.45e-24
Time:                     22:31:41  Log-Likelihood:             -964.53
No. Observations:              153  AIC:                           1967.
Df Residuals:                  134  BIC:                           2025.
Df Model:                       18
================================================================================
========
                          coef    std err          t      P>|t|      [95.0%
Conf. Int.]
--------------------------------------------------------------------------------
--------
Intercept               -5.723e+07   1.04e+08     -0.551      0.582     -2.62e+08
1.48e+08
persons_prosecuted       -133.4874     33.380     -3.999      0.000      -199.508
-67.467
child_victims             -15.8600     59.138     -0.268      0.789      -132.825
101.105
gdp                      -107.1898     47.483     -2.257      0.026      -201.103
-13.276
policy_index             -637.5839    504.583     -1.264      0.209     -1635.560
360.393
females_education          4.34e+07      8e+07      0.542      0.588     -1.15e+08
2.02e+08
life_expectancy          4.193e+05    3.9e+05      1.075      0.284     -3.52e+05
1.19e+06
I(persons_prosecuted ** 2.0)  50.3547    10.252      4.912      0.000        30.079
70.631
I(child_victims ** 2.0)     8.6766     27.626      0.314      0.754       -45.962
63.315
I(gdp ** 2.0)               3.8014      3.751      1.013      0.313        -3.618
11.221
I(policy_index ** 2.0)    601.7357    406.774      1.479      0.141      -202.792
1406.263
I(females_education ** 2.0) -1.108e+07   2.06e+07     -0.538      0.591     -5.18e+07
2.97e+07
I(life_expectancy ** 2.0) -9.996e+04    9.4e+04     -1.064      0.289     -2.86e+05
```

```
8.59e+04

I(persons_prosecuted ** 3.0)     -3.8735      0.737     -5.257     0.000      -5.331

-2.416

I(child_victims ** 3.0)          -0.1597      3.171     -0.050     0.960      -6.431

6.112

I(gdp ** 3.0)                    -0.0361      0.075     -0.480     0.632      -0.185

0.113

I(policy_index ** 3.0)         -131.4920     86.698     -1.517     0.132    -302.965

39.981

I(females_education ** 3.0)    9.435e+05    1.77e+06     0.534     0.594     -2.55e+06

4.44e+06

I(life_expectancy ** 3.0)      7933.6417   7540.870      1.052     0.295    -6980.886

2.28e+04

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

Omnibus:                         69.548    Durbin-Watson:                     1.413

Prob(Omnibus):                    0.000    Jarque-Bera (JB):                283.826

Skew:                             1.678    Prob(JB):                       2.33e-62

Kurtosis:                         8.766    Cond. No.                            nan

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


Warnings:

[1] The smallest eigenvalue is -5.15e-12. This might indicate that there are

strong multicollinearity problems or that the design matrix is singular.

                        OLS Regression Results

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

Dep. Variable:          Adult_victims    R-squared:                        0.673

Model:                           OLS    Adj. R-squared:                    0.621

Method:                Least Squares    F-statistic:                       12.85

Date:             Wed, 08 Oct 2014    Prob (F-statistic):              1.26e-22

Time:                       22:31:42    Log-Likelihood:                  -964.38

No. Observations:                153    AIC:                               1973.

Df Residuals:                    131    BIC:                               2039.

