Applied Data Science
Lab – 1
Peter Varshavsky

1. Which variables have the most explanatory power? Which have the least?

Modeling adult victims

None of the variables have predictive power if measured by R-squared (Table 1.2) when fitting ordinary linear regression of each variable other than `country` on `adult_victims` or `total_victims`. Fitting a multivariate model including all variables other than `country` and numbers of victims variables yields no statistically significant coefficients (Table 1.1).

Table 1.1: adult_victims OLS Regression Results

```
==============================================================================
Dep. Variable:          adult_victims   R-squared:                       0.015
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     1.158
Date:                Thu, 09 Oct 2014   Prob (F-statistic):              0.317
Time:                        14:24:24   Log-Likelihood:                -1264.2
No. Observations:                 156   AIC:                             2534.
Df Residuals:                     153   BIC:                             2543.
Df Model:                           2
====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept             0.0052      0.006      0.824      0.411        -0.007     0.018
gdp                3.251e-12   3.15e-11      0.103      0.918     -5.89e-11  6.54e-11
year                  2.3466      2.003      1.172      0.243        -1.610     6.303
policy_index         14.3208     31.975      0.448      0.655       -48.848    77.490
percent_fem_educ    -99.5681     81.699     -1.219      0.225      -260.973    61.836
life_expectancy       3.5322      8.983      0.393      0.695       -14.214    21.279
==============================================================================
Omnibus:                      190.943   Durbin-Watson:                   0.795
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5228.504
Skew:                           4.948   Prob(JB):                         0.00
Kurtosis:                      29.579   Cond. No.                     3.89e+15
==============================================================================
```

Table 1.2: Individual R-squared for adult_victims as dependent variable

```
========================
year               0.002
gdp                0.001
policy_index       0.004
percent_fem_educ   0.010
life_expectancy    0.003
person_prosecuted  0.002
```

Modeling child victims

For child victims the full main effects model shows that `year` (P = 0.002) and `percent_fem_educ` (P = 0.004) are significant, however the significance of `year` disappears when `gdp` is excluded from the model. The R-squared for the full model is 0.092. Output is given in Table 1.4, and the P-values from bivariate regressions in Table 1.3.

Table 1.3: Individual R-squared for child_victims as dependent variable

```
========================
year               0.005
gdp                0.004
policy_index       0.001
percent_fem_educ   0.058
life_expectancy    0.027
person_prosecuted  0.004
```

Table 1.4: child_victims OLS Regression Results

```
==============================================================================
Dep. Variable:          child_victims   R-squared:                       0.092
Model:                            OLS   Adj. R-squared:                  0.081
Method:                 Least Squares   F-statistic:                     7.793
Date:                Thu, 09 Oct 2014   Prob (F-statistic):           0.000598
Time:                        14:24:25   Log-Likelihood:                -913.48
No. Observations:                 156   AIC:                             1833.
Df Residuals:                     153   BIC:                             1842.
Df Model:                           2
====================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept            0.0014      0.001      2.049      0.042      4.83e-05     0.003
gdp              -5.142e-13   3.32e-12     -0.155      0.877     -7.08e-12  6.05e-12
year                 0.6815      0.211      3.222      0.002         0.264     1.099
policy_index         3.6156      3.377      1.071      0.286        -3.056    10.287
percent_fem_educ   -25.3439      8.628     -2.937      0.004       -42.390    -8.298
life_expectancy     -2.1545      0.949     -2.271      0.025        -4.029    -0.280
==============================================================================
Omnibus:                      153.536   Durbin-Watson:                   0.875
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2502.876
Skew:                           3.691   Prob(JB):                         0.00
Kurtosis:                      21.182   Cond. No.                     3.89e+15
==============================================================================
```

Modeling persons prosecuted victims

The variables *year* and *percent_fem_educ* appear to be significant, but *year* is unstable. Tables 1.5 and 1.6 provide multiple regression and bivariate R-squares respectively.

