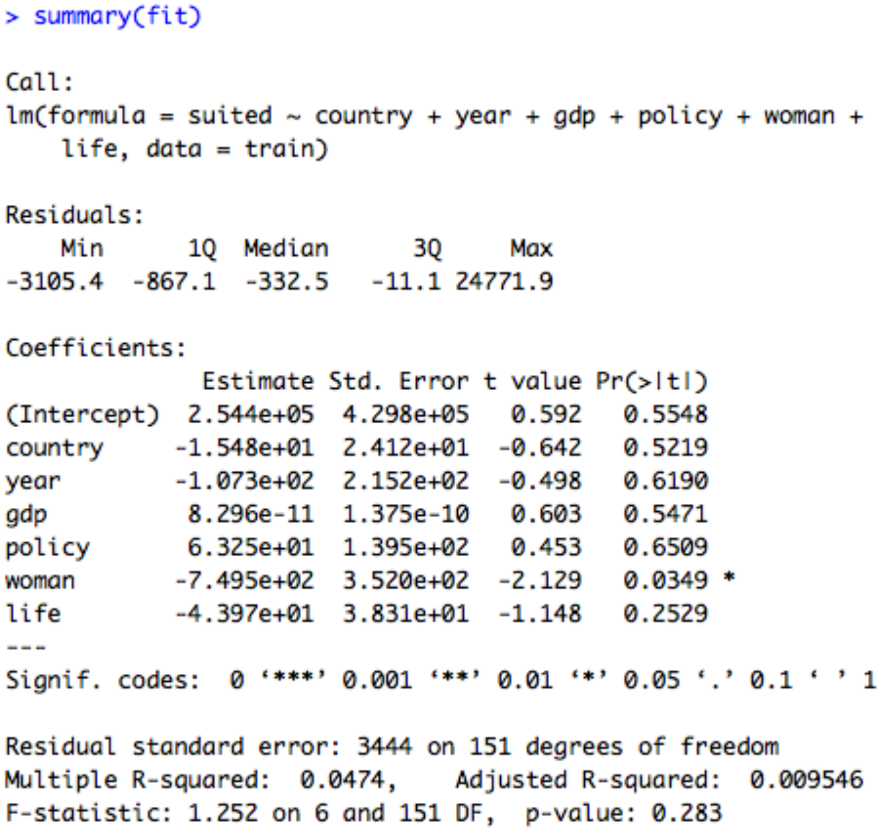
Lab 01

Applied Data Science-Multiple Linear Regression

Jiamin Xuan (jx624) - October 8, 2014

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

Load the csv file into R and clean the NA value, then fit the model, the summary of the model:



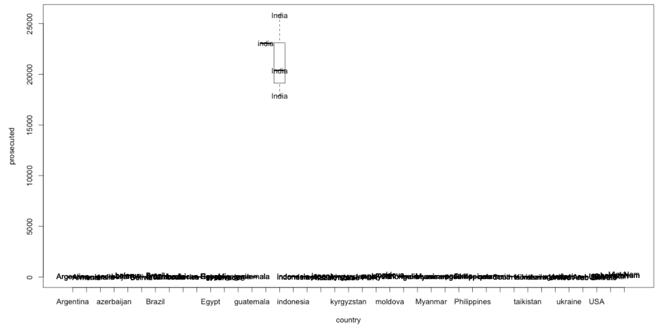
```
> summary(fit)

Call:
lm(formula = suited ~ country + year + gdp + policy + woman + 
    life, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3105.4  -867.1  -332.5   -11.1 24771.9 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  2.544e+05  4.298e+05   0.592   0.5548  
country     -1.548e+01  2.412e+01  -0.642   0.5219  
year        -1.073e+02  2.152e+02  -0.498   0.6190  
gdp          8.296e-11  1.375e-10   0.603   0.5471  
policy       6.325e+01  1.395e+02   0.453   0.6509  
woman       -7.495e+02  3.520e+02  -2.129   0.0349 *
life        -4.397e+01  3.831e+01  -1.148   0.2529  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3444 on 151 degrees of freedom
Multiple R-squared:  0.0474,	Adjusted R-squared:  0.009546 
F-statistic: 1.252 on 6 and 151 DF,  p-value: 0.283
```

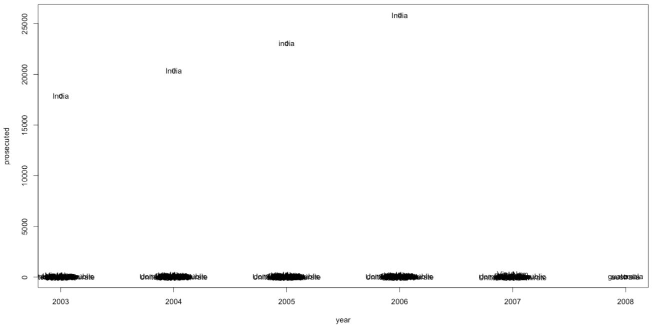
as we can see in the table, the percentage of female has a p-value smaller than 0.05 and policy has the greatest p-value. So “**the percentage of female**” has greatest explanatory power and “**policy index**” has the least explanatory power.

