**Introduction**

The hypothesis we wish to test is that armed forces rate is to investigate whether armed forces rate has an effect on whether a country is a democracy or not (democracy\_cat\_int).

To do this we create a binary response variable by applying the two functions listed in appendix one to first create a categorical variable with three levels and then by applying polityscore\_cat\_int

to create a binary response variable (1 if a country is a democracy, o if it is not).

Polity scores have the following breakdowns:

* Polity score range 6 to10: Democracy
* Polity score range -5 to 5: Anocracy
* Polity score range -10 to -6: Autocracy

After our initial investigation we build a second model including the explanatory variables:

* Female employee rate
* Armed forces rate
* Internet user rate
* Urban rate
* Income per person

The aim of this is twofold:

* To see if the additional variables effect on whether a country is a democracy or not.
* To check for evidence of confounding.

A full script to replicate the code can be downloaded here:

(<https://github.com/brennap3/Gapminder/blob/master/Gapminder_Analysis_2015.py>)

**Pre-processing the data**

Besides converting the response variable to a binary response variable we also centred all our explanatory variables with their means at zero.

**Interpreting our model: Model 1**

A summary is obtained of the logistic model (explanatory: Democracy (1, Non Democracy 0 ) and response variable centred armed forces rate ). From the summary the odds ratio and the confidence intervals can be obtained.

For armed forces rate, the values are:

5% 95% OR

0.559936 0.933543 0.722997

From the summary we can also see that the effect of the variable is statistically significant. So we can say that armed forces rate was significantly associated with democracy such that countries with higher armed forces rate were significantly less likely to be a democracy (OR= 0.722997, 95% CI=0.559936 -0.933543, p=0.013).

**Summary**

>>> print (lreg1.summary())

Logit Regression Results

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

Dep. Variable: polityscore\_cat\_int No. Observations: 146

Model: Logit Df Residuals: 144

Method: MLE Df Model: 1

Date: Sun, 14 Feb 2016 Pseudo R-squ.: 0.03797

Time: 19:42:42 Log-Likelihood: -95.756

converged: True LL-Null: -99.536

LLR p-value: 0.005970

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

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

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

Intercept 0.2986 0.172 1.735 0.083 -0.039 0.636

armedforcesrate\_centred -0.3244 0.130 -2.487 0.013 -0.580 -0.069

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

**Odds ratio Model 1:**

2.5% 97.5% OR

Intercept 1.313153 3.374140 2.104937

armedforcesrate 0.559936 0.933543 0.722997

**Second Model**

In the second model we include all centred explanatory variables in our model, these are:

* Female employee rate
* Armed forces rate
* Internet user rate
* Urban rate
* Income per person

From the summary, we can see that the armed forces rate and internet use rate are statistically significant. Looking at the extracted Odds ratios, we can see that armed forces rate was significantly associated with democracy such that countries with higher armed forces rate were significantly less likely to be a democracy (OR= 0.525738, 95% CI=0.360439-0.766844 , p=0.001). While internet use rate is also significantly associated with democracy such that countries with higher internet usage rates are more likely to be democracies (OR= 1.045871, 95% CI=1.015899 -1.076728 , p=0.003 ).

**Summary second model**

Logit Regression Results

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

Dep. Variable: polityscore\_cat\_int No. Observations: 146

Model: Logit Df Residuals: 140

Method: MLE Df Model: 5

Date: Sun, 14 Feb 2016 Pseudo R-squ.: 0.2189

Time: 19:46:12 Log-Likelihood: -77.750

converged: True LL-Null: -99.536

LLR p-value: 2.830e-08

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

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

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

Intercept 0.4541 0.215 2.117 0.034 0.034 0.875

incomeperperson\_centred 3.998e-06 4.59e-05 0.087 0.931 -8.59e-05 9.39e-05

urbanrate\_centred 0.0016 0.013 0.125 0.901 -0.024 0.027

internetuserate\_centred 0.0449 0.015 3.023 0.003 0.016 0.074

armedforcesrate\_centred -0.6430 0.193 -3.338 0.001 -1.020 -0.265

femaleemployrate\_centred -0.0139 0.015 -0.934 0.350 -0.043 0.015

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

**Odds ratio Model 2**

The odds ratio is shown below:

2.5% 97.5% OR

Intercept 1.034187 2.398046 1.574811

incomeperperson\_centred 0.999914 1.000094 1.000004

urbanrate\_centred 0.976481 1.027406 1.001620

internetuserate\_centred 1.015899 1.076728 1.045871

armedforcesrate\_centred 0.360439 0.766844 0.525738

femaleemployrate\_centred 0.957784 1.015399 0.986171

**Evidence for confounding**

The original analysis looked at the relationship between the democracy response variable (1 is a democracy, 0 is not a democracy) and armed forces rate, the explanatory variable.

However when using multiple explanatory variables, analysis found that armed forces rate is still statistically significant This would be evidence for armed forces rate not being an example of a confounded with the other explanatory variables.