Df Model:                         21

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

                             coef     std err          t     P>|t|      [95.0%
```

```
                             Conf. Int.]
------------------------------------------------------------------------------------------
--------
Intercept                    -5.783e+07  1.05e+08   -0.550   0.583   -2.66e+08
1.5e+08
persons_prosecuted            -132.5277    33.867   -3.913   0.000   -199.525
-65.530
child_victims                  -14.3127    59.915   -0.239   0.812   -132.839
104.214
gdp                           -108.7413    48.593   -2.238   0.027   -204.870
-12.612
policy_index                  -651.9263   511.092   -1.276   0.204  -1662.988
359.135
females_education              4.389e+07  8.11e+07    0.541   0.589   -1.16e+08
2.04e+08
life_expectancy                4.045e+05  3.95e+05    1.023   0.308   -3.78e+05
1.19e+06
new_data                        -0.1280    19.341   -0.007   0.995   -38.389
38.133
I(persons_prosecuted ** 2.0)    50.3116    10.404    4.836   0.000    29.729
70.894
I(child_victims ** 2.0)          7.8654    28.000    0.281   0.779   -47.524
63.255
I(gdp ** 2.0)                    3.9073     3.825    1.021   0.309    -3.660
11.474
I(policy_index ** 2.0)         611.7667   411.725    1.486   0.140   -202.724
1426.257
I(females_education ** 2.0)   -1.121e+07  2.09e+07   -0.537   0.592   -5.25e+07
3.01e+07
I(life_expectancy ** 2.0)     -9.637e+04  9.52e+04   -1.012   0.313   -2.85e+05
9.2e+04
I(new_data ** 2.0)              -5.8822    12.764   -0.461   0.646   -31.132
19.368
I(persons_prosecuted ** 3.0)    -3.8810     0.748   -5.189   0.000    -5.361
-2.401
I(child_victims ** 3.0)         -0.0741     3.213   -0.023   0.982    -6.430
```

```
6.282
I(gdp ** 3.0)                    -0.0383     0.077     -0.500     0.618        -0.190
0.113
I(policy_index ** 3.0)         -133.3417    87.717     -1.520     0.131       -306.866
40.183
I(females_education ** 3.0)    9.544e+05   1.79e+06     0.533     0.595       -2.59e+06
4.5e+06
I(life_expectancy ** 3.0)      7644.0742  7642.478      1.000     0.319       -7474.570
2.28e+04
I(new_data ** 3.0)               1.0261     6.075      0.169     0.866        -10.991
13.044
==============================================================================
Omnibus:                        68.626   Durbin-Watson:                   1.420
Prob(Omnibus):                   0.000   Jarque-Bera (JB):              281.148
Skew:                            1.650   Prob(JB):                      8.90e-62
Kurtosis:                        8.763   Cond. No.                          nan
==============================================================================

Warnings:
[1] The smallest eigenvalue is -1.37e-11. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

For initial testing random numbers was used to test whether the model consistently maintain the R squared given the polynomial effect. We could conclude that adding random variables to this model does not significantly change the R squared.
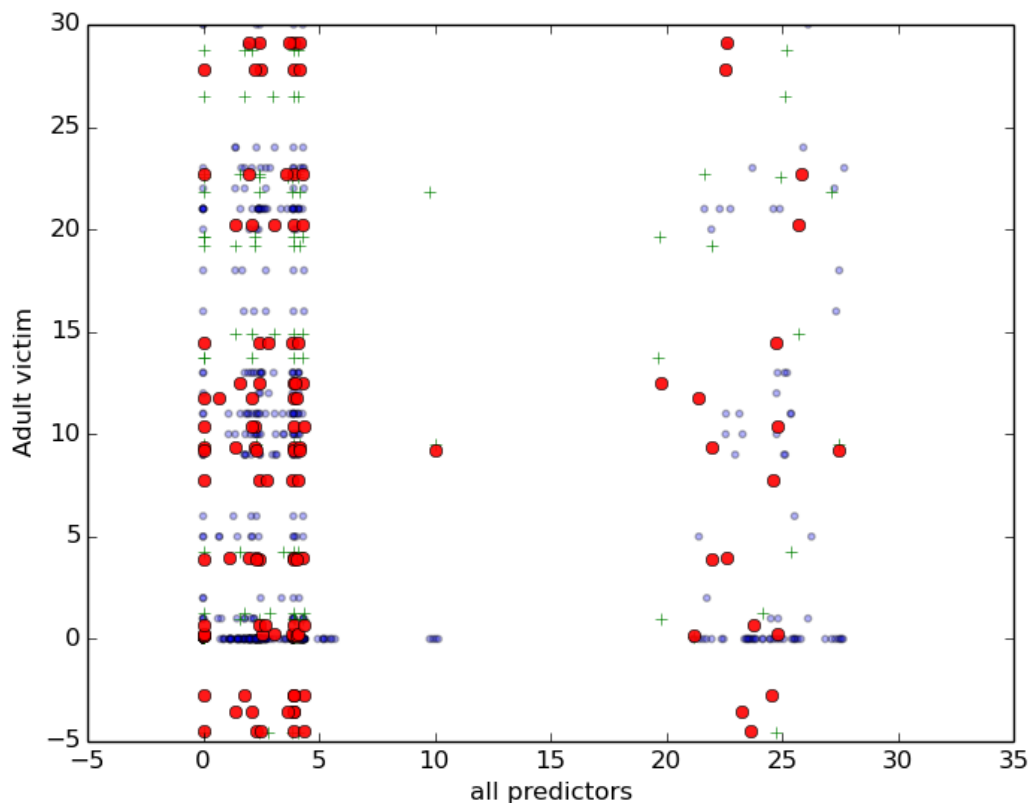
6)   *Using the model and data discussed in class predict how many cases a set of "new countries" would have (data to be provided in a separate csv file) Provide visualizations and a few paragraphs explaining your results.*

