IS 607 Final Project

Predicting internet penetration rates

Adam Stopek

Sunday, December 14, 2014

Introduction:

The internet is the greatest educational tool around today. The fact that I can study at a university that is physically located 6000 miles away from me is astounding. My belief is that internet penetration is strongly correlated to literacy levels. According to a press release by internet.org over 85% of the world is within range of picking up internet signals but only about 35% of the world actually use the internet. I would like to first quantify the correlation between literacy and internet penetration and then study some of the other factors that contribute to a higher internet penetration rate on a country level.

Feature Selection:

In order to find data I have scoured wikipedia and some other sites for country level data which I believe might be relevant towards affecting internet penetration. The features that I have come up with are:

- The literacy rate of a nation
- The GDP per capita (in USD)
- The median age
- The suicide rate
- The unemployment rate
- The population density
- The ratio of males to females
- The happy planet index, which is an index that says how happy the nation is
- The percent of English speakers
- The polution levels
- The life expectancy

In general these features are variables that I thought might have some effect on internet penetration rates. The reasons chosen may or may not be true. I will quickly try to explain some of the rationale for each of the features. Literacy rate is pretty obvious as internet is a text based world. GDP would be to understand if the costs are a barrier. The median age and life expectancy are because maybe internet is more prevelant amongst young people, or maybe old people, or possibly more developed countries have a higher life expectancy and also a higher adoption rate (merely because they re more developed). Suicide rate, unemployment rate (also related to cost barrier), happy planet index and polution rate all have to do with the populations desire to go on the internet. Ratio of males to females is based on an assumption that internet adoption is affected by gender. Population density was included under the assumption that a higher population density might mean better infrastructure. Finally percent English speakers is because most of the internet content is in English and unfortunately not the whole world speaks the language.

It is very possible that other reasons exist that could be predictive features, yet this is what I was able to think of to include.

Data Collection and Retrieval:

I started with scraping the data I found on the internet at the following URL's

```
penetration_url <- "http://data.worldbank.org/indicator/IT.NET.USER.P2"
literacy_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_literacy_rate"
gdp_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_past_and_projected_GDP_(PPP)"
median_age_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_median_age"
life_expectancy_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy"
suicide_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate"
unemployment_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_unemployment_rate"
pop_density_url <- "http://simple.wikipedia.org/wiki/List_of_countries_by_population_density"
male_female_ratio_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_sex_ratio"
hpi_url <- "http://en.wikipedia.org/wiki/Happy_Planet_Index"
c02_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_carbon_dioxide_emissions"
eng_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population"
populations_url <- "http://en.wikipedia.org/wiki/List_of_countries_by_population"
```

Data Cleansing:

Once reading the URL's into R, I had to massage the data to remove unused metrics from the tables as well as remove things like commas from the number fields and notes from the country field. A lot of regular expression work was used to clean up the data. Eventually from each URL my final data included only the country name and the specific metric as a numeric field. Some of the numeric fields were percentages and were displayed as a number between 0 and 100, I converted those to numbers between 0 and 1. Also in the case of GDP I only had the complete GDP and I had to devide it by population figures.

