An Analysis of the Factors that determine Immigration and Population Growth in the Developed Nations

x18179797@student.ncirl.ie

Abstract

We perform linear and logistic multiple regression on United Nations data to identify the factors influencing immigration and population growth in the developed nations. In both instances, we discover that the Gross Domestic Product (GDP) per capita is the dominant factor at play.

I. Introduction

There have been many studies that have identified the ageing demographics in the developed world [1-3]. In these developed nations, there has been a dramatic drop in fertility rates which has been attributed to a variety of factors ranging from the mass introduction of women into the labour markets, the habitual use of contraception and cost of child rearing [3]. This places an increasing burden on the work force to support an ever increasing proportion of senior citizens and there have been numerous articles addressing this pension crisis [4-5]. One solution is to allow the free movement of migrants into a nation to support public services and economic activity [6].

Here we perform our own independent study into the factors that determine the population dynamics of a developed nation. In section II, we discuss the data set and tools used in this work. In section III of this paper, we examine immigration in the developed nations. Why do some countries have a much larger proportion of migrant population than others? We address this question by building a linear regression model on our available data.

In the section IV, we look at the population growth rates in the developed nations. Specifically we split these countries into two groups, those with positive growth rates, and those with negative. We then construct a logistic regression model that gives the probability of a nation having a positive population growth rate, and examine the significant factors in the model. Finally in section V, we end the paper with a brief discussion and conclusion on the work performed.

II. Data Set and Analysis Tools

The raw data sets were sourced from the publicly available United Nations data repositories at http://data.un.org. Data covering population, economic and environment statistics were downloaded. There were 10 data sets in total used in this study (Links to and details of each data set are provided in Appendix i). They each have a similar structure. That is, each record/observation corresponds to a country or region that is uniquely identified by an M49 code [7].

As an illustrative example, here is a sample of a subset of columns from the "UNSD – Methodology.csv" data set:

M49 Code	Country or Area	Region Name	Developed/Developing Countries
818	Egypt	Africa	Developing
756	Switzerland	Europe	Developed
392	Japan	Asia	Developed

Note also that the downloaded files have a *Year* column. That is, data is provided for a subset of years e.g. Employment data for 2010, 2015 and 2018.

All work was performed in Python on a Jupyter Notebook (that is available on request). Each data set was read in and the most recent annual data extracted. The annual records for each data set were then joined on the *M49 code* column to produce a single master data set with each record thus having population, economic and environmental statistics for a specific country or geographical region. It should be noted however that as the year extracted was not identical across all data sets (ranged from 2014-2018), each record is therefore not temporally consistent e.g. We may have 2016 economic data mixed with 2018 population data. (For the details of this inconsistency, see the table in appendix i).

Finally, filtering on the "Developed/Developing Countries" column (See sample table above), we selected only developed countries so that our master data set was reduced down to 46 records/nations. For the remainder of this paper, unless otherwise explicitly stated, when we refer to our *data* or *data set*, it is this data set that we are referencing.

As previously mentioned, all our subsequent analysis was done in Python on a Jupyter notebook with both the linear and logistic regressions performed using the *statsmodels* libraries. It should also be noted that in all our regression work in the proceeding sections that our significance level is 0.05.

III. Immigration in the Developed Nations – A Multiple Linear Regression model

In this section we build a multiple linear regression model in order to infer the dominant factors that determine the proportion of migrants in a nations population. The essential question we are addressing here is: Why does one country contain a large proportion of migrant population while another has a much smaller proportion?

As our dependent variable we shall choose the field "*International migrant stock: Both sexes (% total population)*" which I shall, going forward, refer to by the alias *%Migrants* (Note that a full list of feature names together with their aliases are given in appendix ii). The model was constructed in a systematic fashion by writing bespoke code that essentially implements a version of forward selection. Here we build our model up one feature at a time.

This was done by cycling through all the available features and observing their returned adjusted r-squared and p-values. The appropriate feature was then manually selected and added to the model before proceeding with another automated scan over the remaining features.

