naivereg: a user's manual

Shuilian Cai[†], Qingliang Fan[‡*] and Wei Zhong[†]

[†]Wang Yanan Institute of Studies in Economics (WISE), Xiamen University
[‡] The Chinese University of Hong Kong

1. Introduction

This is a document about the usage of R package named **naivereg** (Fan, He and Zhong, 2020)[1] with the trade and economic growth data as an application. The original paper for this R package is Fan & Zhong (2018). The data is mainly growth, trade and geographic data of 158 countries in 2017. Specifically, the data includes the actual trade share, the total population of each country, and per capita income which are from the Penn World Table, PWT 9.1; Total area, land area, water area, forest cover data, the proportion of cultivated land (the ratio of cultivated land area to the total area), the length of coastlines, the length of inland borders, the number of official and widely used languages in each country, whether it is an inland country are all from Frankel & Romer (1999) (FR99 henceforth) and The World Factbook.

The causal effect of trade on growth is a classic economics question that is still a hot topic in public debates nowadays. In empirical research on trade and growth, an important issue is the endogenous nature of trade variable, which is attributed to some unobserved factors that affect both trade and growth. In order to understand the impact of trade on income, one could consider regressing income per capita on the trade share (the ratio of imports and exports to GDP), but this does not correctly reflect the impact of trade on income, because the variable trade share is an endogenous variable. FR99 studied the relationship between growth and trade using a two-stage least squares method. We know that in the first-stage of the two-stage least squares, the linear reduced form equation is used to fit the trade share with the instrumental variables (which we elaborate in the following), but the true relationship between them may be non-linear and the functional form is unknown to the researcher. Also, given a large number of exogenous variables, it is important to know which variables are strong IV, and which variables are irrelevant.

Following FR99, we use the gravity model of trade to circumvent the endogeneity problem of trade using instrumental variable method. The gravity theory of trade states that the size of the countries (population, area, etc.) and the distance between them (up to a gravity parameter) determine the trade volume between two countries. The instrument is the proxy variable for trade, which is constructed using geographical variables such as the country size, common border, and bilateral distance of two countries. We extend the cross-sectional study of FR99 by considering more potential instruments. Besides the original instruments, that is, proxy for trade, total population, and total land area (the latter two are included exogenous variables), we also include total water area, coastline, the arable land as percentage of total land, land boundaries, forest area as percentage of land area, the number of official and other commonly used languages in a country, and the interaction terms of constructed trade proxy with these variables (in total 15 instruments). Most of these instruments are geographical variables, which are fixed in the dataset (hence exogenous), and they can only affect growth through the channel of trade.

Following FR99, the structural equation we consider here is:

$$lnY_i = \alpha + \beta T_i + \epsilon_i$$

*Qingliang Fan: michaelqfan@gmail.com
[1]https://cran.r-project.org/web/packages/naivereg/index.html

where Y_i is the GDP per worker in country i, T_i is the share of international trade to GDP, and ϵ_i is the unobserved random disturbances, for $i = 1, 2, ..., 158$.

The reduced form model we consider is:

$$T_i = \mu + \sum_{j=1}^{15} f_j(z_{ij}) + \xi_i$$

where $f_j(\cdot)$ is the jth unknown smooth univariate function and z_{ij} is the ith observed value of the aforementioned jth instrument, $j = 1, 2, ..., 15$.

