Problem: The Tax Gap

Ben Mazzotta

Friday, October 24, 2014

Tax Gap

The tax gap is defined as the difference between the true tax burden of a country and its actual tax revenue. In the United States, a quintennial estimate is released at the IRS website using audit data. The most recent applies to tax year 2006, see (The Tax Gap)[http://www.irs.gov/uac/The-Tax-Gap]. Needless to say no comparable international study exists.

Instructions

1. Collect average effective tax rates for as many countries as possible.
2. Collect informal economy to GDP ratios for as many countries as possible.
3. Join the data.
4. Summarize the availability of data.
5. Create a scatterplot of informality and average effective tax rates.
6. Save the joined data in each of the following formats: CSV, DTA, RDA. If you have cloned this repository, make sure to save the data in the `./data` directory

```
##  This is how to pull statistics from World Development Indicators into R.
##      A similar interface exists for Stata.

#       Load the library
require("WDI")
```

```
## Loading required package: WDI
## Loading required package: RJSONIO
```

```
#       Query variables on taxes
WDIsearch("Tax revenue")
```

```
##      indicator
## [1,] "GC.TAX.EXPT.ZS"
## [2,] "GC.TAX.IMPT.ZS"
## [3,] "GC.TAX.TOTL.CN"
## [4,] "GC.TAX.TOTL.GD.ZS"
##      name
## [1,] "Taxes on exports (% of tax revenue)"
## [2,] "Customs and other import duties (% of tax revenue)"
## [3,] "Tax revenue (current LCU)"
## [4,] "Tax revenue (% of GDP)"
```

```
#       Jot down the name of the variable
taxvar <- WDIsearch("Tax revenue")[4][1]  ## Note the subscripts in square brackets
#       Query for all countries, three recent years
tax <- WDI(indicator=taxvar, country="all", start=2010, end=2012, extra=TRUE)
#       Filter for country observations
```

```r
tax <- subset(tax, region!="aggregates")
names(tax)[grep("GC.", names(tax))] <- "tax.gdp"

rm(taxvar)
summary(tax)
```

```
##     iso2c             country             tax.gdp           year     
##  Length:735         Length:735         Min.   : 0.02   Min.   :2010  
##  Class :character   Class :character   1st Qu.:13.02   1st Qu.:2010  
##  Mode  :character   Mode  :character   Median :16.13   Median :2011  
##                                        Mean   :16.74   Mean   :2011  
##                                        3rd Qu.:20.63   3rd Qu.:2012  
##                                        Max.   :37.64   Max.   :2012  
##                                        NA's   :311                   
##      iso3c                                                     region   
##  ABW    :  3   Europe & Central Asia (all income levels)     :171  
##  AFG    :  3   Sub-Saharan Africa (all income levels)        :141  
##  AGO    :  3   Latin America & Caribbean (all income levels) :123  
##  ALB    :  3   East Asia & Pacific (all income levels)       :108  
##  AND    :  3   Aggregates                                    : 96  
##  ARB    :  3   Middle East & North Africa (all income levels): 63  
##  (Other):717   (Other)                                       : 33  
##         capital        longitude        latitude   
##             :111             :111             :111  
##  Abu Dhabi  :  3   -0.126236:  3   -0.229498:  3  
##  Abuja      :  3   -0.20795 :  3   -1.27975 :  3  
##  Accra      :  3   -1.53395 :  3   -1.95325 :  3  
##  Addis ababa:  3   -10.7957 :  3   -11.6986 :  3  
##  Agana      :  3   -13.2134 :  3   -12.0931 :  3  
##  (Other)    :609   (Other)  :609   (Other)  :609  
##                   income            lending   
##  Aggregates           : 96   Aggregates    : 96  
##  High income: nonOECD:114   Blend         : 45  
##  High income: OECD    : 93   IBRD          :186  
##  Low income           :105   IDA           :195  
##  Lower middle income :165   Not classified:213  
##  Not classified       :  3                       
##  Upper middle income :159                       
```

```r
#       Now for informal economy.
WDIsearch("informal sector")
```

```
##      indicator             
## [1,] "IC.CNS.INFM.ZS"      
## [2,] "IC.FRM.COMP.ZS"      
## [3,] "IC.FRM.INFOR.INFOR2 "
## [4,] "IC.FRM.OBS.OBST12   "
## [5,] "SL.TLF.IFRM.UR.FE.ZS"
##      name
## [1,] "Practices Informal Sector (% of managers surveyed ranking this as a major constra
## [2,] "Firms identifying practices of competitors in the Informal Sector as a major const
## [3,] "Percent of firms identifying practices of competitors in the informal sector as a
## [4,] "Percent of firms choosing practices of the informal sector as their biggest obstac
## [5,] "Urban informal sector employment, female (% of total urban female employment)"
```

```
cat("Rats.")
```

```
## Rats.
```

Since the World Bank has no readily available measures of employment in the informal sector, here is the ILO website that measures the same.

http://laborsta.ilo.org/informal_economy_E.html

Good instructions on reading data formats are available with Google searches. I recommend bookmarking

- UCLA pages on learning R
- Quick-R
- StackExchange

Merge data in R.