**Conclusion – Whether the results gathered support the hypothesis for the association between your primary explanatory variable (Democracy) and the primary response variable (armed forces rate ).**

In conclusion based on the results from the multiple models, particularly the second model, we can see from the extracted Odds ratios that armed forces rate was significantly associated with democracy such that countries with higher armed forces rate were significantly less likely to be a democracy (OR= 0.525738, 95% CI=0.360439-0.766844 , p=0.001). While internet use rate is also significantly associated with democracy such that countries with higher internet usage rates are more likely to be democracies (OR= 1.045871, 95% CI=1.015899 -1.076728 , p=0.003 ).

Appendix 1:

def polityscore\_cat (row):

if (row['polityscore'] >=6 and row['polityscore'] <= 10 ) :

return 'Democracy'

elif (row['polityscore'] >=-5 and row['polityscore'] <= 5 ) :

return 'Anocracy'

elif (row['polityscore'] >=-10 and row['polityscore'] <= -6 ) :

return 'Autocracy'

else :

return 'NA'

def polityscore\_cat\_int (row):

if (row['polityscore\_cat']=='Democracy') :

return 1

else :

return 0

Appendix 2 preprocessing and modelling code

####week 4 logistic rergression

##select the values of interest

##'polityscore\_cat'

##'incomeperperson'

##'urbanrate'

##'internetuserate'

##'armedforcesrate'

datalogmodeltdf=data[['polityscore\_cat','incomeperperson'

,'urbanrate'

,'internetuserate','femaleemployrate'

,'armedforcesrate']].dropna()

datalogmodeltdfnona=datalogmodeltdf[(data.polityscore\_cat!='NA')]

datalogmodeltdf.dtypes

##build logistic model

datalogmodeltdfnona['polityscore\_cat']=datalogmodeltdfnona['polityscore\_cat'].astype(str)

datalogmodeltdfnona.dtypes

datalogmodeltdfnona=datalogmodeltdfnona.reset\_index()

del datalogmodeltdfnona['index']

datalogmodeltdfnona

datalogmodeltdfnonav1=datalogmodeltdfnona[['polityscore\_cat','incomeperperson','urbanrate'

,'internetuserate','femaleemployrate'

,'armedforcesrate'

]].dropna()

##recode variables

def polityscore\_cat\_int (row):

if (row['polityscore\_cat']=='Democracy') :

return 1

else :

return 0

##recode if democracy 1 else (it its anocracy or autocracy)

##calculate the age of NATO countries

##data['Years\_In\_Nato'] = data.apply (lambda row: AGE\_YEARS (row),axis=1)

datalogmodeltdfnonav1['polityscore\_cat\_int'] = datalogmodeltdfnonav1.apply (lambda row: polityscore\_cat\_int (row),axis=1)

#####Pre-processing data

datalogmodeltdfnonav1\_centered = preprocessing.scale(datalogmodeltdfnonav1[['incomeperperson','urbanrate','internetuserate','femaleemployrate','armedforcesrate']], with\_mean=True, with\_std=False) ##corrected this had wrong version of code had True and False in qoutes now its in correct format

##cast it as a dataframe

datalogmodeltdfnonav1\_centered\_df = pd.DataFrame(datalogmodeltdfnonav1\_centered)

## set the columns

datalogmodeltdfnonav1\_centered\_df.columns=['incomeperperson\_centred','urbanrate\_centred','internetuserate\_centred','femaleemployrate\_centred','armedforcesrate\_centred']

## check the count

datalogmodeltdfnonav1\_centered\_df.count()

##all look sfine

##data

##data 3 is our second subset we will use to do some analysis

##concatanate once the indexs are reset

datalogmodeltdfnonav1\_centered\_df\_cntred = pd.concat([datalogmodeltdfnonav1['polityscore\_cat\_int'], datalogmodeltdfnonav1\_centered\_df], axis=1)

datalogmodeltdfnonav1\_centered\_df\_cntred.columns.values

##

## preprocessing ended

lreg1 = smf.logit(formula='polityscore\_cat\_int ~ armedforcesrate\_centred',data = datalogmodeltdfnonav1\_centered\_df\_cntred).fit()

print (lreg1.summary())

##odds ratio

print np.exp(lreg1.params)

##little or no effect

params = lreg1.params

conf = lreg1.conf\_int()

conf['OR'] = params

conf.columns = ['2.5%', '97.5%', 'OR']

print np.exp(conf)

lreg3 = smf.logit(formula='polityscore\_cat\_int ~ incomeperperson\_centred+urbanrate\_centred+internetuserate\_centred+armedforcesrate\_centred+femaleemployrate\_centred',data = datalogmodeltdfnonav1\_centered\_df\_cntred).fit()

print (lreg3.summary())

##odds ratio

params = lreg3.params

conf = lreg3.conf\_int()

conf['OR'] = params

conf.columns = ['2.5%', '97.5%', 'OR']

print np.exp(conf)