2. naivereg usage: an R code implementation of trade and growth example

First, we should load the package **naivereg** and load the data **"TradeAndGrowthData"** which is included in package **naivereg**. If you don't have this package installed, you should run the command **install.packages("naivereg")** in the console window first.

```
> library(naivereg)
> data("TradeAndGrowthData")
> data1=TradeAndGrowthData[-1,] # delete the country Aruba for existing NA
> data2=data1[,c("country","code","y","T","T_hat","N","A","water","coast","arable","border",
+       "forest","lang","in_water","in_coast","in_arable","in_border","in_forest","in_lang")]
> dim(data2)
```

```
## [1] 158  19
```

```
> summary(data2)
```

```
##       country         code           y                T          
##  Albania  :  1   AGO    :  1   Min.   : 7.463   Min.   :0.1981  
##  Algeria  :  1   ALB    :  1   1st Qu.: 9.385   1st Qu.:0.5474  
##  Angola   :  1   ARE    :  1   Median :10.416   Median :0.7579  
##  Argentina:  1   ARG    :  1   Mean   :10.177   Mean   :0.8657  
##  Armenia  :  1   ARM    :  1   3rd Qu.:11.081   3rd Qu.:1.0327  
##  Australia:  1   AUS    :  1   Max.   :12.026   Max.   :4.1287  
##  (Other)  :152   (Other):152                                    
##      T_hat               N                A              water          
##  Min.   :0.01525   Min.   :-3.037   Min.   : 5.697   Min.   :     0.0  
##  1st Qu.:0.05745   1st Qu.: 0.327   1st Qu.:10.462   1st Qu.:   192.5  
##  Median :0.07918   Median : 1.480   Median :12.015   Median :  2365.0  
##  Mean   :0.09253   Mean   : 1.382   Mean   :11.726   Mean   : 25378.4  
##  3rd Qu.:0.12223   3rd Qu.: 2.618   3rd Qu.:13.251   3rd Qu.: 10515.0  
##  Max.   :0.29681   Max.   : 6.674   Max.   :16.611   Max.   :891163.0  
##                                                                        
##      coast              arable           border            forest       
##  Min.   :     0.0   Min.   : 0.5577   Min.   :    0.0   Min.   : 0.00  
##  1st Qu.:    56.5   1st Qu.:24.5123   1st Qu.:  574.5   1st Qu.:10.74  
##  Median :   523.0   Median :42.0623   Median : 1899.5   Median :30.62  
##  Mean   :  4268.6   Mean   :40.9477   Mean   : 2837.8   Mean   :29.89  
##  3rd Qu.:  2433.5   3rd Qu.:57.3548   3rd Qu.: 3932.0   3rd Qu.:45.38  
##  Max.   :202080.0   Max.   :82.5597   Max.   :22147.0   Max.   :98.26  
##                                                                        
```

```
##       lang           in_water           in_coast           in_arable
##  Min.   : 1.000   Min.   :    0.00   Min.   :    0.000   Min.   : 0.03339
##  1st Qu.: 1.000   1st Qu.:   16.84   1st Qu.:    3.869   1st Qu.: 1.64871
##  Median : 1.000   Median :  201.54   Median :   40.365   Median : 3.18411
##  Mean   : 1.873   Mean   : 1884.56   Mean   :  354.902   Mean   : 3.82292
##  3rd Qu.: 2.000   3rd Qu.:  861.49   3rd Qu.:  142.774   3rd Qu.: 5.35419
##  Max.   :16.000   Max.   :87556.26   Max.   :19854.247   Max.   :19.40765
##
##    in_border         in_forest          in_lang
##  Min.   :   0.00   Min.   : 0.0000   Min.   :0.01680
##  1st Qu.:  64.84   1st Qu.: 0.7589   1st Qu.:0.06649
##  Median : 185.00   Median : 1.9619   Median :0.11257
##  Mean   : 243.75   Mean   : 2.7024   Mean   :0.16973
##  3rd Qu.: 307.08   3rd Qu.: 3.4932   3rd Qu.:0.16903
##  Max.   :2231.55   Max.   :20.5731   Max.   :1.48030
##
```

The dependent variable is log income per person (y), the endogenous variable is actual trade share (T), the instruments are the proxy for trade (T_hat), log total population (N), log total land area (A), total water area (water), the length of coastline (coast), the arable land as percentage of total land (arable), the length of land boundaries (border), forest area as percentage of land area (forest), the number of official and other commonly used languages in a country (lang), and the interaction terms of constructed trade proxy with these variables (in_water, in_coast, in_arable, in_border, in_forest, in_lang). Other exogenous variables in the structural equation include total population (N), and total land area (A).