I will give an example readily available in WDI, rather than showing you how to import ILO data here.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
##      Pull income per capita
incvar <- WDIsearch("GNI per capita")[2][1]
inccap <- WDI(incvar, country="all", start=2012, end=2012)
names(inccap)[grep("NY.", names(inccap))] <- "inccap"
names(inccap); names(tax)
```

```
## [1] "iso2c"   "country" "inccap"  "year"
```

```
##  [1] "iso2c"     "country"   "tax.gdp"   "year"      "iso3c"
##  [6] "region"    "capital"   "longitude" "latitude"  "income"
## [11] "lending"
```

```
##      This is the merge command
df <- merge(inccap, tax, by=c("country","iso2c"))
names(df)
```

```
##  [1] "country"   "iso2c"     "inccap"    "year.x"    "tax.gdp"
##  [6] "year.y"    "iso3c"     "region"    "capital"   "longitude"
## [11] "latitude"  "income"    "lending"
```

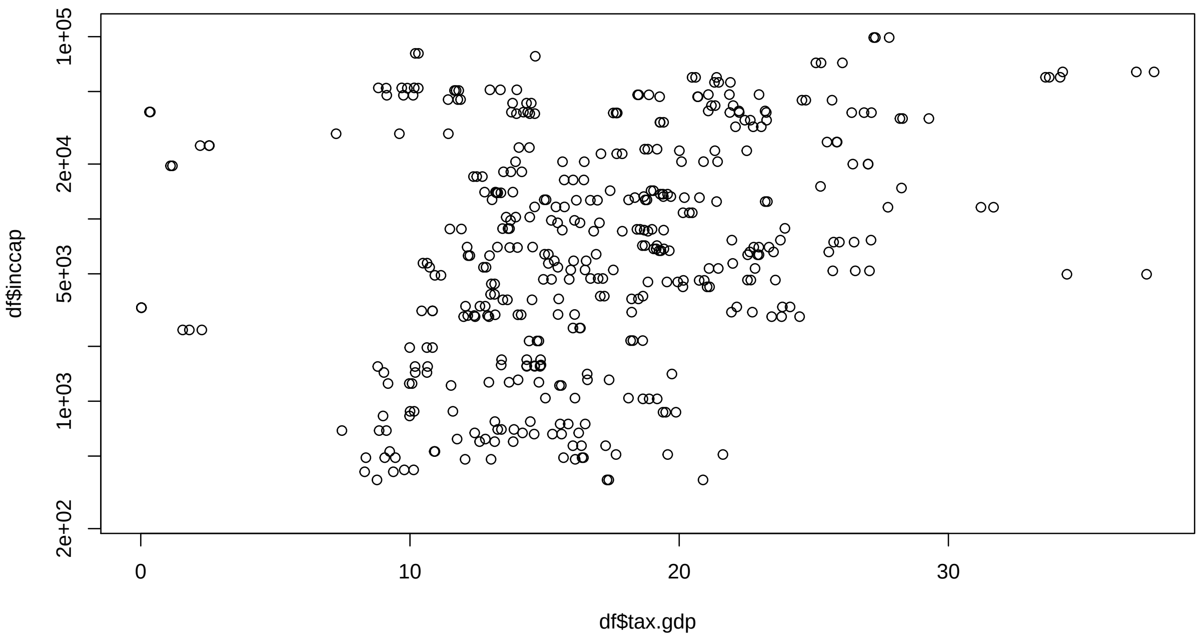
```
##      Keep only the useful variables
df <- df[,c("iso2c","inccap","tax.gdp","region","income","lending","year.y")]
###     Rename the awkward "year.y"
names(df)[names(df) %in% "year.y"] <- "year"
summary(df)
```

```
##      iso2c               inccap          tax.gdp
##   Length:735         Min.   :   240   Min.   : 0.02
##   Class :character   1st Qu.:  1580   1st Qu.:13.02
##   Mode  :character   Median :  5430   Median :16.13
##                      Mean   : 13243   Mean   :16.74
##                      3rd Qu.: 14040   3rd Qu.:20.63
##                      Max.   :104610   Max.   :37.64
##                      NA's   :96       NA's   :311
##                                                region
##   Europe & Central Asia (all income levels)     :171
##   Sub-Saharan Africa (all income levels)        :141
##   Latin America & Caribbean (all income levels) :123
##   East Asia & Pacific (all income levels)       :108
##   Aggregates                                    : 96
##   Middle East & North Africa (all income levels): 63
##   (Other)                                       : 33
##                    income               lending          year
##   Aggregates          : 96   Aggregates    : 96   Min.   :2010
##   High income: nonOECD:114   Blend         : 45   1st Qu.:2010
##   High income: OECD   : 93   IBRD          :186   Median :2011
##   Low income          :105   IDA           :195   Mean   :2011
##   Lower middle income :165   Not classified:213   3rd Qu.:2012
##   Not classified      :  3                        Max.   :2012
##   Upper middle income :159
```