Note that new.csv does not have Adult victim, child victim and people prosecution value, and to get the expected value, we recalculate the model using the same number of x variables (fields) as the new sets. Note that the model predictive power decreased dramatically because it is used against different set of x values (this time all other expected values such as persons prosecuted):

```
                        OLS Regression Results
===============================================================================
Dep. Variable:        Adult_victims   R-squared:                    -0.315

Model:                          OLS   Adj. R-squared:               -0.315
```

```
Method:              Least Squares  F-statistic:                  -inf

Date:            Thu, 09 Oct 2014  Prob (F-statistic):            nan

Time:                    00:17:07  Log-Likelihood:            -1070.9

No. Observations:             153  AIC:                          2144.

Df Residuals:                 152  BIC:                          2147.

Df Model:                       0
================================================================================
=======

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

Int.]
--------------------------------------------------------------------------------
-------

Intercept              7.208e-48   1.28e-47      0.563      0.574     -1.81e-47

3.25e-47

gdp                   -3.066e-37   5.44e-37     -0.563      0.574     -1.38e-36

7.69e-37

policy_index          -2.808e-74   6.99e-74     -0.402      0.688     -1.66e-73

1.1e-73

females_education      3.482e-46   6.18e-46      0.563      0.574     -8.73e-46

1.57e-45

life_expectancy        4.694e-46   8.34e-46      0.563      0.574     -1.18e-45

2.12e-45

I(gdp ** 2.0)          3.778e-24   6.71e-24      0.563      0.574     -9.48e-24

1.7e-23

I(policy_index ** 2.0) 6.994e-46   1.24e-45      0.563      0.574     -1.75e-45

3.15e-45

I(females_education ** 2.0) 1.683e-44   2.99e-44      0.563      0.574     -4.22e-44

7.59e-44

I(life_expectancy ** 2.0)    3.1e-44   5.51e-44      0.563      0.574     -7.78e-44

1.4e-43

I(gdp ** 3.0)         -2.197e-37   5.47e-37     -0.402      0.688      -1.3e-36

8.61e-37

I(policy_index ** 3.0)  7.37e-45   1.31e-44      0.563      0.574     -1.85e-44

3.32e-44

I(females_education ** 3.0) 8.132e-43   1.44e-42      0.563      0.574     -2.04e-42

3.67e-42
```

```
I(life_expectancy ** 3.0)     2.075e-42    3.69e-42       0.563     0.574    -5.21e-42
9.36e-42

==============================================================================
Omnibus:                       110.367   Durbin-Watson:                   0.569
Prob(Omnibus):                   0.000   Jarque-Bera (JB):              656.897
Skew:                            2.782   Prob(JB):                    2.27e-143
Kurtosis:                       11.490   Cond. No.                          nan
==============================================================================
```

Warnings:

[1] The smallest eigenvalue is -0.028. This might indicate that there are

strong multicollinearity problems or that the design matrix is singular.
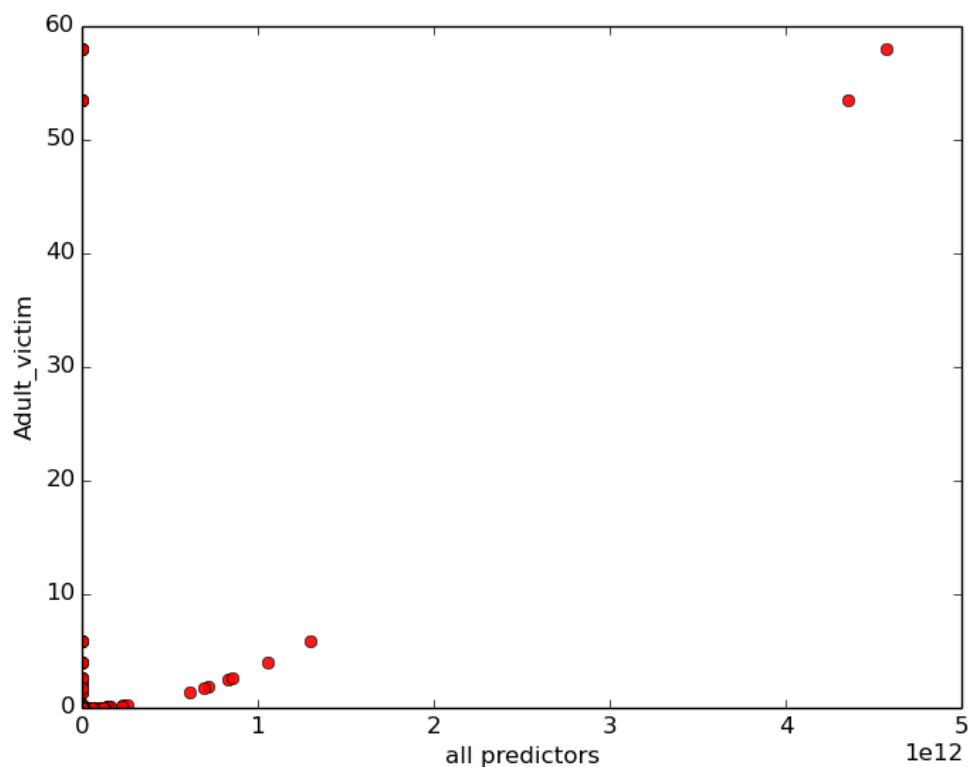
7)   *Try other models discussed from class.   What do these models predict and how do they differ from the linear regression model?*

In this problem I used weighted least squares, to get the statistic summary as follows:

```
                        WLS Regression Results
==============================================================================
Dep. Variable:          Adult_victims   R-squared:                      0.128
Model:                            WLS   Adj. R-squared:                 0.116
Method:                 Least Squares   F-statistic:                    11.02
Date:                Thu, 09 Oct 2014   Prob (F-statistic):          3.44e-05
Time:                        01:13:24   Log-Likelihood:               -654.63
No. Observations:                 153   AIC:                            1315.
Df Residuals:                     150   BIC:                            1324.
Df Model:                           2
===================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
const             -2951.2832   1089.960     -2.708      0.008    -5104.940  -797.626
gdp                8.351e-12   8.44e-12      0.989      0.324    -8.33e-12    2.5e-11
policy_index          0.9748      8.546      0.114      0.909      -15.912     17.861
females_education    74.1763     22.284      3.329      0.001       30.146    118.207
life_expectancy      -7.8345      2.436     -3.217      0.002      -12.647     -3.022
==============================================================================
Omnibus:                       88.244   Durbin-Watson:                  0.797
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             379.683
Skew:                           2.229   Prob(JB):                    3.57e-83
Kurtosis:                       9.299   Cond. No.                    1.54e+14
==============================================================================
```

Warnings:

[1] The condition number is large, 1.54e+14. This might indicate that there are
strong multicollinearity or other numerical problems.
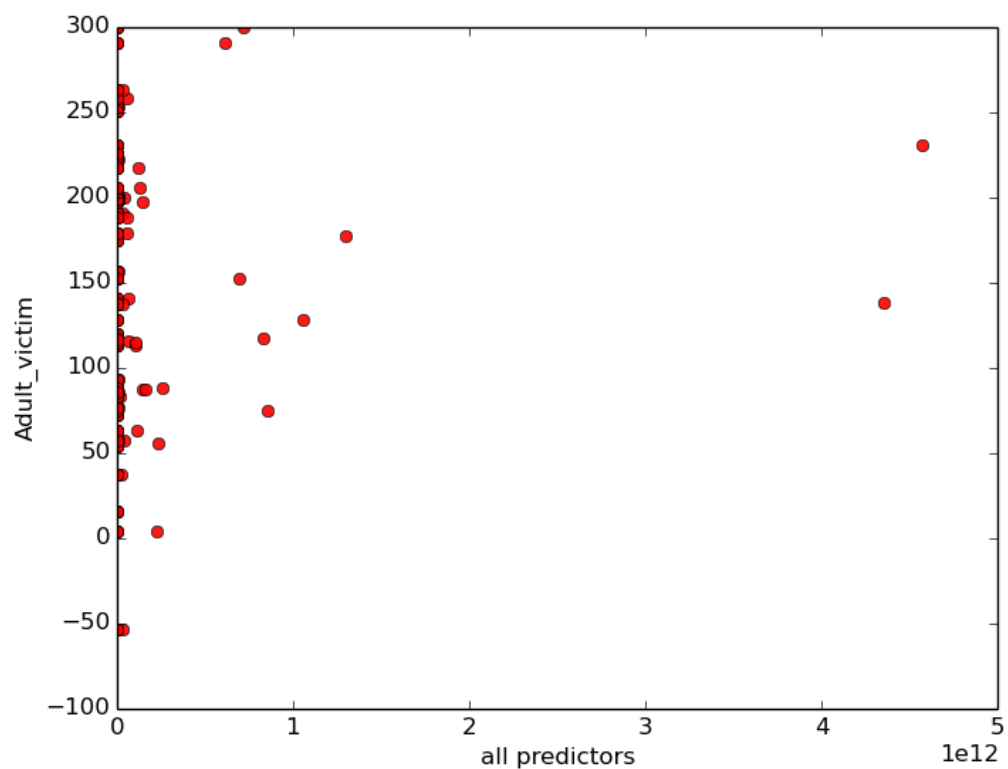
This prediction

The WLS is capable of resulting higher R squared value compared to the previous 3rd order polynomial model, but not as good as the original linier model.

8) ***Now remove the variables with the least explanatory power.   Does your linear regression improve compared to the other models?   Does it do worse?   Why? Please provide visuals and a few paragraphs of explanation***

From the summary we could see that gdb has the least explanatory power since it has smallest coefficient:

```
                       WLS Regression Results
==============================================================================
Dep. Variable:          Adult_victims   R-squared:                    0.122
Model:                            WLS   Adj. R-squared:               0.105
Method:                 Least Squares   F-statistic:                  6.925
Date:                Thu, 09 Oct 2014   Prob (F-statistic):        0.000214
```

```
Time:                        01:22:09  Log-Likelihood:                -655.12

No. Observations:                 153  AIC:                             1318.

Df Residuals:                     149  BIC:                             1330.

Df Model:                           3

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

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

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

const             -3161.9193  1076.032     -2.939      0.004   -5288.172 -1035.667

policy_index          4.1584     7.969      0.522      0.603     -11.589    19.906

females_education    77.4102    22.188      3.489      0.001      33.566   121.255

life_expectancy      -7.3936     2.410     -3.067      0.003     -12.157    -2.631

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

Omnibus:                       87.565  Durbin-Watson:                    0.798

Prob(Omnibus):                  0.000  Jarque-Bera (JB):               373.048

Skew:                           2.213  Prob(JB):                      9.85e-82

Kurtosis:                       9.239  Cond. No.                       4.99e+03

==================================================================================
```

Warnings:

[1] The condition number is large, 4.99e+03. This might indicate that there are strong multicollinearity or other numerical problems.

From here we know that WLS gives relatively equal predictive power compared to last result since we only omitted the variable that does not have significant effect to the whole model.

9) *Now add in the extra data you found.   Does your linear regression improved compared to the other models? Does it do worse? Why? Please provide visuals and a few paragraphs of explanation*

By adding unemployment variable it could be seen that the R squared and adjusted Rsquared does add up the value of R squared from 0.122 to 0.124.

```
                            WLS Regression Results
==============================================================================
Dep. Variable:         Adult_victims   R-squared:                       0.124
Model:                           WLS   Adj. R-squared:                  0.101
Method:                Least Squares   F-statistic:                     5.254
Date:               Thu, 09 Oct 2014   Prob (F-statistic):           0.000551
Time:                       11:24:36   Log-Likelihood:                -654.95
No. Observations:                153   AIC:                             1320.
Df Residuals:                    148   BIC:                             1335.
Df Model:                          4
==============================================================================
```

```
                    coef      std err        t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
const            -1.324e+04   4281.373    -3.092    0.002     -2.17e+04  -4777.917

policy_index       59.0554     64.946     0.909     0.365      -69.286    187.397

females_education 3918.4535  1097.999     3.569     0.000     1748.673   6088.234

life_expectancy  -473.9677    153.584    -3.086     0.002     -777.469   -170.466

unemployment      -27.6415     31.278    -0.884     0.378      -89.450     34.167

================================================================================
Omnibus:                      87.335   Durbin-Watson:                    0.798

Prob(Omnibus):                 0.000   Jarque-Bera (JB):               370.310

Skew:                          2.208   Prob(JB):                      3.88e-81

Kurtosis:                      9.212   Cond. No.                       1.63e+03

================================================================================
```