Now we use **naivereg** function to do regression analysis.

```
> y1=data2[,"y"]
> x1=as.matrix(data2[,c("T","N","A")])
> z1=as.matrix(data2[,c(5,8:19)])
> # The default is regressing with the intercept and apply the BIC criterion
> naive = naivereg(y1,x1,z1,endogenous.index = c(1,0,0))
```

```
## $beta.endogenous
## [1] 0.9101234
##
## $beta.exogenous
## [1]  0.11356320 -0.07114204 10.08781978
##
## $std.endogenous
## [1] 0.3159906
##
## $std.exogenous
## [1] 0.08117524 0.05856197 0.67232469
##
## $n
## [1] 158
##
## $degree
## [1] 1
##
## $criterion
## [1] "BIC"
##
```

```
## $ind
## [1]  1 13 14 15
##
## $ind.b
## [1]  1 13 14 15
##
## $res
##   [1] -0.676053122  0.313725468  0.964360627  0.931173901  0.101542682
##   [6]  1.833903403  1.148412349  0.013589995 -2.380407961  0.542479545
##  [11] -1.529865075 -1.520393888 -0.947113728  0.026506614  0.198276897
##  [16]  0.904471301  0.837634096 -0.070520651 -0.325030894 -0.241791818
##  [21]  0.559346217  0.212812072 -0.136690812  0.783659941 -2.064202294
##  [26]  1.508136421  0.989288020  1.014948601  0.005370403 -0.704210980
##  [31] -0.832068007  0.448735177 -0.389043916 -0.585002084  0.577113563
##  [36]  0.864677450  0.315975059  1.048646984 -1.111510273  1.113519172
##  [41]  0.490199958  0.883829241  0.147401511  0.546625582  1.319713275
##  [46]  0.383359329 -1.716349615  1.455760524 -0.270705421  1.413123975
##  [51]  0.973984753  1.216417685  0.020668792 -0.931270899 -1.638752083
##  [56] -0.818575078 -1.443778775  0.637973053  1.055453376  0.125917238
##  [61] -0.146336349 -1.126019972  0.838447241 -1.923908385  0.015194257
##  [66]  0.089478957 -0.457631246  0.741606483  1.087000805  1.185530727
##  [71]  1.514720537  1.208846649  1.369445031 -0.327000398  0.313998151
##  [76]  1.252892390  0.930279348 -0.925596019 -1.086425855 -1.906650500
##  [81]  0.824436193  0.893226626 -0.604490961  0.536360969 -2.174505319
##  [86]  0.172120072  0.492076374 -1.342045548  0.290733799 -1.349495384
##  [91]  0.569447077 -0.217633337 -0.739620576 -1.846481003 -0.379599211
##  [96]  0.433460715  0.017065571 -1.147511458 -0.362165046  0.151308693
## [101] -2.344819868 -0.660761765  0.480150371 -2.194880028  0.195731085
## [106]  0.670405057 -2.169409614 -0.373608965 -0.871422722  0.446227866
## [111]  1.669189242 -1.660277285  1.391231346  0.773523548 -0.105118590
## [116]  0.634344311  0.192194190 -0.328769469  0.660491986  0.776187405
## [121] -0.304424279  1.337957006  0.824481830  0.993380898 -1.790196174
## [126]  1.671612546 -0.778021922 -0.891586027 -1.681569881 -0.447922246
## [131] -0.571658344  0.626127352 -0.001884794  0.366682769  1.331793149
## [136] -0.159589879 -0.113882381 -1.514030831 -2.088966165 -0.385085086
## [141] -0.483309658  0.348408392  1.116956192  0.248972527  1.294466943
## [146] -1.181263319 -1.452752164 -0.227439923  0.998508218  1.911842343
## [151] -0.037271250  0.040428502  0.215680609 -2.113052902 -0.497952041
## [156]  0.603246066 -0.461947576 -1.783269222
```

The **naivereg** will return the estimated coefficients and the standard deviations of the variables in design matrix, the degree of B-spline, the criterion for selecting the IVs, the index of chosen instruments and the B-spline basis functions, and the residuals of the structural equation. Besides that, **naivereg** can also return the t value and the 95% CI (confidence interval). The users could customize the function parameters (see details in the help file in the package).

```
> naive$t.endogenous
```

```
## [1] 2.880223
```

```
> naive$t.exogenous
```

```
## [1]  1.398988 -1.214816 15.004387
```

```
> naive$endogenous.conf.interval.lower
```

```
## [1] 0.2907818
```

```
> naive$endogenous.conf.interval.upper
```

```
## [1] 1.529465
```

```
> naive$exogenous.conf.interval.lower
```

```
## [1] -0.04554028 -0.18592350  8.77006339
```

```
> naive$exogenous.conf.interval.upper
```

```
## [1]  0.27266667  0.04363943 11.40557617
```

0.91 is the estimated coefficient of interested parameter β.