Once I had each of the tables seperately imported and cleaned, I merged the data together on the country field performing an "Inner Join". Unfortunately since my data did not include all the information for each country, my data set was widdled down to the lowest common denominator for countries included in each data set. If each of the data sets was more complete I would have ended up with 200+ observations for my final analysis, but alas it was not, and I was left with roughly 60 obs. The reason I performed an inner join was because I did not want to perform some sort of bootstrapping method for filling in null values as those methods may be statiscally questionable and might affect the final analysis. Here is the final data set after merging them together.

```
##                country penetration_rate literacy_rate gdp_per_capita
## 1            Argentina            0.599         0.979          21746
## 2            Australia            0.830         0.960          44413
## 3              Austria            0.806         0.980          44179
## 4             Barbados            0.750         0.997          15660
## 5              Belgium            0.822         0.990          40529
## 6               Belize            0.317         0.769           8015
## 7               Bhutan            0.299         0.528           7185
## 8               Brazil            0.516         0.913          14798
## 9             Bulgaria            0.531         0.984          16501
## 10              Canada            0.858         0.990          42561
## 12               Chile            0.665         0.986          22202
## 13            Colombia            0.517         0.936          12565
## 14             Croatia            0.667         0.992          20286
## 15              Cyprus            0.655         0.987          29534
## 16      Czech Republic            0.741         0.990          27333
## 17             Denmark            0.946         0.990          42590
## 18             Estonia            0.800         0.998          26138
## 19             Finland            0.915         1.000          39890
## 20              France            0.819         0.990          38356
## 21             Germany            0.840         0.990          43484
## 22              Greece            0.599         0.980          25286
## 23             Grenada            0.350         0.960          11749
## 24              Guyana            0.330         0.918           6668
## 25             Hungary            0.726         0.990          23236
## 26               India            0.151         0.744           5360
## 27             Ireland            0.782         0.990          46275
## 28              Israel            0.708         0.971          31144
## 29               Italy            0.585         0.990          33495
## 30             Jamaica            0.378         0.879           8693
## 31              Jordan            0.442         0.934          11423
## 32          Kazakhstan            0.540         0.995          22756
## 33              Latvia            0.752         0.998          23369
## 34           Lithuania            0.685         0.997          25760
## 35          Luxembourg            0.938         1.000          88252
## 36               Malta            0.689         0.928          30957
## 37           Mauritius            0.390         0.898          17677
## 38              Mexico            0.435         0.934          17199
## 39               Nepal            0.133         0.660           2257
## 40         Netherlands            0.940         0.990          46222
## 41         New Zealand            0.828         0.990          33131
## 42            Pakistan            0.109         0.690           4432
## 43         Philippines            0.370         0.954           6385
## 44              Poland            0.628         0.997          23295
## 45            Portugal            0.621         0.954          25652
## 46             Romania            0.498         0.977          18615
## 47           Singapore            0.730         0.959          77747
## 48            Slovenia            0.727         0.997          28427
## 49        South Africa            0.489         0.931          12270
## 50               Spain            0.716         0.977          32012
## 51           Sri Lanka            0.219         0.981           9841
## 52            Suriname            0.374         0.926          16475
## 53              Sweden            0.948         0.990          42949
## 54         Switzerland            0.867         0.990          52603
## 55            Thailand            0.289         0.935          14868
## 56 Trinidad and Tobago            0.638         0.986          30563
## 57              Turkey            0.463         0.953          18828
## 58      United Kingdom            