Table 1.5: persons_prosecuted OLS Regression Results

```
==============================================================================
Dep. Variable:     persons_prosecuted   R-squared:                       0.043
Model:                            OLS   Adj. R-squared:                  0.030
Method:                 Least Squares   F-statistic:                     3.401
Date:                Thu, 09 Oct 2014   Prob (F-statistic):             0.0359
Time:                        14:32:17   Log-Likelihood:                -1489.7
No. Observations:                 156   AIC:                             2985.
Df Residuals:                     153   BIC:                             2995.
Df Model:                           2
====================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept            0.0477      0.027      1.796      0.075        -0.005     0.100
```

```
gdp                 7.089e-11   1.34e-10       0.531      0.596     -1.93e-10  3.35e-10
year                  20.5959      8.500       2.423      0.017         3.803    37.388
policy_index          47.6318    135.723       0.351      0.726      -220.501   315.764
percent_fem_educ    -795.7608    346.788      -2.295      0.023     -1480.871  -110.650
life_expectancy      -42.3887     38.129      -1.112      0.268      -117.717    32.939
============================================================================== Omnibus:
214.526   Durbin-Watson:                   0.543  Prob(Omnibus):                0.000   Jarque-
Bera (JB):              8663.042  Skew:                             5.824   Prob(JB):
0.00 Kurtosis:                        37.599   Cond. No.                          3.89e+15
==============================================================================
```

Table 1.6: Individual R-squared for child_victims as dependent variable

```
========================
year                  0.002
gdp                   0.000
policy_index          0.001
percent_fem_educ      0.034
life_expectancy       0.005
```

2. Remove some the outlier countries, how does this affect your model?

For the remainder of the assignment I will focus on modeling `child_victims` using all available variables not including `persons_prosecuted`. I removed the following outliers: `gdp` (Japan and USA) (Plot 2.1). I chose not to use `persons_`prosecuted as a predictor since we are asked to use other predictors to estimate `persons_prosecuted` in another question. *`Figure 2.1`*

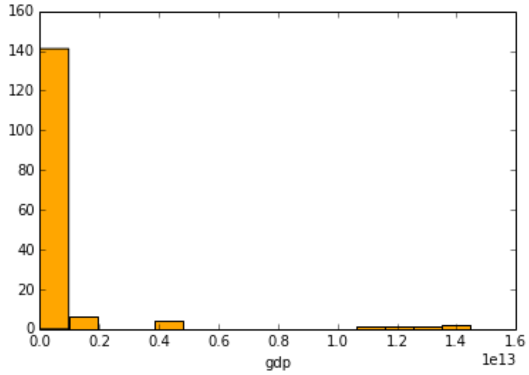

Figure 2.1. Histogram of GDP.

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

The predictive power is increased as measured by the higher R-squared and Adj. R-squared.

Table 3.1 child_victim log-log model OLS Regression Results

```
==============================================================================
Dep. Variable:      log_child_victims   R-squared:                       0.126
Model:                            OLS   Adj. R-squared:                  0.096
Method:                 Least Squares   F-statistic:                     4.200
Date:                Thu, 09 Oct 2014   Prob (F-statistic):            0.00135
Time:                        14:11:42   Log-Likelihood:                -275.06
No. Observations:                 152   AIC:                             562.1
Df Residuals:                     146   BIC:                             580.3
Df Model:                           5
==========================================================================================
                             coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------------
Intercept                 81.9845     29.518      2.777      0.006        23.648   140.321
log_gdp                    0.0301      0.031      0.968      0.335        -0.031     0.092
```

```
log_life_expectancy       -4.0978      1.116     -3.672      0.000        -6.303    -1.892
log_percent_fem_educ     -16.7676      7.628     -2.198      0.030       -31.843    -1.693
log_persons_prosecuted    -0.0351      0.086     -0.408      0.684        -0.205     0.135
log_policy_index           0.7247      0.513      1.412      0.160        -0.290     1.739
==============================================================================
Omnibus:                       19.328   Durbin-Watson:                   0.715
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.360
Skew:                           0.958   Prob(JB):                     8.46e-06
Kurtosis:                       3.124   Cond. No.                     6.26e+03
==============================================================================
```

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.

Classification methods can be used if the dependent variable is split into bins.

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.

Measures of poverty, employment, inequality, ratios of urban/rural populations, presence of conflicts or civil wars, racial and ethnic diversity, technology penetration and accessibility, educational attainment can be tried to improve the model.

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.

7. Try other models discussed from class. What do these models predict and how do they differ from the linear regression 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.

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.

10. Download (or scrape) data from the websites

- Sources of internet usage:
  http://www.internetworldstats.com/
  http://data.worldbank.org/indicator/IT.NET.USER.P2/countries
- Number of connected devices: http://www.internetlivestats.com/internet-users/

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?

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?