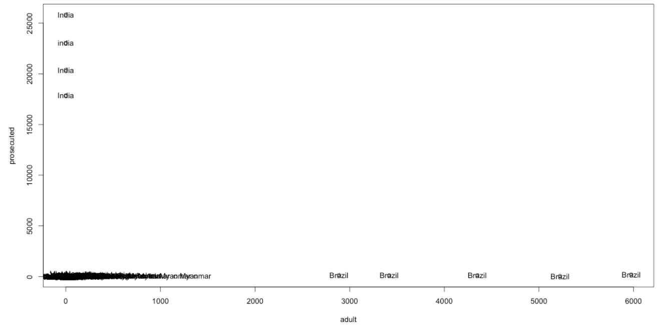
Please see attached R code.

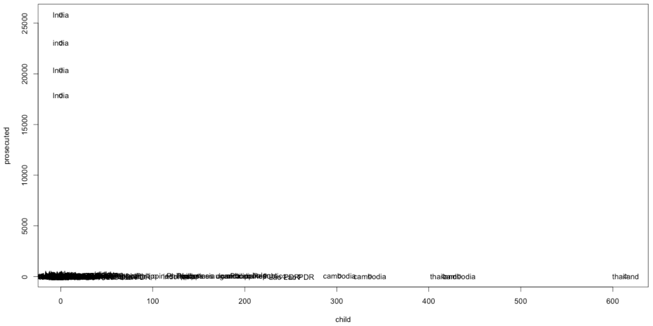
2) Remove some the outlier countries, how does this effect your model?

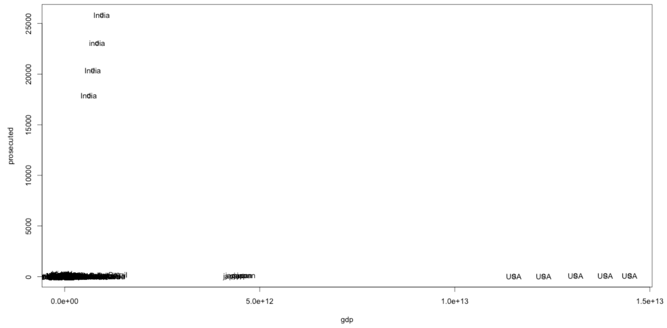
Plotting the variables like this:

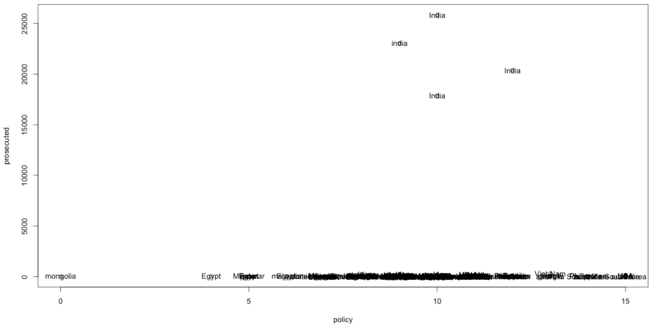


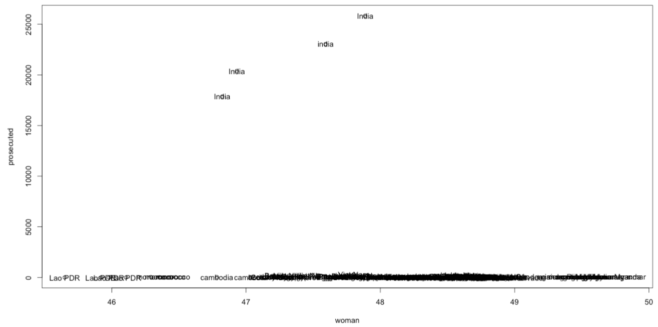


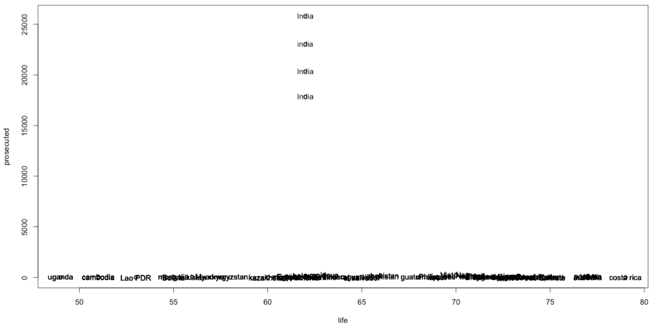












country: India

year: India

GDP: India, USA, Japan

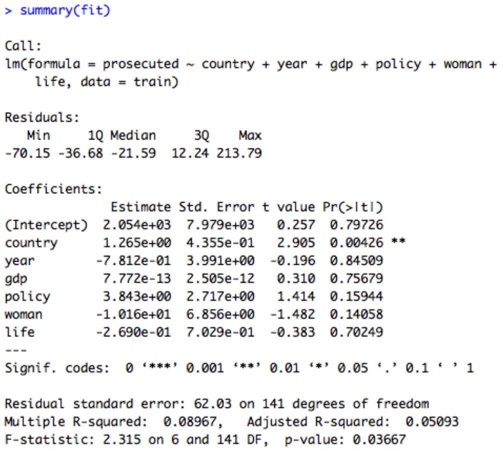
policy: Mongolia, India

Woman: India

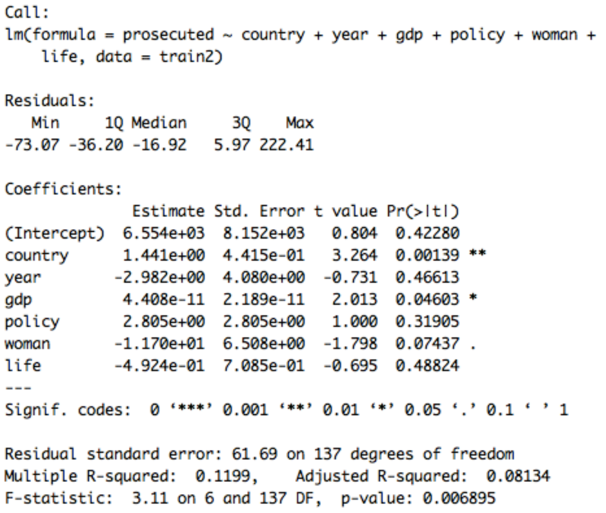
Life: India

so delete entire row of India but leave other rows alone. we can actually take a deeper look of the data, just use table(), summary(), str() functions to discover the outliers. Actually, we can use KNN or other ML algorithm to detect the outliers but too much for this task, I'll stick to the basic one.

so I also find that Mongolia 2003 is obvious wrong as well as Myanmar.



```
> summary(fit)

Call:
lm(formula = prosecuted ~ country + year + gdp + policy + woman + 
    life, data = train)

Residuals:
   Min     1Q Median     3Q    Max 
-70.15 -36.68 -21.59  12.24 213.79 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.054e+03  7.979e+03   0.257  0.79726   
country      1.265e+00  4.355e-01   2.905  0.00426 **
year        -7.812e-01  3.991e+00  -0.196  0.84509   
gdp          7.772e-13  2.505e-12   0.310  0.75679   
policy       3.843e+00  2.717e+00   1.414  0.15944   
woman       -1.016e+01  6.856e+00  -1.482  0.14058   
life        -2.690e-01  7.029e-01  -0.383  0.70249   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.03 on 141 degrees of freedom
Multiple R-squared:  0.08967,	Adjusted R-squared:  0.05093 
F-statistic: 2.315 on 6 and 141 DF,  p-value: 0.03667
```



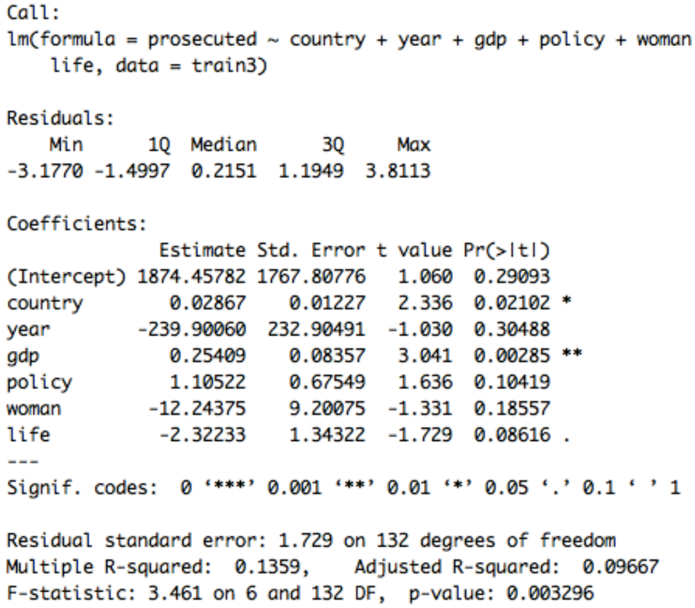
```
Call:
lm(formula = prosecuted ~ country + year + gdp + policy + woman + 
    life, data = train2)

Residuals:
   Min     1Q Median     3Q    Max 
-73.07 -36.20 -16.92   5.97 222.41 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  6.554e+03  8.152e+03   0.804  0.42280   
country      1.441e+00  4.415e-01   3.264  0.00139 **
year        -2.982e+00  4.080e+00  -0.731  0.46613   
gdp          4.408e-11  2.189e-11   2.013  0.04603 * 
policy       2.805e+00  2.805e+00   1.000  0.31905   
woman       -1.170e+01  6.508e+00  -1.798  0.07437 . 
life        -4.924e-01  7.085e-01  -0.695  0.48824   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.69 on 137 degrees of freedom
Multiple R-squared:  0.1199,	Adjusted R-squared:  0.08134 
F-statistic:  3.11 on 6 and 137 DF,  p-value: 0.006895
```