0.898         0.990          36197
## 59       United States            0.842         0.990          52519
## 60            Zimbabwe            0.185         0.907           1963
##    median_age suicide_rate unemployment_rate population_per_sq_km
## 1        30.3       0.0770             0.075                 14.0
## 2        37.5       0.1000             0.064                  3.2
## 3        42.6       0.1545             0.048                100.0
## 4        36.2       0.0350             0.115                595.0
## 5        42.0       0.1700             0.085                355.0
## 6        20.7       0.0370             0.113                 14.0
## 7        24.3       0.1620             0.040                 46.0
## 8        30.5       0.0480             0.049                 24.0
## 9        41.6       0.1230             0.131                 66.0
## 10       40.7       0.1150             0.065                  3.4
## 12       31.7       0.1120             0.061                 24.0
## 13       27.6       0.0490             0.078                 42.0
## 14       41.2       0.1970             0.176                 79.0
## 15       34.5       0.0360             0.153                 87.0
## 16       40.4       0.1280             0.067                134.0
## 17       40.7       0.1130             0.070                128.0
## 18       40.2       0.1480             0.087                 29.0
## 19       41.6       0.1600             0.082                 16.0
## 20       39.7       0.1470             0.104                114.0
## 21       43.7       0.1220             0.051                229.0
## 22       42.2       0.0350             0.259                 86.0
## 23       28.2       0.0000             0.245                302.0
## 24       23.6       0.2640             0.090                  3.5
## 25       41.3       0.2110             0.071                108.0
## 26       25.9       0.1050             0.088                368.0
## 27       35.4       0.1030             0.110                 65.0
## 28       29.3       0.0580             0.059                371.0
## 29       44.3       0.0630             0.126                200.0
## 30       23.9       0.0010             0.113                247.0
## 31       21.8       0.0010             0.119                 71.0
## 32       29.9       0.2560             0.061                  6.2
## 33       40.4       0.2080             0.116                 35.0
## 34       39.7       0.3100             0.115                 47.0
## 35       39.3       0.0780             0.061                194.0
## 36       39.7       0.0340             0.069               1322.0
## 37       32.3       0.0680             0.079                631.0
## 38       26.7       0.0400             0.049                 57.0
## 39       21.2       0.0000             0.460                199.0
## 40       40.8       0.0880             0.073                409.0
## 41       36.8       0.1150             0.056                 16.0
## 42       21.2       0.0110             0.066                234.0
## 43       22.7       0.0275             0.070                307.0
## 44       38.2       0.1750             0.097                122.0
## 45       39.7       0.0960             0.153                115.0
## 46       38.1       0.1190             0.072                 90.0
## 47       39.6       0.1030             0.019               7148.0
## 48       42.1       0.2180             0.098                106.0
## 49       24.7       0.1540             0.255                 41.0
## 50       41.5       0.0760             0.256                 91.0
## 51       31.3       0.2130             0.042                308.0
## 52       28.3       0.1440             0.090                  3.2
## 53       41.7       0.1200             0.081                 21.0
## 54       41.3       0.1120             0.031                188.0
## 55       33.7       0.0610             0.009                125.0
## 56       32.6       0.1070             0.037                261.0
## 57       28.1       0.0419             0.088                 93.0
## 58       40.5       0.1180             0.060                255.0
## 59       36.9       0.1250             0.058                 32.0
## 60       17.8       0.0790             0.700                 33.0
##    male_female_ratio   hpi percent_english_speakers co2_per_capita
## 1               0.97 51.96                   0.0652          4.471
## 2               1.00 34.06                   0.9703         16.934
## 3               0.95 48.77                   0.7300          7.974
## 4               0.94 52.73                   0.9857          5.362
## 5               0.96 44.04                   0.5900          9.977
## 6               1.03 51.32                   0.8165          1.367
## 7               1.10 61.08                   0.1140          0.665
## 8               0.98 48.59                   