GHS implementation status Logistic Regression Analysis

Steve Fick

Aug 8, 2017

# Introduction

The purpose of this exercise is to determine which variables in the GHS dataset best 'explain' variation GHS implementation status.

To do this we can use a multiple regression technique called *logistic regression*, in which the dependent response variable is modeled as '1' or '0' via a bionomial distribution.

This method essentially takes a regular multiple regression formula and performs a log transformation on it (thus 'logistic regression') so the response will always fall between 0 and 1.

One side-effect of this transformation is that interpretation of the effects become a bit different. Normally coefficients correspond to linear increases in the response variable ( eg in the formula y = mx + b , 1 unit increase in x results in m units increase in y). However in logistic regression, 1 unit increase in x results in a *nonlinear* increase in y which is related to m.

However, this is not a big problem for us because we are mostly interested in identifying the importance of each variable, holding others constant, rather than the direct effect of each variable on the actual probability of GHS implementation occurring.

# Method

## 0. load data

#setwd('c:/Users/stephen.fick/Dropbox/GHS 2017/scripts/20170808/')  
#setwd('c:/Users/stephen.fick/Dropbox/GHS 2017/scripts/20170808/')  
library(knitr)  
library(Hmisc)  
library(car)  
d <- read.csv(file.path('..', outdir,'data.csv'), check.names=F, strings=F)

*Recode variables*

# ghs = 1 if partial or full implementation, 0 if no implementation  
d$ghs <- as.numeric(d$`GHS implementation` >0)  
  
# reduction to trade bariers  
d$wto <- d$`WTO membership`  
d$wto <- as.factor(d$wto)  
  
d$trade <- as.numeric(as.character(d$`Trade Open-ness for 2013`))

## Warning: NAs introduced by coercion

# Commitment to safety  
  
# ILO = 1 if \*ANY\* ratifications signed  
d$ILO <- rowSums(d[, grep('ILO Conv', names(d))])  
d$ILO <- as.factor(d$ILO > 0)  
  
# Commitment to environment  
  
# env = number of environmental documents ratified  
d$env <- rowSums(d[, c("Sthlm ratification", "Rotterdam ratificaiton","Basel ratification","SAICM focal point in place=1")])  
  
#globalization  
d$global <- d[['KOF sub-index C']]  
  
#HDI  
d$hdi <- d$`HDI for 2015`  
  
# GDP per person  
d$GDPP <- d$`GDP per Capita PPP for 2015`  
  
# GDP Total  
d$GDPT <- d$`Total GDP, PPP, for 2015`

*Scale continuous variables* to promote inter-comparisons

cc <- c('GDPP', 'GDPT', 'hdi', 'global', 'trade', 'env')  
ds <-d   
ds[,cc] <- scale(d[,cc])

## 1. Evaluate correlations

x <- rcorr(as.matrix(d[,c('hdi', 'GDPP', 'GDPT','global', 'trade')]))

*Pearsons correlation*

kable(x$r)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | hdi | GDPP | GDPT | global | trade |
| hdi | 1.0000000 | 0.7251108 | 0.1897388 | 0.2918295 | 0.1354204 |
| GDPP | 0.7251108 | 1.0000000 | 0.1507982 | 0.2392583 | 0.3225656 |
| GDPT | 0.1897388 | 0.1507982 | 1.0000000 | 0.2767525 | -0.2441668 |
| global | 0.2918295 | 0.2392583 | 0.2767525 | 1.0000000 | -0.2322094 |
| trade | 0.1354204 | 0.3225656 | -0.2441668 | -0.2322094 | 1.0000000 |

HDI and GDP per capita are highly correlated (.72). Next highest correlation is trade and GD per capita (.32)

*Significance (P)*

kable(x$P)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | hdi | GDPP | GDPT | global | trade |
| hdi | NA | 0.0000000 | 0.0114243 | 0.0000672 | 0.1132703 |
| GDPP | 0.0000000 | NA | 0.0433166 | 0.0014748 | 0.0001359 |
| GDPT | 0.0114243 | 0.0433166 | NA | 0.0002184 | 0.0043185 |
| global | 0.0000672 | 0.0014748 | 0.0002184 | NA | 0.0063245 |
| trade | 0.1132703 | 0.0001359 | 0.0043185 | 0.0063245 | NA |

All correlations are highly significant except trade open-ness and HDI, as wella s GDPT and GDP-percapita

This could be problematic for our model, because it will be difficult to determine what the effect of HDI is 'independent of' GDP and vice versa, because we wont have (any) examples where GDP is is high but HDI is low to see the effects on GHS, for example.