Then I keep doing that to remove USA and japan, then here's the outcome:

As we comparing the R-squared(adj) the model performance is getting better.

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

Calculate the log scale of the training data and train the model,



```
Call:
lm(formula = prosecuted ~ country + year + gdp + policy + woman +
    life, data = train3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1770 -1.4997  0.2151  1.1949  3.8113 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1874.45782 1767.80776   1.060  0.29093   
country        0.02867    0.01227   2.336  0.02102 * 
year        -239.90060  232.90491  -1.030  0.30488   
gdp            0.25409    0.08357   3.041  0.00285 **
policy         1.10522    0.67549   1.636  0.10419   
woman        -12.24375    9.20075  -1.331  0.18557   
life          -2.32233    1.34322  -1.729  0.08616 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.729 on 132 degrees of freedom
Multiple R-squared:  0.1359,    Adjusted R-squared:  0.09667 
F-statistic: 3.461 on 6 and 132 DF,  p-value: 0.003296
```

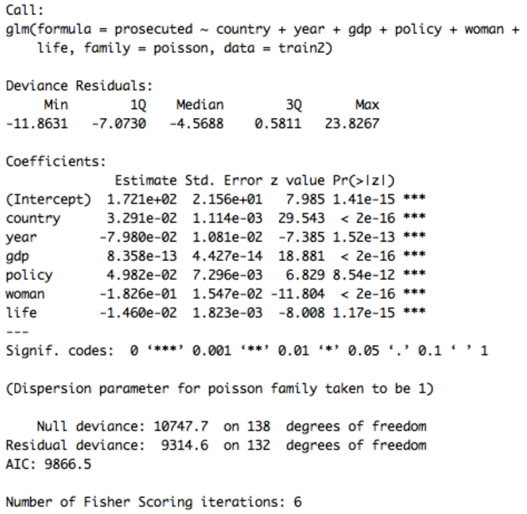
as we can see from the result, the GDP, which has large number when it is not in a log scale, now is more significant. Moreover, if we check the R-squared(adj), it’s 0.097 which is greater than earlier model.

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

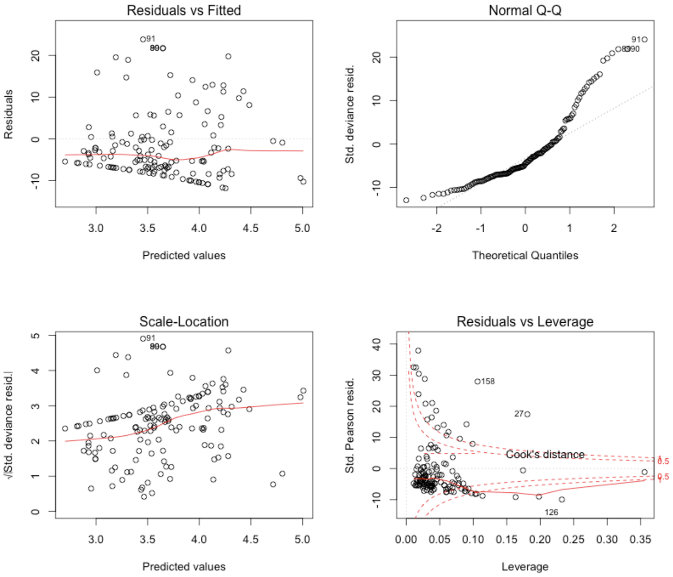
considering poisson distribution rather than normal distribution will be more reasonable for this data set.

comparing the difference between the poisson fit and linear fit:

the statistical number is great but the graph shows that the model definitely has some problems