```
str(df)
```

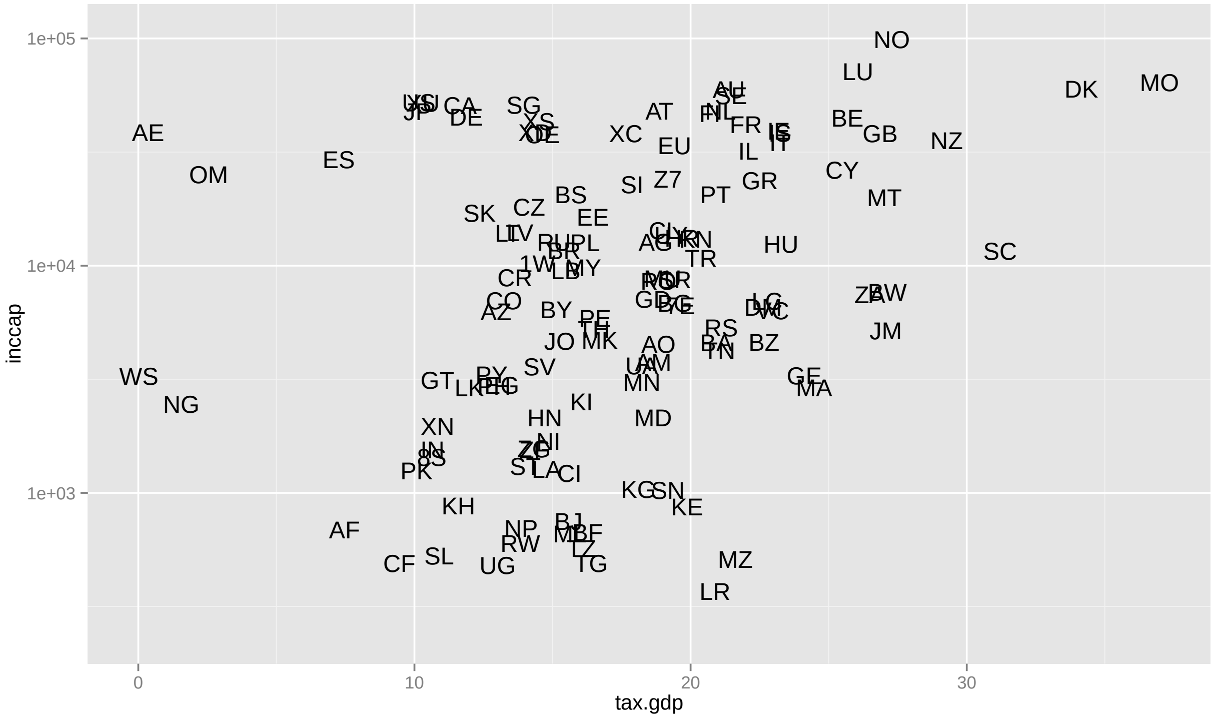
```
## 'data.frame':    735 obs. of  7 variables:
##  $ iso2c  : chr  "AF" "AF" "AF" "AL" ...
##  $ inccap : num  690 690 690 4520 4520 4520 4970 4970 4970 NA ...
##  $ tax.gdp: num  7.47 9.12 8.85 NA NA ...
##  $ region : Factor w/ 8 levels "Aggregates","East Asia & Pacific (all income levels)",..
##  $ income : Factor w/ 7 levels "Aggregates","High income: nonOECD",..: 4 4 4 7 7 7 7 7
##  $ lending: Factor w/ 5 levels "Aggregates","Blend",..: 4 4 4 3 3 3 3 3 3 5 ...
##  $ year   : num  2012 2010 2011 2010 2012 ...
```

```
##      Simple plot with the base functions
plot(df$tax.gdp, df$inccap, log="y")
```



```
##      Fancier GGPLOT2 package
qplot(tax.gdp, inccap, data=subset(df, year==2012), geom="text", label=iso2c, log="y")
```

```
## Warning: Removed 120 rows containing missing values (geom_text).
```



Next steps

After you've done that, consider what you would do about incomplete data. How could you model missing data?

Ben Mazzotta is a postdoc at IBGC. Fork me on Github. Check out the CCglobal repository Fletcher.