Warnings:

[1] The condition number is large, 1.63e+03. This might indicate that there are strong multicollinearity or other numerical problems.
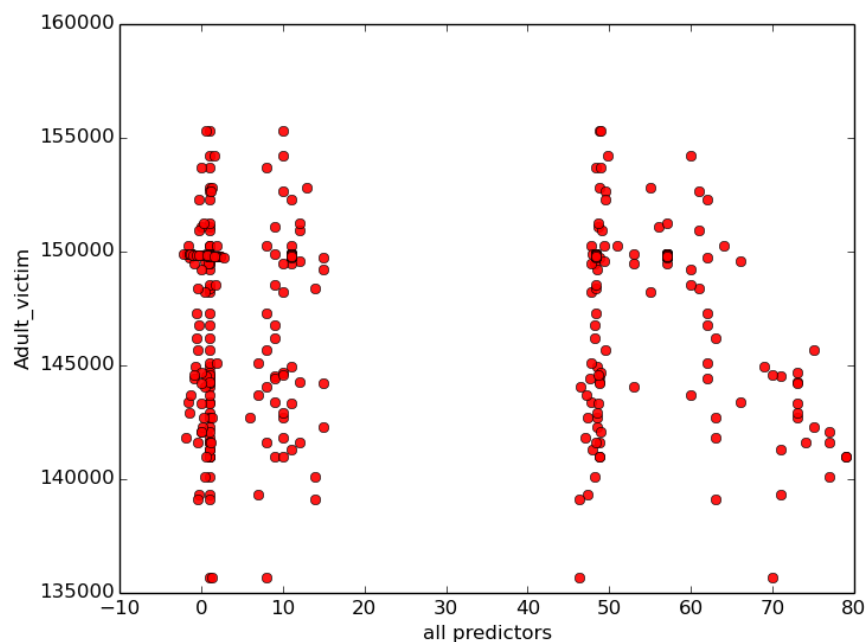

http://www.internetworldstats.com/

http://data.worldbank.org/indicator/IT.NET.USER.P2/countries

--sources of internet usage

http://www.internetlivestats.com/internet-users/

--number of connected devices

10)   *download (or scrape) data from the above websites.*

Downloaded from the website into CSV then run scripts to add to adding into main csv.

11)   *How much explanatory power does the model gain by adding the amount of internet penetration in a given country?   How much does adding the total number of connected devices add?*

We would like to know how much the know the explanatory power the model gain, so we run a multivariate linear regression:

```
                        OLS Regression Results
==============================================================================
Dep. Variable:        Adult_victims   R-squared:                    0.153
Model:                          OLS   Adj. R-squared:               0.130
Method:               Least Squares   F-statistic:                  6.683
Date:              Thu, 09 Oct 2014   Prob (F-statistic):        5.68e-05
Time:                      09:51:59   Log-Likelihood:              -1037.2
No. Observations:               153   AIC:                          2084.
Df Residuals:                   148   BIC:                          2100.
Df Model:                         4
==============================================================================
                       coef    std err          t      P>|t|     [95.0% Conf. Int.]
------------------------------------------------------------------------------------
const              -3407.2646   1153.837     -2.953      0.004    -5687.387 -1127.142
gdp                 8.744e-12   8.37e-12      1.044      0.298      -7.8e-12  2.53e-11
policy_index          -0.3599      8.523     -0.042      0.966      -17.203    16.483
females_education     82.6786     23.400      3.533      0.001       36.437   128.920
life_expectancy       -7.1480      2.466     -2.899      0.004      -12.021    -2.275
persons_prosecuted    -0.0029      0.005     -0.555      0.580       -0.013     0.007
child_victims          0.3980      0.208      1.913      0.058       -0.013     0.809
==============================================================================
```

```
Omnibus:                      91.217   Durbin-Watson:              0.738

Prob(Omnibus):                 0.000   Jarque-Bera (JB):         417.756

Skew:                          2.289   Prob(JB):                1.93e-91

Kurtosis:                      9.676   Cond. No.                1.64e+14

==============================================================================
```