0.0790          2.150
## 9               0.92 31.59                   0.2500          6.041
## 10              0.98 39.76                   0.8563         14.678
## 12              0.98 52.20                   0.0953          4.213
## 13              0.98 67.24                   0.0422          1.629
## 14              0.93 43.71                   0.4900          4.727
## 15              1.04 45.99                   0.7300          6.984
## 16              0.95 36.50                   0.2700         10.669
## 17              0.98 41.40                   0.8600          8.346
## 18              0.84 22.68                   0.5000         13.773
## 19              0.96 37.36                   0.7000         11.531
## 20              0.96 36.42                   0.3900          5.556
## 21              0.97 43.83                   0.6400          9.115
## 22              0.96 35.71                   0.5100          7.775
## 23              1.02 48.96                   0.9091          2.487
## 24              1.00 56.65                   0.9055          2.164
## 25              0.91 37.64                   0.2000          5.058
## 26              1.08 42.46                   0.1035          1.666
## 27              0.99 39.38                   0.9837          8.772
## 28              1.00 39.07                   0.8497          9.268
## 29              0.96 48.26                   0.3400          6.854
## 30              0.98 51.01                   0.9764          2.660
## 31              1.10 42.05                   0.4500          3.444
## 32              0.93 36.92                   0.1540         15.239
## 33              0.86 27.27                   0.4600          3.631
## 34              0.89 29.29                   0.3800          4.378
## 35              0.97 45.62                   0.5600         21.360
## 36              0.99 53.26                   0.8900          6.246
## 37              0.97 49.65                   0.1597          3.215
## 38              0.96 54.39                   0.1290          3.764
## 39              0.96 49.95                   0.4649          0.140
## 40              0.98 46.00                   0.9000         10.958
## 41              0.99 41.92                   0.9782          7.224
## 42              1.09 39.40                   0.4900          0.932
## 43              1.00 59.17                   0.5663          0.873
## 44              0.94 39.29                   0.3300          8.309
## 45              0.95 34.83                   0.2700          4.952
## 46              0.95 37.72                   0.3100          3.889
## 47              0.95 36.14                   0.8000          2.663
## 48              0.95 44.03                   0.5900          7.482
## 49              0.99 27.80                   0.3100          9.041
## 50              0.96 43.04                   0.2200          5.790
## 51              0.97 60.31                   0.0990          0.615
## 52              0.99 55.03                   0.8709          4.540
## 53              0.98 38.17                   0.8600          5.600
## 54              0.97 48.30                   0.6128          4.953
## 55              0.98 55.39                   0.2716          4.447
## 56              1.02 51.87                   0.8774         38.161
## 57              1.02 41.40                   0.1700          4.131
## 58              0.98 40.29                   0.9774          7.863
## 59              0.97 28.83                   0.9420         17.564
## 60              0.91 16.64                   0.4158          0.721
##    life_expectancy
## 1             75.3
## 2             81.4
## 3             80.2
## 4             76.2
## 5             79.7
## 6             75.3
## 7             65.7
## 8             75.5
## 9             72.7
## 10            80.5
## 12            78.6
## 13            72.9
## 14            76.0
## 15            78.9
## 16            77.0
## 17            78.2
## 18            73.9
## 19            79.3
## 20            81.0
## 21            79.8
## 22            79.5
## 23            75.3
## 24            68.7
## 25            73.6
## 26            64.1
## 27            79.6
## 28            80.6
## 29            81.3
## 30            72.2
## 31            72.9
## 32            65.7
## 33            72.2
## 34            71.3
## 35            79.3
## 36            78.8
## 37            72.8
## 38            76.1
## 39            67.4
## 40            80.2
## 41            80.1
## 42            64.5
## 43            67.8
## 44            75.5
## 45            78.5
## 46            73.1
## 47            80.6
## 48            78.5
## 49            51.2
## 50            80.7
## 51            74.2
## 52            69.6
## 53            80.8
## 54            81.8
## 55            73.5
## 56            69.4
## 57            72.9
## 58            79.5
## 59            77.9
## 60            46.5
```