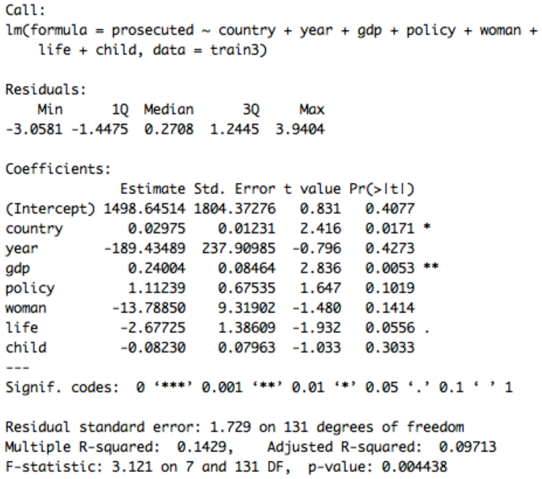


```
Call:
glm(formula = prosecuted ~ country + year + gdp + policy + woman +
    life, family = poisson, data = train2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-11.8631   -7.0730   -4.5688    0.5811   23.8827

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.721e+02  2.156e+01   7.985 1.41e-15 ***
country      3.291e-02  1.114e-03  29.543  < 2e-16 ***
year        -7.980e-02  1.081e-02  -7.385 1.52e-13 ***
gdp          8.358e-13  4.427e-14  18.881  < 2e-16 ***
policy       4.982e-02  7.296e-03   6.829 8.54e-12 ***
woman       -1.826e-01  1.547e-02 -11.804  < 2e-16 ***
life        -1.460e-02  1.823e-03  -8.008 1.17e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10747.7  on 138  degrees of freedom
Residual deviance:  9314.6  on 132  degrees of freedom
AIC: 9866.5

Number of Fisher Scoring iterations: 6
```



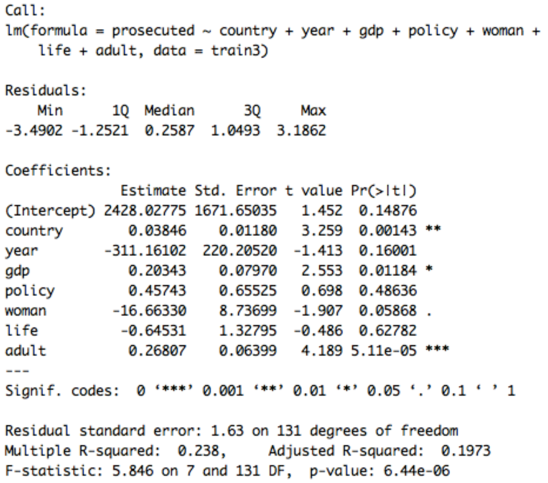
5) Think about how this model might be improved by adding more data. Then add this data to the model and test your hypothesis. What did you find. Provide descriptive statistics and visualizations as well as a few paragraphs explaining how you chose what data you did and why.

I don't have an exactly clue now, I think crime rate, population, etc, would be great for building the model for predicting the prosecuted people. But I would try to add the child victims data first since the child victims would be a great indicator for number of people prosecuted. It's a very serious crime for child trafficking and it's more rare and mostly would be independent incidents. Adult trafficking sometimes involve prostitution and others and make things complicated.



```
Call:
lm(formula = prosecuted ~ country + year + gdp + policy + woman +
    life + child, data = train3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0581 -1.4475  0.2708  1.2445  3.9404 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1498.64514 1804.37276   0.831   0.4077   
country        0.02975    0.01231   2.416   0.0171 * 
year        -189.43489  237.90985  -0.796   0.4273   
gdp            0.24004    0.08464   2.836   0.0053 **
policy         1.11239    0.67535   1.647   0.1019   
woman        -13.78850    9.31902  -1.480   0.1414   
life          -2.67725    1.38609  -1.932   0.0556 . 
child         -0.08230    0.07963  -1.033   0.3033   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.729 on 131 degrees of freedom
Multiple R-squared:  0.1429,	Adjusted R-squared:  0.09713 
F-statistic: 3.121 on 7 and 131 DF,  p-value: 0.004438
```