Warnings:

[1] The condition number is large, 1.64e+14. This might indicate that there are strong multicollinearity or other numerical problems.

```
                           OLS Regression Results
==============================================================================
Dep. Variable:         Adult_victims   R-squared:                  0.164

Model:                           OLS   Adj. R-squared:             0.135

Method:                Least Squares   F-statistic:                5.760

Date:               Thu, 09 Oct 2014   Prob (F-statistic):      6.94e-05

Time:                       09:51:59   Log-Likelihood:           -1036.3

No. Observations:                153   AIC:                        2085.

Df Residuals:                    147   BIC:                        2103.

Df Model:                          5
===================================================================================
                      coef    std err          t      P>|t|    [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
const              -3437.2616  1150.537     -2.988      0.003  -5710.992 -1163.532

gdp                 1.536e-11  9.63e-12      1.595      0.113  -3.67e-12  3.44e-11

policy_index          -0.3827     8.497     -0.045      0.964    -17.175    16.410

females_education     80.6406    23.376      3.450      0.001     34.445   126.836

life_expectancy       -4.7612     3.006     -1.584      0.115    -10.702     1.180

persons_prosecuted    -0.0039     0.005     -0.746      0.457     -0.014     0.006

child_victims          0.3724     0.208      1.788      0.076     -0.039     0.784

internet_penet        -1.8272     1.324     -1.380      0.170     -4.445     0.790
==============================================================================
Omnibus:                      90.949   Durbin-Watson:              0.747

Prob(Omnibus):                 0.000   Jarque-Bera (JB):         417.275

Skew:                          2.279   Prob(JB):                2.45e-91

Kurtosis:                      9.684   Cond. No.                1.64e+14

==============================================================================
```

Warnings:

[1] The condition number is large, 1.64e+14. This might indicate that there are strong multicollinearity or other numerical problems.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Adult_victims   R-squared:                       0.156
Model:                            OLS   Adj. R-squared:                  0.127
Method:                 Least Squares   F-statistic:                     5.435
Date:                Thu, 09 Oct 2014   Prob (F-statistic):           0.000128
Time:                        09:51:59   Log-Likelihood:                 -1037.0
No. Observations:                 153   AIC:                             2086.
Df Residuals:                     147   BIC:                             2104.
Df Model:                           5
=================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
const              -3556.1615   1173.716     -3.030      0.003    -5875.698 -1236.625
gdp                -9.496e-12    2.65e-11     -0.359      0.720     -6.18e-11  4.28e-11
policy_index          -1.1579      8.607     -0.135      0.893       -18.168    15.852
females_education     85.7290     23.811      3.600      0.000        38.674   132.784
life_expectancy       -7.1508      2.470     -2.895      0.004       -12.032    -2.269
persons_prosecuted    -0.0114      0.013     -0.888      0.376        -0.037     0.014
child_victims          0.4081      0.209      1.954      0.053        -0.005     0.821
connected_dev       8.891e-07    1.22e-06      0.726      0.469     -1.53e-06  3.31e-06
==============================================================================
Omnibus:                       92.312   Durbin-Watson:                   0.741
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              433.177
Skew:                           2.310   Prob(JB):                     8.65e-95
Kurtosis:                       9.826   Cond. No.                     1.67e+14
==============================================================================
```

Warnings:

[1] The condition number is large, 1.67e+14. This might indicate that there are strong multicollinearity or other numerical problems.

Similarly, this shows that internet penetration was the only additional variable that has positive correlation to adult victim.

12) ***Can you give an explanation of why or why not this does not add to the model's explanatory power?   Is there another variable you might take away that is related to these variables?***

The pearson test was conducted and resulted in the following:

```
Internet penetration VS adult victim:
(-0.17378602791144523, 0.03168508364587453)
Connected device VS adult victim:
(-0.023572270737661622, 0.77241153794652262)
```

Varies from -1 to 1, close to zero implies no correlation). This explained that the number of internet penetration have higher correlation to the object that we are interested in observing, Adult value.