Data Exploration:

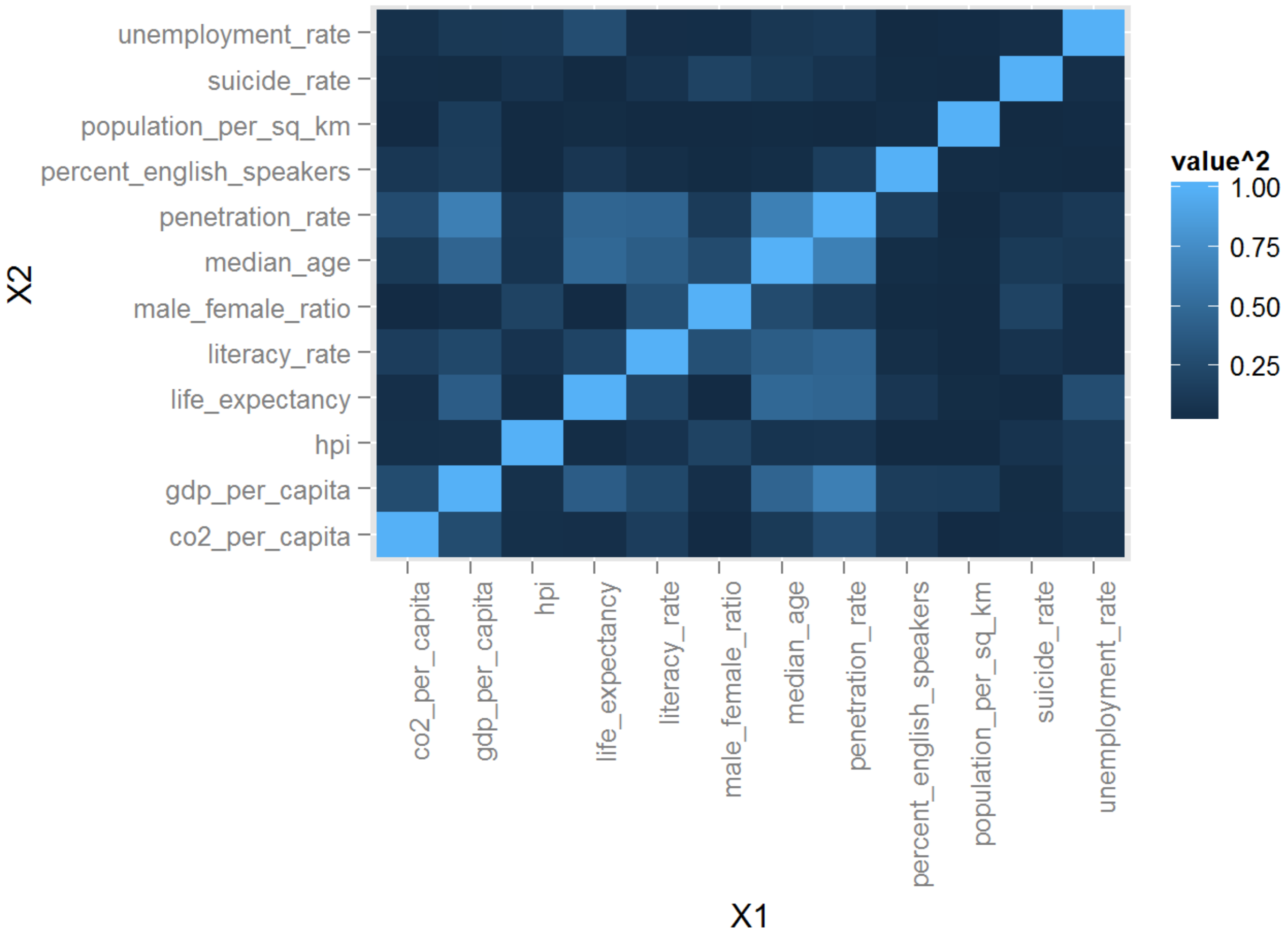
Lets begin by taking a quick look at the summary statistics of the data.

```
##    country          penetration_rate literacy_rate   gdp_per_capita
##  Length:59          Min.   :0.109    Min.   :0.528   Min.   : 1963
##  Class :character   1st Qu.:0.438    1st Qu.:0.934   1st Qu.:14833
##  Mode  :character   Median :0.655    Median :0.981   Median :23369
##                     Mean   :0.608    Mean   :0.945   Mean   :26687
##                     3rd Qu.:0.803    3rd Qu.:0.990   3rd Qu.:37277
##                     Max.   :0.948    Max.   :1.000   Max.   :88252
##    median_age    suicide_rate    unemployment_rate population_per_sq_km
##  Min.   :17.8   Min.   :0.0000   Min.   :0.009     Min.   :   3
##  1st Qu.:28.2   1st Qu.:0.0535   1st Qu.:0.061     1st Qu.:  34
##  Median :36.9   Median :0.1070   Median :0.079     Median :  93
##  Mean   :34.4   Mean   :0.1080   Mean   :0.109     Mean   : 278
##  3rd Qu.:40.7   3rd Qu.:0.1475   3rd Qu.:0.114     3rd Qu.: 232
##  Max.   :44.3   Max.   :0.3100   Max.   :0.700     Max.   :7148
##  male_female_ratio      hpi       percent_english_speakers co2_per_capita
##  Min.   :0.840     Min.   :16.6   Min.   :0.0422           Min.   : 0.14
##  1st Qu.:0.950     1st Qu.:37.5   1st Qu.:0.2700           1st Qu.: 3.33
##  Median :0.970     Median :43.0   Median :0.5000           Median : 5.36
##  Mean   :0.974     Mean   :43.5   Mean   :0.5332           Mean   : 6.86
##  3rd Qu.:0.990     3rd Qu.:50.5   3rd Qu.:0.8582           3rd Qu.: 8.56
##  Max.   :1.100     Max.   :67.2   Max.   :0.9857           Max.   :38.16
##  life_expectancy
##  Min.   :46.5
##  1st Qu.:72.8
##  Median :76.0
##  Mean   :74.7
##  3rd Qu.:79.5
##  Max.   :81.8
```