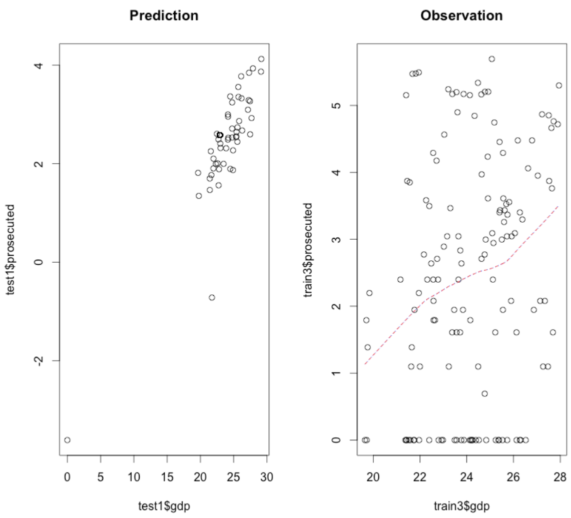
However, it's not improving the significance a lot. Then I tried the opposite theory using the adult victims, then:



```
Call:
lm(formula = prosecuted ~ country + year + gdp + policy + woman +
    life + adult, data = train3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4902 -1.2521  0.2587  1.0493  3.1862 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2428.02775 1671.65035   1.452  0.14876    
country        0.03846    0.01180   3.259  0.00143 ** 
year        -311.16102  220.20520  -1.413  0.16001    
gdp            0.20343    0.07970   2.553  0.01184 *  
policy         0.45743    0.65525   0.698  0.48636    
woman        -16.66330    8.73699  -1.907  0.05868 .  
life          -0.64531    1.32795  -0.486  0.62782    
adult          0.26807    0.06399   4.189 5.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.63 on 131 degrees of freedom
Multiple R-squared:  0.238,	Adjusted R-squared:  0.1973 
F-statistic: 5.846 on 7 and 131 DF,  p-value: 6.44e-06
```

It's doing great. I guess that adult victim are more likely to be find since people can speak and friends would call the police.Anyway it's a really good indicator for the number of prosecution.

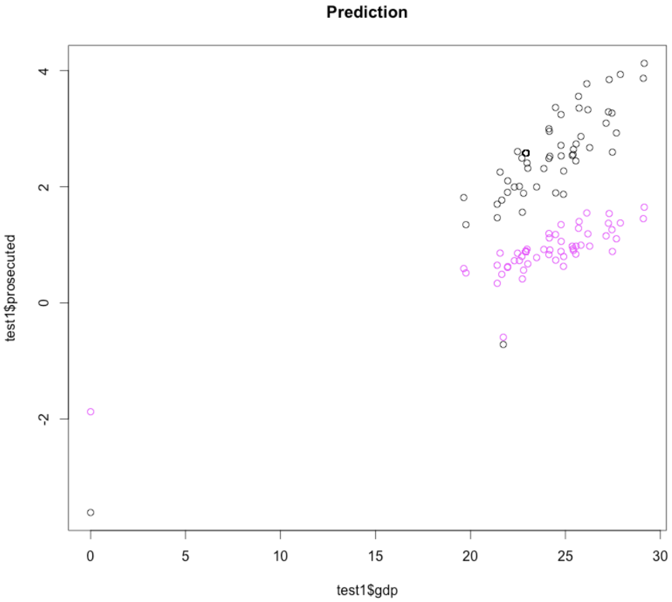
6) Using the model and data discussed in class predict how many cases a set of "new countries" would have (data to be provided in a separate csv file) Provide visualizations and a few paragraphs explaining your results.



using linear regression model on log scale, I made the prediction for the number of prosecuted people.

we can say the observation and prediction has the same positive effects, but Prediction seems to be more obvious.

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

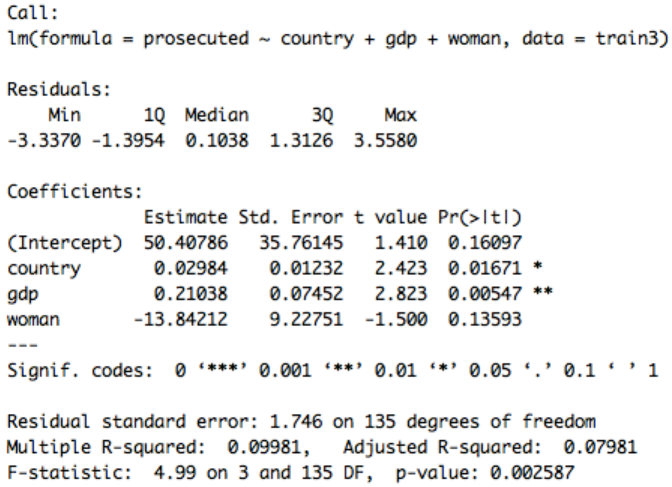


black is linear regression model and pink is Poisson regression model.