There are three important points of note regarding this semi-automated procedure. First, regarding the manual selection of the appropriate feature, preference was given to those features having large adjusted r-squared and low p-value. But another important criterion was that it was also not strongly correlated (|correlation|<0.5) to the current features in the model.

Secondly, the list of candidate features did not cover all of the available features in our data set as some had missing values and others in my opinion had some redundancy. The final set of 35 candidate features are listed in appendix ii.

Lastly, all the values of these variables were standardised (to z-scores) before commencing the regression to avoid numerical complications and to more easily identify the dominant features/factors in our model.

MODEL RESULTS: It is informative to show the results of the forward selection at the first stage i.e. Selecting the first model feature. The automated scan returned the following results that are ordered by adjusted r-squared:

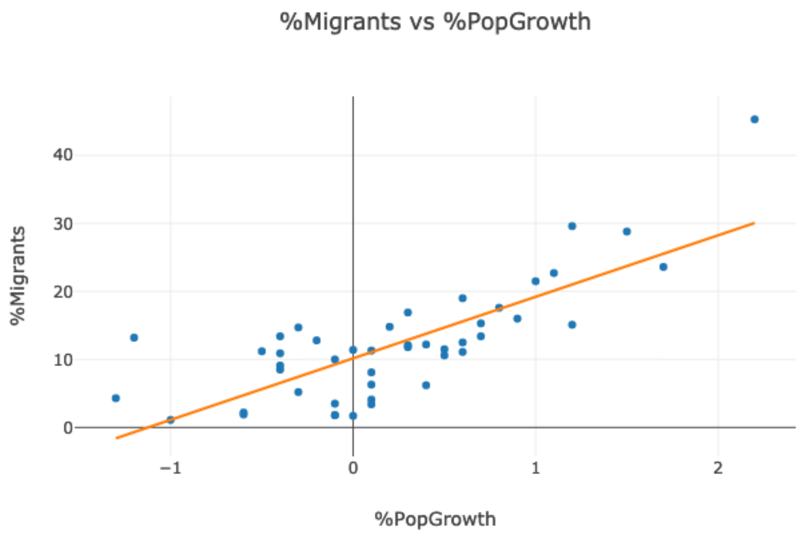
Adjusted r-squared	p-value	Feature Name
0.580	4.70e-10	%PopGrowth
0.545	2.93e-09	GDPperCapita
0.434	3.85e-07	%UrbanPopGrowth

Clearly, the features %PopGrowth and GDPperCapita are the leading candidates for selection, having significantly better stats (i.e. adjusted r-squared and p-value) than all the remaining candidates (Note that only the top three candidates are shown above).

I show the summary stats (and plots for illustration) for both simple linear regression models below:

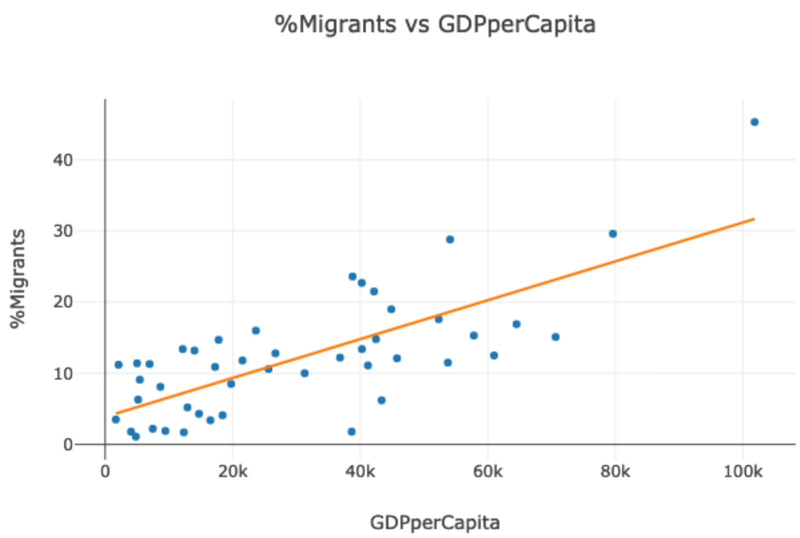
Model A: *%Migrants ~ %PopGrowth*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              %Migrants   R-squared:                       0.590
Model:                            OLS   Adj. R-squared:                  0.580
Method:                 Least Squares   F-statistic:                     63.26
Date:                Sun, 07 Apr 2019   Prob (F-statistic):           4.70e-10
Time:                        17:45:24   Log-Likelihood:                -142.76
No. Observations:                  46   AIC:                             289.5
Df Residuals:                      44   BIC:                             293.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         12.1630      0.813     14.968      0.000      10.525      13.801
%PopGrowth     6.4633      0.813      7.954      0.000       4.826       8.101
==============================================================================
Omnibus:                        4.836   Durbin-Watson:                   1.591
Prob(Omnibus):                  0.089   Jarque-Bera (JB):                3.815
Skew:                           0.683   Prob(JB):                        0.148
Kurtosis:                       3.357   Cond. No.                         1.00
==============================================================================
```



Model B: *%Migrants ~ GDPperCapita*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              %Migrants   R-squared:                       0.555
Model:                            OLS   Adj. R-squared:                  0.545
Method:                 Least Squares   F-statistic:                     54.84
Date:                Sun, 07 Apr 2019   Prob (F-statistic):           2.93e-09
Time:                        17:46:45   Log-Likelihood:                -144.64
No. Observations:                  46   AIC:                             293.3
Df Residuals:                      44   BIC:                             296.9
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           12.1630      0.847     14.368      0.000      10.457      13.869
GDPperCapita     6.2688      0.847      7.405      0.000       4.563       7.975
==============================================================================
Omnibus:                        0.486   Durbin-Watson:                   1.869
Prob(Omnibus):                  0.784   Jarque-Bera (JB):                0.634
Skew:                           0.171   Prob(JB):                        0.728
Kurtosis:                       2.537   Cond. No.                         1.00
==============================================================================
```



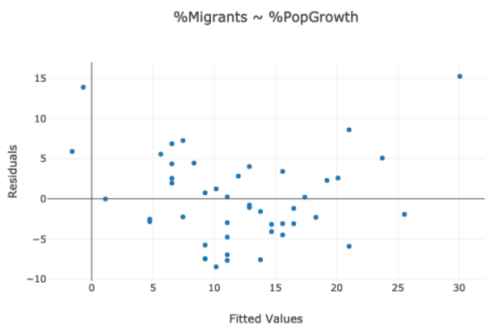
Both models above have single features with p-values that indicate their statistical significance.

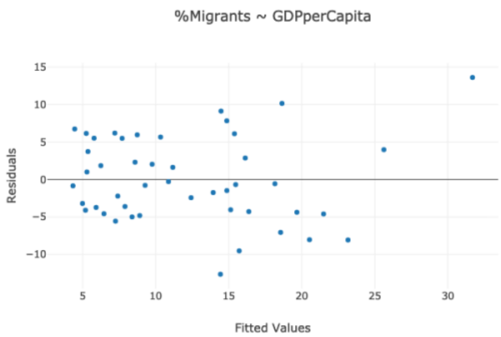
Shown below are also the results of some statistical tests executed to verify the requirements needed for our regression analysis to be sound and valid.

Model	Harvey-Collier	Jarque-Bera	Durbin-Watson	Breusch-Pagan
%Migrants ~ %PopGrowth	0.027	0.148	1.591	0.544
%Migrants ~ GDPperCapita	0.175	0.728	1.869	0.00098

The p-values of the Jarque-Bera statistic in both cases indicate the errors/residuals are normally distributed and the Durbin-Watson statistic in both models is also within the acceptable range of 1.5 to 2.5 thus suggesting the absence of any significant auto-correlation in these residuals.

However, though visually it would seem that both models display linearity, the low p-value for the Harvey-Collier test of linearity on model A would suggest otherwise.