My original assumption was that literacy levels were highly correlated to internet penetration. In order to check if that is true we can calculate the Pearson Coefficient between those two variables.

```
## [1] "The Correlation Coefficient between Literacy rate and Internet Penetration is : 0.67"
```

Now that we know that there is a strong correlation, it would be interesting to look at the cross correlations between all the variable. Below is a heatmap of the R-squared for each pair of variables. The lighter the color the more correlated the variables are. Notice that the diaganol is completely light blue as each variable’s correlation with itself is 1.



Predictive Model Building

In order to build a predictive model, I will devide my dataset into a training set and a test set. I will use 70% of the observations to train on and the other 30% to validate the model. R-squared is a good indication, but physically plotting the actual versus the expected is something I always find to be great way to visualize the results.

Once I seperate out the test set from the training set I will run a simple linear regression as a naive model just to see how well each of the variables are at helping to predict the penetration rate in each country. As you can see from the regression result, some of the variables below are not significant.

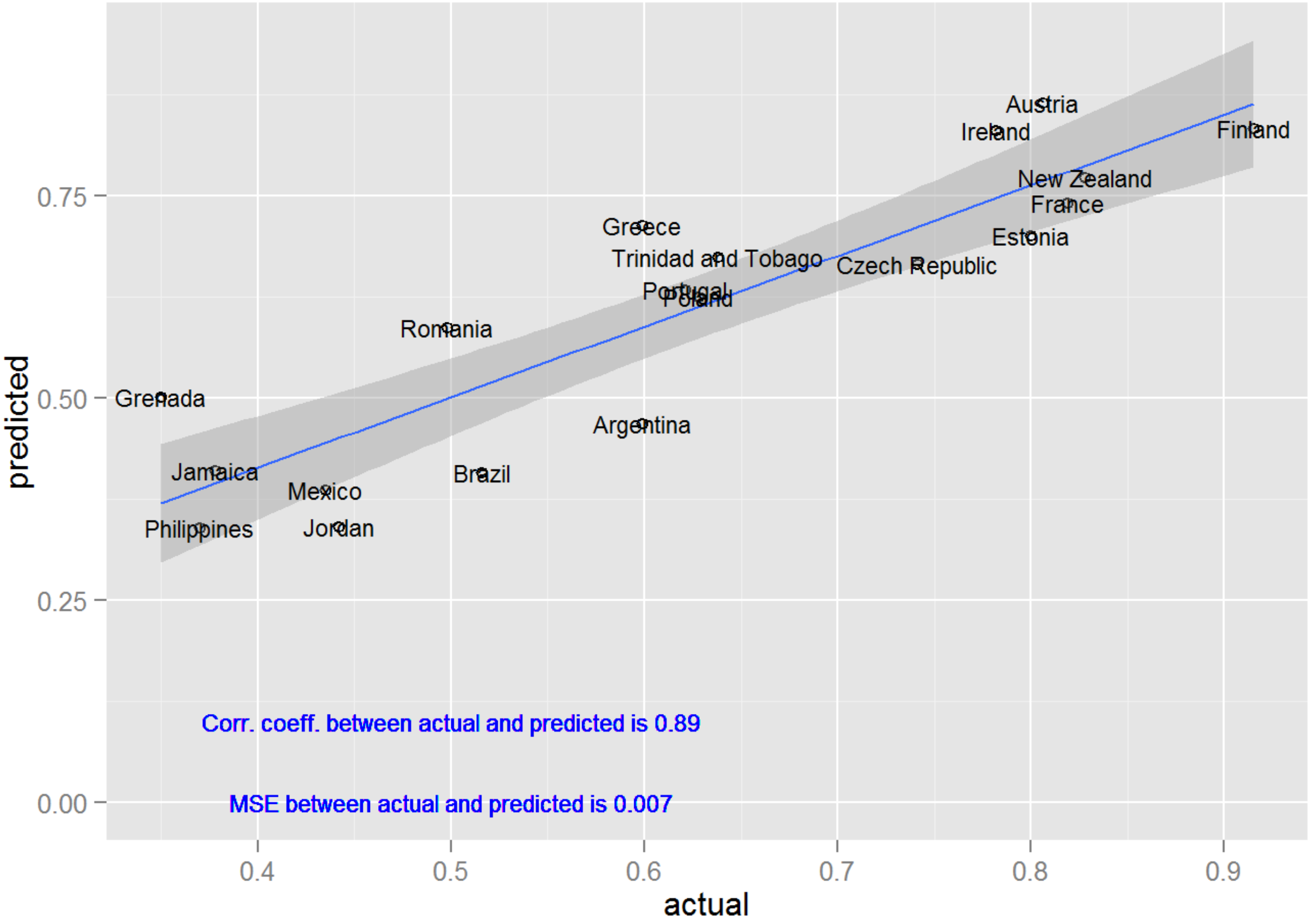
```
##
## Call:
## lm(formula = penetration_rate ~ literacy_rate + gdp_per_capita +
##     median_age + suicide_rate + unemployment_rate + population_per_sq_km +
##     male_female_ratio + hpi + percent_english_speakers + co2_per_capita +
##     life_expectancy, data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18458 -0.03599  0.00514  0.05499  0.15314
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.26e-01   6.79e-01   -0.48   0.6348
## literacy_rate            1.93e-01   2.23e-01    0.87   0.3933
## gdp_per_capita           5.03e-06   1.63e-06    3.08   0.0044 **
## median_age               9.56e-03   5.74e-03    1.66   0.1064
## suicide_rate             3.37e-02   3.09e-01    0.11   0.9137
## unemployment_rate       -1.37e-01   2.06e-01   -0.66   0.5116
## population_per_sq_km    -4.18e-05   8.52e-05   -0.49   0.6270
## male_female_ratio       -1.53e-01   5.08e-01   -0.30   0.7658
## hpi                     -3.95e-03   2.51e-03   -1.58   0.1256
## percent_english_speakers  1.16e-01   5.68e-02    2.05   0.0491 *
## co2_per_capita           1.28e-03   3.20e-03    0.40   0.6923
## life_expectancy          7.58e-03   5.98e-03    1.27   0.2145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0951 on 30 degrees of freedom
## Multiple R-squared:  0.891,  Adjusted R-squared:  0.851
## F-statistic: 22.2 on 11 and 30 DF,  p-value: 1.83e-11
```