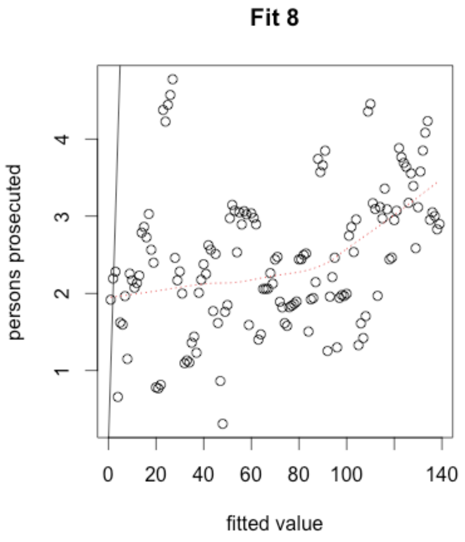
we can see the GLM model is lower, means smaller than the prediction of the LM model.

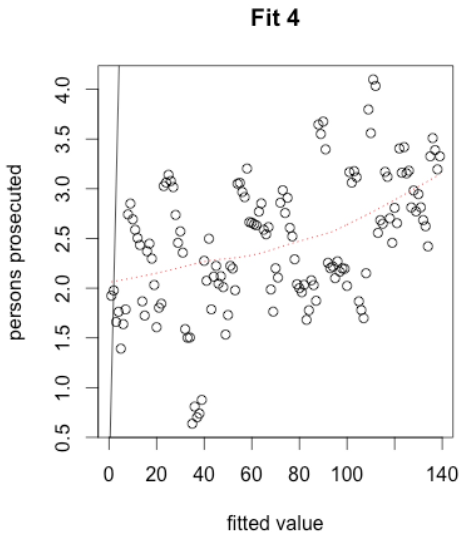
8) Now remove the variables with the least explanatory power. Does your linear regression improve compared to the other models? Does it do worse? Why? Please provide visuals and a few paragraphs of explanation

I delete the year and the life expectancy. Here's the statistical results:

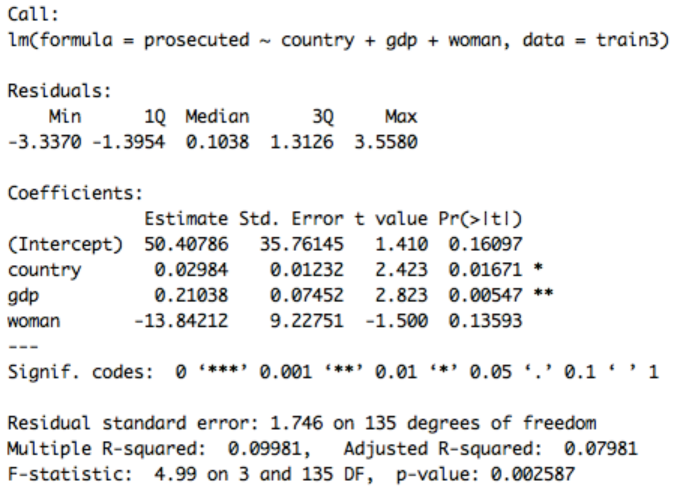


```
Call:
lm(formula = prosecuted ~ country + gdp + woman, data = train3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3370 -1.3954  0.1038  1.3126  3.5580 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  50.40786   35.76145   1.410  0.16097   
country       0.02984    0.01232   2.423  0.01671 * 
gdp           0.21038    0.07452   2.823  0.00547 **
woman       -13.84212    9.22751  -1.500  0.13593   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.746 on 135 degrees of freedom
Multiple R-squared:  0.09981,	Adjusted R-squared:  0.07981 
F-statistic:  4.99 on 3 and 135 DF,  p-value: 0.002587
```

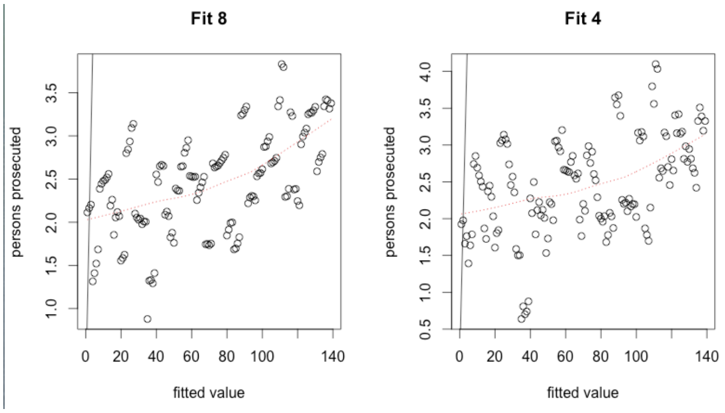




9) Now add in the extra data you found. Does your linear regression improved compared to the other models? Does it do worse? Why? Please provide visuals and a few paragraphs of explanation