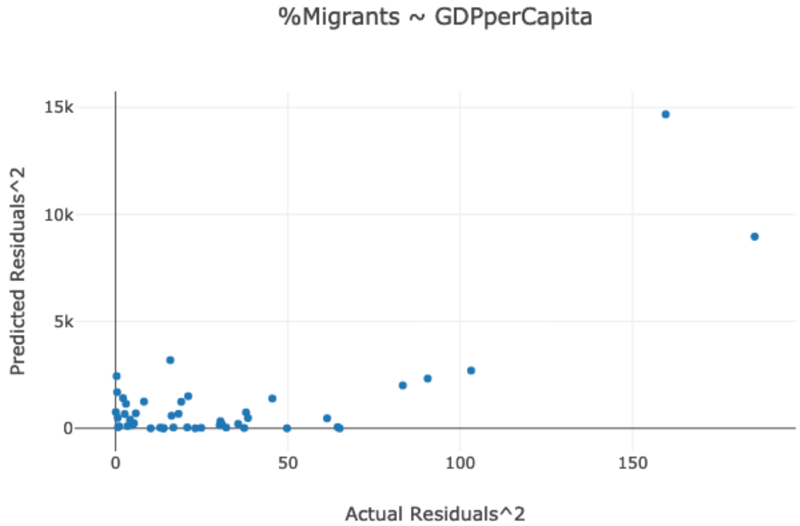


The very low p-value on the Breusch-Pagan test for model B also indicates the possibility that its residuals suffer from heteroskedasticity.

It transpires that, after removing any candidate feature that was strongly correlated (|corr|>0.5) to *%PopGrowth*, executing the semi-automated forward selection process yielded no statistically significant feature to add to the model.

We were therefore left with only model B to work with. It was decided to therefore address the problematic heteroskedasticity before proceeding any further. To obtain the plot below, we regressed the square of the actual residuals (from model B) on our feature *%GDPperCapita* to obtain the predicted residuals.

The plot below shows two clear outliers existing on the far right. These points correspond to the countries Japan and Luxembourg. Removing these two countries from our data set (so that we are left with 44 developed countries) yielded the test statistics shown on the next page.



Model	Harvey-Collier	Jarque-Bera	Durbin-Watson	Breusch-Pagan
%Migrants ~ GDPperCapita	0.886	0.332	1.548	0.068

The p-value for the Breusch-Pagan test has increased dramatically from 0.00098 to 0.068 which lies just above our significance level of 0.05. The removal of the two outlier countries has alleviated any concerns around the heteroskedasticity in the residuals while all the other stats above remain within acceptable bounds.

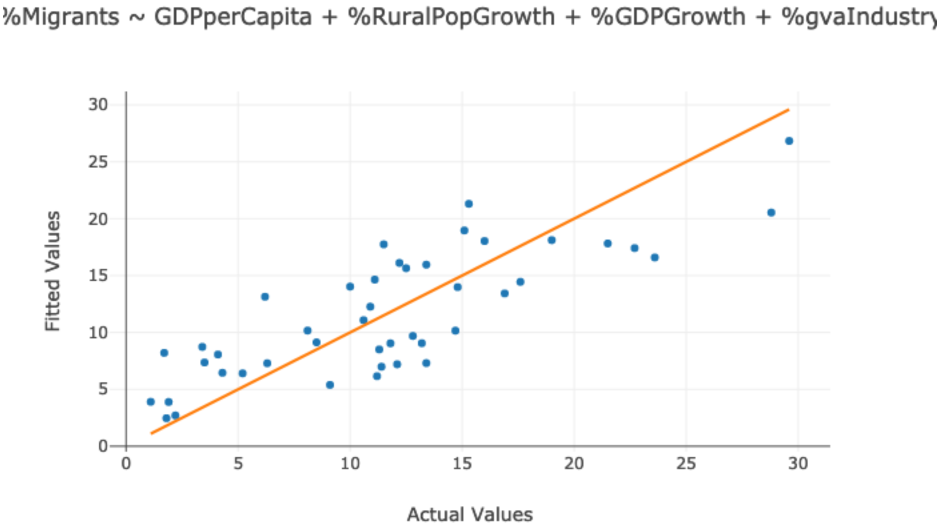
Therefore, building recursively via forward selection on model B with more confidence yielded the following model and associated stats:

%Migrants ~ GDPperCapita + %RuralPopGrowth + %GDPGrowth + %gvaIndustry

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              %Migrants   R-squared:                       0.655
Model:                            OLS   Adj. R-squared:                  0.620
Method:                 Least Squares   F-statistic:                     18.55
Date:                Sun, 07 Apr 2019   Prob (F-statistic):           1.30e-08
Time:                        20:30:45   Log-Likelihood:                -123.40
No. Observations:                  44   AIC:                             256.8
Df Residuals:                      39   BIC:                             265.7
Df Model:                           4
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              11.6455      0.640     18.192      0.000      10.351      12.940
GDPperCapita        4.0598      0.662      6.130      0.000       2.720       5.399
%RuralPopGrowth     2.4898      0.681      3.657      0.001       1.113       3.867
%GDPGrowth         -1.7621      0.660     -2.668      0.011      -3.098      -0.426
%gvaIndustry       -1.6233      0.653     -2.486      0.017      -2.944      -0.303
==============================================================================
Omnibus:                        5.745   Durbin-Watson:                   1.810
Prob(Omnibus):                  0.057   Jarque-Bera (JB):                2.183
Skew:                           0.131   Prob(JB):                        0.336
Kurtosis:                       1.941   Cond. No.                         1.45
==============================================================================
```

Harvey-Collier	Jarque-Bera	Durbin-Watson	Breusch-Pagan
NaN	0.336	1.810	0.418

Other than the Harvey-Collier test statistic above, all others illustrate that the underlying assumptions for linear regression hold true. Unfortunately, I did not have the time to investigate why the *statsmodels* routine *linear_harvey_collier()* produced a non-result. As an alternative test for linearity, I have provided the plot directly below.



One should also check for the presence of multicollinearity. Below we display the correlation matrix of the independent variables in the model along with the variance inflation factors (which are all reassuringly low in value). Again, no red flags present themselves.

	GDPperCapita	**%RuralPopGrowth**	**%GDPGrowth**	**%gvaIndustry**
GDPperCapita	1.000000	0.238307	0.025116	-0.061384
%RuralPopGrowth	0.238307	1.000000	0.202180	0.106595
%GDPGrowth	0.025116	0.202180	1.000000	-0.113240
%gvaIndustry	-0.061384	0.106595	-0.113240	1.000000

Feature Name (Alias)	Variance Inflation Factor
GDPperCapita	1.07
%RuralPopGrowth	1.13
%GDPGrowth	1.07
%gvaIndustry	1.04

We will discuss the above model and other points of note in more detail in section V.

IV. A Logistic Regression Model for Population Growth

In this section, we will be addressing the issue of population growth rate in the developed world. We will attempt to do this by constructing a logistic regression model using the same data set of 46 developed countries as used in the previous section.

The independent variable for this regression was constructed from the values contained under the "*Population annual rate of increase (percent)*" column in our data set. Essentially the independent variable, which we shall henceforth refer to as our *target* variable, is given a value of 1 if the population growth rate is greater than zero, otherwise its value is set to 0. A value of 1 therefore indicates that the population of a country is growing.

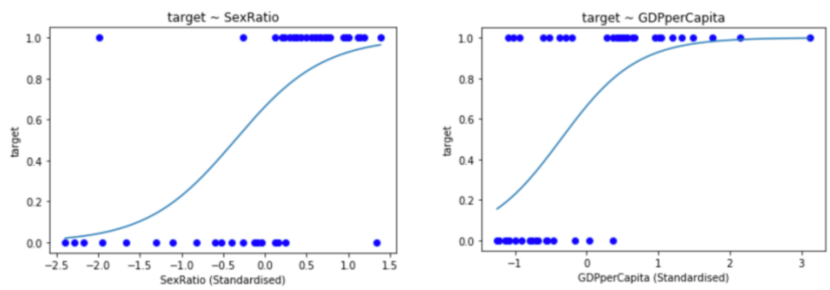
Looking at our data set of 46 countries, we find that 27 have positive growth rates with the remaining 19 either decreasing (or stagnant).

We use the same semi-automated forward selection approach (with the same criterion) to build our model. On the first stage of construction, the following candidate features present themselves:

	aic	bic	#dataPoints	p-value	featureName
1	42.319787	45.977070	46	0.000546	GDPperCapita
6	45.735555	49.392837	46	0.002076	SexRatio
22	45.944924	49.602207	46	0.005099	%Agriculture
24	47.166602	50.823885	46	0.000791	%Services

As in the previous section, *GDPperCapita* presents itself as a dominant factor on which to build our logistic regression model. Note above that we are using the Akaike and Bayesian Information Criterion (AIC & BIC) to ascertain the goodness of fit.

For illustrative purposes only I show below the fit to our data for single feature models using *GDPperCapita* and *SexRatio* (*SexRatio* is essentially the ratio of men to women in a population).



As *GDPperCapita* produced both a lower AIC (and BIC) and also had stronger statistical significance than *SexRatio*, a model of the following form was first built on the former feature:

Target ~ GDPperCapita + FertilityRate