Becuase of this, I will now run a stepwise regression (bi-directional) in order to remove some of the variables which dont provide enough predictive power to the model. Here are the final results of the stepwise model.

```
## Start:  AIC=-187.8
## penetration_rate ~ literacy_rate + gdp_per_capita + median_age +
##     suicide_rate + unemployment_rate + population_per_sq_km +
##     male_female_ratio + hpi + percent_english_speakers + co2_per_capita +
##     life_expectancy
##
##                            Df Sum of Sq   RSS  AIC
## - suicide_rate              1    0.0001 0.271 -190
## - male_female_ratio         1    0.0008 0.272 -190
## - co2_per_capita            1    0.0014 0.273 -190
## - population_per_sq_km      1    0.0022 0.273 -190
## - unemployment_rate         1    0.0040 0.275 -189
## - literacy_rate             1    0.0068 0.278 -189
## <none>                                  0.271 -188
## - life_expectancy           1    0.0145 0.286 -188
## - hpi                       1    0.0224 0.294 -186
## - median_age                1    0.0250 0.296 -186
## - percent_english_speakers  1    0.0380 0.309 -184
## - gdp_per_capita            1    0.0858 0.357 -178
## 
## Step:  AIC=-189.8
## penetration_rate ~ literacy_rate + gdp_per_capita + median_age + 
##     unemployment_rate + population_per_sq_km + male_female_ratio + 
##     hpi + percent_english_speakers + co2_per_capita + life_expectancy
## 
##                            Df Sum of Sq   RSS  AIC
## - male_female_ratio         1    0.0010 0.272 -192
## - co2_per_capita            1    0.0013 0.273 -192
## - population_per_sq_km      1    0.0036 0.275 -191
## - unemployment_rate         1    0.0053 0.277 -191
## - literacy_rate             1    0.0068 0.278 -191
## <none>                                  0.271 -190
## - life_expectancy           1    0.0186 0.290 -189
## - hpi                       1    0.0232 0.294 -188
## + suicide_rate              1    0.0001 0.271 -188
## - median_age                1    0.0343 0.305 -187
## - percent_english_speakers  1    0.0401 0.311 -186
## - gdp_per_capita            1    0.0860 0.357 -180
## 
## Step:  AIC=-191.6
## penetration_rate ~ literacy_rate + gdp_per_capita + median_age + 
##     unemployment_rate + population_per_sq_km + hpi + percent_english_speakers + 
##     co2_per_capita + life_expectancy
## 
##                            Df Sum of Sq   RSS  AIC
## - co2_per_capita            1    0.0012 0.273 -194
## - population_per_sq_km      1    0.0040 0.276 -193
## - unemployment_rate         1    0.0045 0.277 -193
## - literacy_rate             1    0.0116 0.284 -192
## <none>                                  0.272 -192
## - life_expectancy           1    0.0177 0.290 -191
## - hpi                       1    0.0238 0.296 -190
## + male_female_ratio         1    0.0010 0.271 -190
## + suicide_rate              1    0.0003 0.272 -190
## - percent_english_speakers  1    0.0393 0.311 -188
## - median_age                1    0.0578 0.330 -186
## - gdp_per_capita            1    0.0855 0.358 -182
## 
## Step:  AIC=-193.5
## penetration_rate ~ literacy_rate + gdp_per_capita + median_age +
##     unemployment_rate + population_per_sq_km + hpi + percent_english_speakers +
##     life_expectancy
##
##                            Df Sum of Sq   RSS  AIC
## - population_per_sq_km      1    0.0043 0.278 -195
## - unemployment_rate         1    0.0063 0.280 -194
## <none>                                  0.273 -194
## - literacy_rate             1    0.0135 0.287 -193
## - life_expectancy           1    0.0171 0.290 -193
## - hpi                       1    0.0231 0.296 -192
## + co2_per_capita            1    0.0012 0.272 -192