## 2. Model fitting

Lets try to create a first model with all terms thrown in.

m0 <- glm(ghs ~ hdi + GDPP + GDPT + global + env + ILO + trade + wto, data = ds , family = 'binomial')  
  
summary(m0)

##   
## Call:  
## glm(formula = ghs ~ hdi + GDPP + GDPT + global + env + ILO +   
## trade + wto, family = "binomial", data = ds)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.31679 -0.31395 -0.00893 0.22368 2.28035   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.1975 1.4747 -2.168 0.03014 \*   
## hdi 5.7886 1.3790 4.198 2.7e-05 \*\*\*  
## GDPP -2.1543 0.7599 -2.835 0.00458 \*\*   
## GDPT 0.9452 0.5767 1.639 0.10121   
## global 1.5205 0.4857 3.130 0.00175 \*\*   
## env 0.9776 0.6179 1.582 0.11363   
## ILOTRUE 0.9726 0.6578 1.479 0.13922   
## trade 1.3829 0.4719 2.930 0.00338 \*\*   
## wto1 -0.4221 1.2827 -0.329 0.74213   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 184.298 on 133 degrees of freedom  
## Residual deviance: 65.702 on 125 degrees of freedom  
## (59 observations deleted due to missingness)  
## AIC: 83.702  
##   
## Number of Fisher Scoring iterations: 7

Ok, starting from the top we see our formula, then the deviance residuals. The deviance residuals appear approximately normal based on the quantiles, which is a good thing.

Next we see the coefficients. The intercept indicates the overall likelihood for a country to have GHS (without ILO ratification or wto ratification) The fact that this is negative means that overall, likelihood is small that a random country will have GHS implementation.

Following this we see that hdi has a very high coefficient of 5.78. This means that a 1 standard deviation increase in hdi results in 5.7 times the increase of the log odds of GHS being implemented.

The significance of this coefficient is high (small pr > z), indicating that we are very confident that it is important.

GDP per capita interestingly has a negative coefficient. This can happen, however when predictors are highly correlated -- one has to compensate for the other. In this case I think GDPP is compensating for HDI.

GDPT is positive, but not significant

globalization is positive and significant

environmental ratification is positive but not significant

ILO is the same as environmental ratification trade openness is large, positive and significant.

Interestingly WTO membership is slightly negative. However the standard error for this coefficient is high making interpretation difficult.

Lets explore this idea of correlation between GDPP and HDI, by examinging the Variance Inflation Factor (VIF) for each variable. This gives us a sense of how much the overall uncertainty in the model increases when adding a variable, compared to adding a completely uncorrelated variable. Any value greater than 2 is considered high, and > 5 is very high

vif(m0)

## hdi GDPP GDPT global env ILO trade wto   
## 3.673072 4.229146 1.483320 1.407598 1.264248 1.064238 2.068321 1.164667

We see that hdi and GDPP have high VIF, likely due to their high correlation. This is a problem in linear modeling called multi-colinearity.

As a result it makes interpretation of the colinear variables problematic, because they have uncertain standard errors. This makes the model highly sensitive to the underlying data as well.

One simple solution to this problem is to simply remove a colinear variable. I think it is reasonable to drop GDPP because it provides similar information to HDI.

## 3. Refit model dropping GDPP

m <- glm(ghs ~ hdi + GDPT + global + env + ILO + trade + wto, data = ds , family = 'binomial')

lets check the VIF

vif(m)

## hdi GDPT global env ILO trade wto   
## 1.115219 1.221259 1.576629 1.211151 1.088573 1.516519 1.164489

much more reasonable!

summary(m)

##   
## Call:  
## glm(formula = ghs ~ hdi + GDPT + global + env + ILO + trade +   
## wto, family = "binomial", data = ds)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.82926 -0.40427 -0.04208 0.27215 2.25732   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.9629 1.2206 -1.608 0.10779   
## hdi 2.7886 0.6444 4.328 1.51e-05 \*\*\*  
## GDPT 0.4518 0.3988 1.133 0.25724   
## global 1.5675 0.4841 3.238 0.00121 \*\*   
## env 0.7102 0.5712 1.243 0.21376   
## ILOTRUE 1.2134 0.6031 2.012 0.04421 \*   
## trade 0.9029 0.4341 2.080 0.03752 \*   
## wto1 -0.6808 1.1547 -0.590 0.55548   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 184.298 on 133 degrees of freedom  
## Residual deviance: 77.584 on 126 degrees of freedom  
## (59 observations deleted due to missingness)  
## AIC: 93.584  
##   
## Number of Fisher Scoring iterations: 6

This looks better to me... The HDI effect is now much lower (although highly significant) trade and globalization are both significant Also, ILO ratification is now a large and significant explanatory variable

Outputting results to table.

kable(coef(summary(m)))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -1.9629290 | 1.2205698 | -1.6082071 | 0.1077898 |
| hdi | 2.7885882 | 0.6443588 | 4.3276944 | 0.0000151 |
| GDPT | 0.4517658 | 0.3987576 | 1.1329335 | 0.2572422 |
| global | 1.5674686 | 0.4841337 | 3.2376772 | 0.0012051 |
| env | 0.7101629 | 0.5711961 | 1.2432910 | 0.2137606 |
| ILOTRUE | 1.2134138 | 0.6030558 | 2.0121086 | 0.0442085 |
| trade | 0.9028845 | 0.4340596 | 2.0800934 | 0.0375170 |
| wto1 | -0.6807724 | 1.1547022 | -0.5895653 | 0.5554821 |

Other statistics of interest: AIC (a measure of goodness of fit) -- decent, in this case 1- (Residual Deviance / Null deviance) -- analogous to R2. In this case it is 0.5790285 which is not bad.