```
                        Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.505
Dependent Variable: target           AIC:              36.8705
Date:               2019-04-07 21:53 BIC:              42.3564
No. Observations:   46               Log-Likelihood:   -15.435
Df Model:           2                LL-Null:          -31.186
Df Residuals:       43               LLR p-value:      1.4445e-07
Converged:          1.0000           Scale:            1.0000
No. Iterations:     8.0000
------------------------------------------------------------------
                  Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
------------------------------------------------------------------
const             1.9231   0.9541  2.0156 0.0438  0.0531  3.7931
GDPperCapita      2.4638   0.9816  2.5099 0.0121  0.5398  4.3878
FertilityRate     2.4746   1.1733  2.1091 0.0349  0.1750  4.7742
=================================================================
```

As you can see both the features in this model are statistically significant and are weakly correlated with a linear correlation value of 0.361.

For completeness, a second alternative model was built on the *SexRatio* feature:

Target ~ SexRatio + ParticipationTotal

```
                        Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.520
Dependent Variable: target           AIC:              35.9300
Date:               2019-04-07 22:00 BIC:              41.4159
No. Observations:   46               Log-Likelihood:   -14.965
Df Model:           2                LL-Null:          -31.186
Df Residuals:       43               LLR p-value:      9.0261e-08
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                    Coef.  Std.Err.   z    P>|z|   [0.025 0.975]
-----------------------------------------------------------------
const               0.8308   0.5205 1.5962 0.1104 -0.1893 1.8510
SexRatio            1.9569   0.5937 3.2964 0.0010  0.7934 3.1205
ParticipationTotal  1.9549   0.7494 2.6087 0.0091  0.4861 3.4236
=================================================================
```

Again, checking for the presence of multicollinearity, we find the correlation between the independent variable *SexRatio* and *ParticipationTotal* to be 0.163.

V. Discussion and Conclusion

Let us begin this section with a discussion on the multiple linear regression model constructed in section III, namely:

%Migrants ~ GDPperCapita + %RuralPopGrowth + %GDPGrowth + %gvaIndustry

The primary question we wished to address was why do some developed countries have a relatively large proportion of migrants while others have relatively few? For example, the population of Luxembourg and Switzerland consist of 45.3% and 29.6% migrants whereas that of Poland and Bosnia & Herzegovina contain only a mere 1.7% and 1.1% respectively.

The above model identifies some important factors that can explain this difference. To understand precisely how, we need the fitted values to the coefficients of the above independent variables which are stated here:

Feature Name (Alias)	Fitted Coefficient Value
GDPperCapita	4.0598
%RuralPopGrowth	2.4898
%GDPGrowth	-1.7621
%gvaIndustry	-1.6233

Clearly, the GDP per capita, having the largest coefficient, is the dominant factor determining migration levels in a nation (Recall that we standardised the independent variables before performing the regression). We have a clear positive correlation so that the richer a country, the higher the level of immigration.

Surprisingly maybe, there is also a relatively strong positive correlation with annual growth rate of the rural population *(%RuralPopGrowth*) in a country. One possibility is that in recent times, there has been an exodus of people from the cities to more affordable accommodation in more rural areas. That is, this effect may be linked to GDP per capita in that the richer a country, the more unaffordable the housing is to the younger generation. This is a however rather loose conjecture on my part.