## + male_female_ratio         1    0.0008 0.273 -192
## + suicide_rate              1    0.0001 0.273 -192
## - percent_english_speakers  1    0.0454 0.319 -189
## - median_age                1    0.0599 0.333 -187
## - gdp_per_capita            1    0.1303 0.404 -179
##
## Step:  AIC=-194.8
## penetration_rate ~ literacy_rate + gdp_per_capita + median_age +
##     unemployment_rate + hpi + percent_english_speakers + life_expectancy
##
##                            Df Sum of Sq   RSS  AIC
## - unemployment_rate         1    0.0085 0.286 -196
## <none>                                  0.278 -195
## - literacy_rate             1    0.0144 0.292 -195
## - life_expectancy           1    0.0204 0.298 -194
## + population_per_sq_km      1    0.0043 0.273 -194
## + suicide_rate              1    0.0017 0.276 -193
## + co2_per_capita            1    0.0015 0.276 -193
## + male_female_ratio         1    0.0012 0.276 -193
## - hpi                       1    0.0394 0.317 -191
## - percent_english_speakers  1    0.0412 0.319 -191
## - median_age                1    0.0562 0.334 -189
## - gdp_per_capita            1    0.1360 0.414 -180
##
## Step:  AIC=-195.5
## penetration_rate ~ literacy_rate + gdp_per_capita + median_age +
##     hpi + percent_english_speakers + life_expectancy
##
##                            Df Sum of Sq   RSS  AIC
## <none>                                  0.286 -196
## - literacy_rate             1    0.0158 0.302 -195
## + unemployment_rate         1    0.0085 0.278 -195
## + population_per_sq_km      1    0.0065 0.280 -194
## + suicide_rate              1    0.0049 0.281 -194
## + co2_per_capita            1    0.0040 0.282 -194
## - life_expectancy           1    0.0282 0.314 -194
## + male_female_ratio         1    0.0000 0.286 -194
## - hpi                       1    0.0310 0.317 -193
## - percent_english_speakers  1    0.0365 0.323 -192
## - median_age                1    0.0656 0.352 -189
## - gdp_per_capita            1    0.1378 0.424 -181
```

```
##
## Call:
## lm(formula = penetration_rate ~ literacy_rate + gdp_per_capita +
##     median_age + hpi + percent_english_speakers + life_expectancy,
##     data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17757 -0.05220 -0.00476  0.05962  0.17201
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -5.78e-01   2.34e-01   -2.47  0.01869 *
## literacy_rate             2.58e-01   1.86e-01    1.39  0.17349
## gdp_per_capita            5.36e-06   1.31e-06    4.11  0.00023 ***
## median_age                1.05e-02   3.71e-03    2.83  0.00761 **
## hpi                      -3.66e-03   1.88e-03   -1.95  0.05962 .
## percent_english_speakers  1.01e-01   4.80e-02    2.11  0.04169 *
## life_expectancy           7.41e-03   3.99e-03    1.86  0.07149 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0904 on 35 degrees of freedom
## Multiple R-squared:  0.885,  Adjusted R-squared:  0.865
## F-statistic: 44.8 on 6 and 35 DF,  p-value: 5.48e-15
```

Using the final model, I will now apply the model estimates to the test set in order to see how accurate I predicted Internet Penetration. Here I have plotted the actual internet penetration rates versus the predicted rates by the model. As you can see, the model predicts internet pentration extremely well and most observations fall within the confidence interval.



Conclusion

We now see that we can use predictive features to understand internet penetration rates in various countries. Now it is still unsure as to whether the features are a cause to internet penetration or vice versa, but indeed a relationship exists. And understanding the relationship is the first step to making the world more open and connected.