Table 1: Estimation results for the trade and income data

	Intercept	Trade Share	Ln Population	Ln Area
coefficient	0.11	0.91	-0.07	10.08
sd	0.08	0.31	0.05	0.67
t	1.39	2.88	-1.21	15.00
CI	[-0.04,0.27]	[0.29,1.52]	[-0.18,0.04]	[8.77,11.4]

3. Combine IVselect and 2SLS

This section shows how to use the IVselect function in package naivereg to select the IVs and get the predicted value of T (trade). We then use the predicted value of T and the exogenous variables in TSLS.

```
> N=data2$N
> A=data2$A
> T=data2$T
> y=data2$y
> #### select IVs
> fun1=IVselect(z1,x1,endogenous.index = c(1,0,0))
> fun1$degree
```

```
## [1] 1
```

```
> fun1$ind
```

```
## [1]  1 13 14 15
```

```
> colnames(cbind(z1,N,A))[fun1$ind]
```

```
## [1] "T_hat"   "in_lang" "N"       "A"
```

The selected IVs are the proxy for trade (T_hat), the interaction terms of constructed trade proxy with the number of official and other commonly used languages in a country (in_lang), the log total population (N) and the log total land area (A).

```
> z=fun1$IVselect
> z=data.frame(z)
> ##2sls
> #step1:regress endogenous variable trade on selected IVs
> d1=data.frame(T,z)
> f1=lm(data=d1, d1$T~.)
> hatx1=f1$fitted.values ## predicted trade
> summary(f1)
```

```
##
## Call:
## lm(formula = d1$T ~ ., data = d1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95943 -0.22478 -0.00371  0.16254  2.33202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.55053    0.27122   5.717 5.52e-08 ***
## X1           3.07840    0.71137   4.327 2.71e-05 ***
## X13          0.28875    0.18706   1.544 0.124758
## X14         -0.03760    0.03046  -1.234 0.218914
## X15         -0.08245    0.02413  -3.417 0.000812 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4207 on 153 degrees of freedom
## Multiple R-squared:  0.3616, Adjusted R-squared:  0.3449
## F-statistic: 21.67 on 4 and 153 DF,  p-value: 3.516e-14
```

```
> #step2:regress y on predicted trade and exogenous variables
> d2=data.frame(y,hatx1,N,A)
> f2=lm(data=d2,d2$y~.)
> summary(f2)
```

```
##
## Call:
## lm(formula = d2$y ~ ., data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6965 -0.7515  0.1545  0.8520  1.9648
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.08635    1.18300   6.835  1.8e-10 ***
## hatx1        1.28987    0.46251   2.789  0.00596 **
## N            0.01581    0.07585   0.208  0.83517
## A            0.08120    0.07895   1.029  0.30531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.079 on 154 degrees of freedom
## Multiple R-squared:  0.05919,    Adjusted R-squared:  0.04086
## F-statistic:  3.23 on 3 and 154 DF,  p-value: 0.02413
```

The estimated coefficient of interested parameter β is 1.29.

4. References

Fan, Q. and Zhong, W. (2018), "Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective." Journal of Business & Economic Statistics, 36(3), 388-399.

Frankel, J. and Romer, D. (1999), "Does Trade Cause Growth?" The American Economic Review, 89(3), 379-399.

Fan, Q., He, K. and Zhong, W. (2020), R package naivereg: Nonparametric Additive Instrumental Variable Estimator and Related IV Methods. Version 1.0.5. Published on 2020-03-18. https://cran.r-project.org/web/packages/naivereg/index.html.