The remaining factors above are negatively correlated to immigration levels with their effect being approximately equal in strength. They are the annual growth rate of GDP *(%GDPGrowth*) and the output (gross value added) of the industrial sector *(%gvaIndustry*) in a country. Again, this seems counterintuitive. One might have thought that high immigration into a country would boost its GDP although there may well exist a correlation between high GDP countries and low GDP growth rates (Looking at the entire 181 countries in my data set for example, there is a linear correlation of -0.127 between *GDPperCapita* and *%GDPGrowth.* This would support this argument if the magnitude of the correlation was not so weak.

Again the negative correlation of industrial output and immigration might seem strange. However, as a country becomes more developed and richer, the more it moves from industrial production to a more services based economy [8].

Taken at face value, the above model suggests that a country which is rich, has a growing rural population, a weakly growing economy and a small proportion of industrial output is more likely to have a high proportion of immigrants.

Let us now discuss the related logistic models that were constructed to address the factors determining population growth in developed countries. The first model *Target ~ GDPperCapita + FertilityRate* has a relatively simple interpretation. Taking the exponential of the fitted coefficients yields values of 11.75 and 11.88 for the respective independent variables. The fact that the fertility rate should be the more significant factor to population growth here is unsurprising. But why would a larger GDP per capita increase the growth rate? The answer here is again obvious as we established in section III that a high GDP per capita is very strongly predictive of high immigration levels. That is, our logistic model is simply saying that a nation with large fertility rate and immigration levels is more likely to have a growing population.

This is hardly revelatory. The second logistic model *Target ~ SexRatio + ParticipationTotal* is however more intriguing. First note that it is a slightly better fit to our data set than our first model (having a lower AIC and BIC). The associated factors for *SexRatio* and *ParticipationTotal* have almost identical values of 7.08 and 7.06. The model thus suggests that a country is more likely to have a positive population growth rate if there are more males than females in the country and the labour force participation is large. Interpreting this is difficult, but it must somehow be linked to the first simpler logistic model. That is, related to immigration and fertility rates in some fashion.

To conclude, I want to briefly mention some drawbacks to the present study. Firstly, the semi-automated procedure that was implemented in this paper may have been too rigid. That is, a strict correlation cut-off of 0.5 was used to discard candidate variables. The approach may also be too “greedy” in the sense that the first feature added to the model was the feature with essentially the smallest p-value.

Neither does the study take into account cultural and geographical factors. For example, Japan and the USA being large and rich countries have vastly different immigration levels primarily due to cultural reasons.

In short, the investigation reported here can easily be expanded upon in future work.

References

[1] Bloom, D.E., Canning, D. and Fink, G. (2010). Implications of population ageing for economic growth. *Oxford Review of Economic Policy* **26** (4), pp. 583-612.
[2] Adsera, A. (2004). Changing fertility rates in developed countries. The impact of labor market institutions. *Journal of Population Economics* **17** (1), pp. 17-43.
[3] Baird, D.T. et. Al. (2010). Europe the continent with the lowest fertility. *Human Reproduction Update* **16** (6), pp. 590-601.