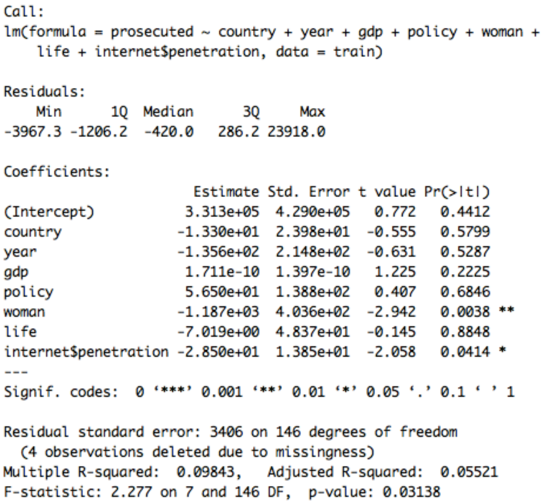
```
Call:
lm(formula = prosecuted ~ country + gdp + woman, data = train3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3370 -1.3954  0.1038  1.3126  3.5580 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  50.40786   35.76145   1.410  0.16097   
country       0.02984    0.01232   2.423  0.01671 * 
gdp           0.21038    0.07452   2.823  0.00547 **
woman       -13.84212    9.22751  -1.500  0.13593   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.746 on 135 degrees of freedom
Multiple R-squared:  0.09981,	Adjusted R-squared:  0.07981 
F-statistic:  4.99 on 3 and 135 DF,  p-value: 0.002587
```



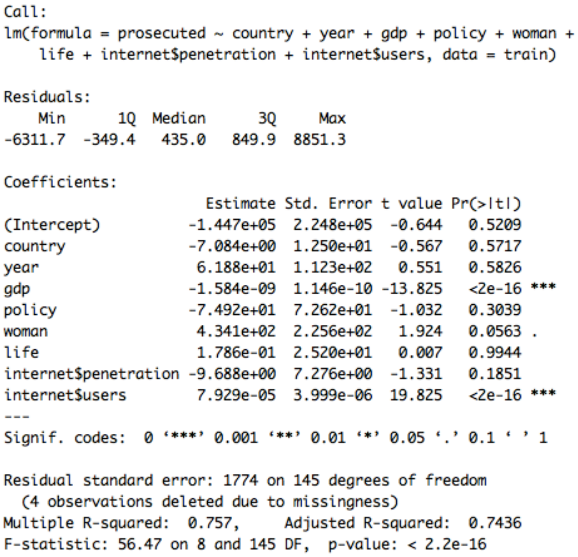
10) download (or scrape) data from the above websites.

see attached code.

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?



```
Call:
lm(formula = prosecuted ~ country + year + gdp + policy + woman + 
    life + internet$penetration, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3967.3 -1206.2  -420.0   286.2 23918.0 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)   
(Intercept)           3.313e+05  4.290e+05   0.772   0.4412   
country              -1.330e+01  2.398e+01  -0.555   0.5799   
year                 -1.356e+02  2.148e+02  -0.631   0.5287   
gdp                   1.711e-10  1.397e-10   1.225   0.2225   
policy                5.650e+01  1.388e+02   0.407   0.6846   
woman                -1.187e+03  4.036e+02  -2.942   0.0038 **
life                 -7.019e+00  4.837e+01  -0.145   0.8848   
internet$penetration -2.850e+01  1.385e+01  -2.058   0.0414 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3406 on 146 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.09843,	Adjusted R-squared:  0.05521 
F-statistic: 2.277 on 7 and 146 DF,  p-value: 0.03138
```



```
Call:
lm(formula = prosecuted ~ country + year + gdp + policy + woman + 
    life + internet$penetration + internet$users, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-6311.7  -349.4   435.0   849.9  8851.3 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -1.447e+05  2.248e+05  -0.644   0.5209    
country              -7.084e+00  1.250e+01  -0.567   0.5717    
year                  6.188e+01  1.123e+02   0.551   0.5826    
gdp                  -1.584e-09  1.146e-10 -13.825   <2e-16 ***
policy               -7.492e+01  7.262e+01  -1.032   0.3039    
woman                 4.341e+02  2.256e+02   1.924   0.0563 .  
life                  1.786e-01  2.520e+01   0.007   0.9944    
internet$penetration -9.688e+00  7.276e+00  -1.331   0.1851    
internet$users        7.929e-05  3.999e-06  19.825   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1774 on 145 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.757,	Adjusted R-squared:  0.7436 
F-statistic: 56.47 on 8 and 145 DF,  p-value: < 2.2e-16
```

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?

I say these variable add the explanatory power. The reason is obvious. the internet usage represent the economy and civilization level and the number of device is simply represent the population of the country. bigger population means more prosecution even with a lower crime rate.