[4] Blake, D. and Mayhew, L. (2006). On the Sustainability of the UK State Pension System in the Light of Population Ageing and Declining Fertility. *The Economic Journal* **116** (512), pp. F286-F305.
[5] Novy-Marx, R. and Rauh, J.D. (2009). The Liabilities and Risks of State-Sponsored Pension Plans. *Journal of Economic Perspectives* **23** (4), pp. 191-210.
[6] Poddar, S. (2016). European Migrant Crisis: Financial Burden or Economic Opportunity? *Social Impact Research Experience* **43**.
[7] https://unstats.un.org/unsd/methodology/m49/
[8] Buera, F.J. and Kaboski J.P. (2012). The Rise of the Service Economy. *American Economic Review* **102** (6), pp. 2540-69.

Appendix i

The following data files were downloaded from the United Nations website and used in this study. Note that we also declare the year that we extracted from each csv file.

URL to csv file	Year Extracted
https://unstats.un.org/unsd/methodology/m49/overview/	N/A
http://data.un.org/_Docs/SYB/CSV/SYB61_T02_Population,%20Surface%20Area%20and%20Density.csv	2018
http://data.un.org/_Docs/SYB/CSV/SYB60_T03_Population%20Growth,%20Fertility%20and%20Mortality%20Indicators.csv	2015
http://data.un.org/_Docs/SYB/CSV/SYB61_T04_International%20Migrants%20and%20Refugees.csv	2017
http://data.un.org/_Docs/SYB/CSV/SYB61_T03_Population%20Growth%20Rates%20in%20Urban%20areas%20and%20Capital%20cities.csv	2015
http://data.un.org/_Docs/SYB/CSV/SYB61_T13_GDP%20and%20GDP%20Per%20Capita.csv	2016
http://data.un.org/_Docs/SYB/CSV/SYB61_T14_Gross%20Value%20Added%20by%20Economic%20Activity.csv	2016
http://data.un.org/_Docs/SYB/CSV/SYB61_T17_Labour%20Force%20and%20Unemployment.csv	2018
http://data.un.org/_Docs/SYB/CSV/SYB61_T18_Employment.csv	2018
http://data.un.org/_Docs/SYB/CSV/SYB60_T25_Carbon%20Dioxide%20Emission%20Estimates.csv	2014

Appendix ii

List of candidate features (with their aliases) that were used in regressions

Alias	Original Feature Name in csv file
%Migrants	International migrant stock: Both sexes (% total population)
GDPperCapita	GDP per capita (US dollars)
%Pop0-14yrs	Population aged 0 to 14 years old (percentage)
%Pop>60yrs'	Population aged 60+ years old (percentage)
PopDensity	Population density
PopMidYear	Population mid-year estimates(millions)
'SexRatio'	Sex ratio (males per 100 females)
'InfantDeaths'	Infant mortality for both sexes (per 1,000 live births)
'LifeExpectancy'	Life expectancy at birth for both sexes (years)
'MaternalDeaths'	Maternal mortality ratio (deaths per 100,000 population)
%PopGrowth'	Population annual rate of increase (percent)
FertilityRate	Total fertility rate (children per women)
%PopInCapital'	Capital city population (as a percentage of total population)
%UrbanPopInCapital'	Capital city population (as a percentage of total urban population)
PopInCapital	Capital city population (thousands)
%RuralPopGrowth'	Rural population (percent growth rate per annum)
%UrbanPopGrowth'	Urban population (percent growth rate per annum)
%PopInUrban'	Urban population (percent)
'GDPinCurrentPrices'	GDP in current prices (millions of US dollars)
%GDPGrowth'	GDP real rates of growth (percent)
%gvaAgriculture'	Agriculture, hunting, forestry and fishing (% of gross value added)
%gvaIndustry'	Industry (% of gross value added)
%gvaServices'	Services (% of gross value added)
'ParticipationTotal'	Labour force participation - Total
'UnemploymentTotal'	Unemployment rate - Total
%Agriculture'	Employment by industry: Agriculture (%) Male and Female
%Industry'	Employment by industry: Industry (%) Male and Female
'%Services'	Employment by industry: Services (%) Male and Female
'%ArableLand'	Arable land (% of total land area)
%ForestCover'	Forest cover (% of total land area)
%ProtectedAreas'	Important sites for terrestrial biodiversity protected (% of total sites protected)
LandArea	Land area (thousand hectares)
%PermCrops'	Permanent crops (% of total land area)
Emissions	Emissions (thousand metric tons of carbon dioxide)
EmissionsPerCapita	Emissions per capita (metric tons of carbon dioxide)