ECO 102: Topics in Economics
Ecole Polytechnique, Spring 2022
Professors: Geoffroy Barrows, Benoit Schmutz
Teaching Assistants: Arnault Chatelain, Maddalena Conte

TD4: Instrumental Variables

Today, we will follow Albouy (2012)'s critic of Acemoglu et al. (2001)' Instrumental Variable (IV). As usual, download the folder TD4.zip, type your answers directly in the Tex script and your Stata commands in a do file.

Exercise

1. A refresher on IVs:

   (a) What is an Instrumental Variable? Why do we use them?
   Answer: When the $E[\epsilon|X] \neq 0$ which might lead to the biased result of OLS.Therefore, in order to solve this problem we introduce a IV, A valid instrument induces changes in the explanatory variable but has no independent effect on the dependent variable, allowing a researcher to uncover the causal effect of the explanatory variable on the dependent variable.

   Correction: It is a variable used to correct endogeneity issues with one of the explanatory variable. We use them to get correct estimates when we have an endogeneity issue.

   (b) In the case of Acemoglu et al. (2001), can you recall what they study, what variable they use as an IV and why?
   Answer: Acemoglu at al. (2001) want to research the causality of political institution and the impact on Economics development. We choose to use log GDP per capita PPP as a economics dependent variable and Expropriation risk to approximately evaluate the effect of political institution. Since most countries they research on was former European colonies and we use the settlers mortality rate as an IV for today's expropriation risk. They argue that it should have impacted the institutions European put in place in the colonies: in places where Europeans faced high mortality rates, they could not settle and were more likely to set up extractive institutions.

   We choose settler mortality rates in the period 1500 - 1800 as our IV which is correlated to expropriation risk and not correlated to lnGDP. Since the settler mortality are correlated to the political institutions that European establish on that time period and it's uncorrelated to the GDP development since it's 200years ago. However there is more precise prove to this conclusion

   Correction: They try to assess causaly how institutions can affect economic performance. They use log gdp per capita PPP as a dependent variable and approximate the effect of institutions using Expropriation Risk. They focus on former European colonies (broadly conceived)and use the settlers mortality rate as an IV for today's expropriation risk.

(c) What are the two conditions that should be met when using an IV?

Answer:
$cov(Exp.risk, settler) \neq 0$
$cov(settler, u) = 0$
Since in this research we consider the endogenous explanatory variable is the expropriation risk therefore the IV we choose have to be correlated to this variable and have no direct effect on the dependent variables which is the log GDP per capita.

Correction:
1. Strong correlation with the endogenous explanatory variable.
2. Exclusion restriction: the instrument is not correlated with the error term of the explanatory equation controling on other covariates.
This implies in particular that there is no omitted variable that is both correlated with the instrument and the dependent variable (see graph on board during the TD) and that the instrument has no direct effect on the dependent variable (the only effect it has is through the explanatory variable which is instrumented.

If the model is y =a+bx+ϵ and z is the instrument, the two conditions can be rewritten:
$cov(u, x) \neq 0$
$cov(z, u) = 0$).

2. Go to Moodle and download the dataset ajrcomment.dta. Let us start with a short comment on the use of log in economics.

(a) In both papers the authors use log GDP per capita in 1995. Create a variable gdp that contains the GDP per capita in 1995. (We are referring to the natural logarithm here).

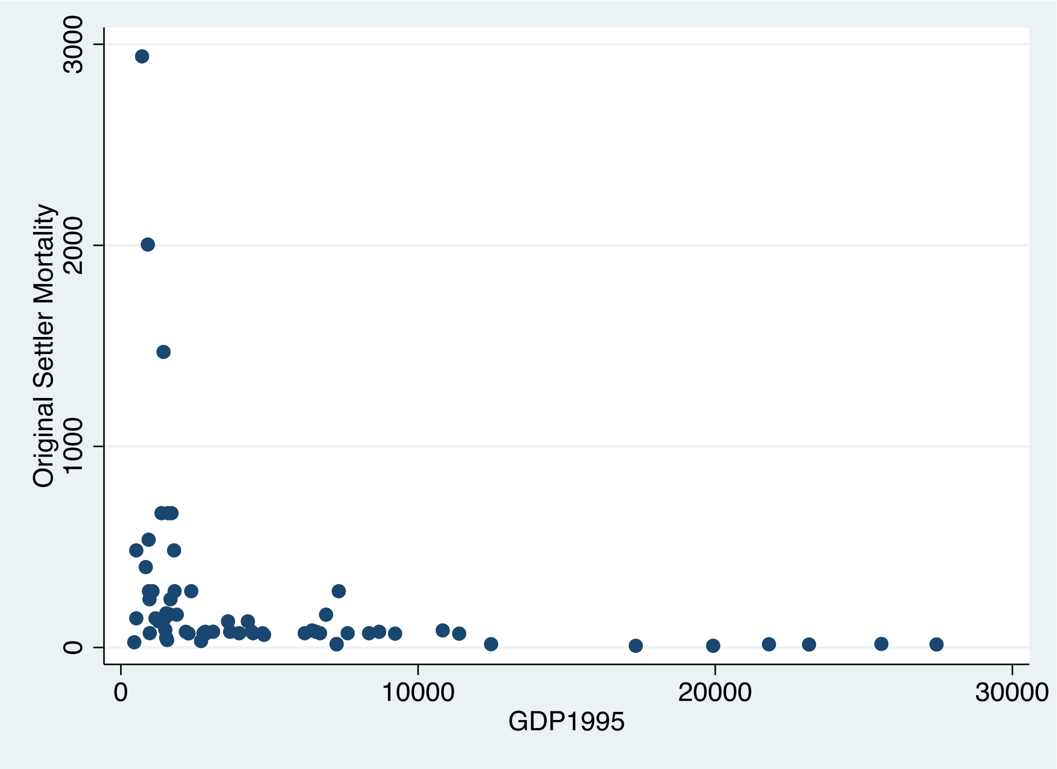


Figure 1: Graph without using nature log

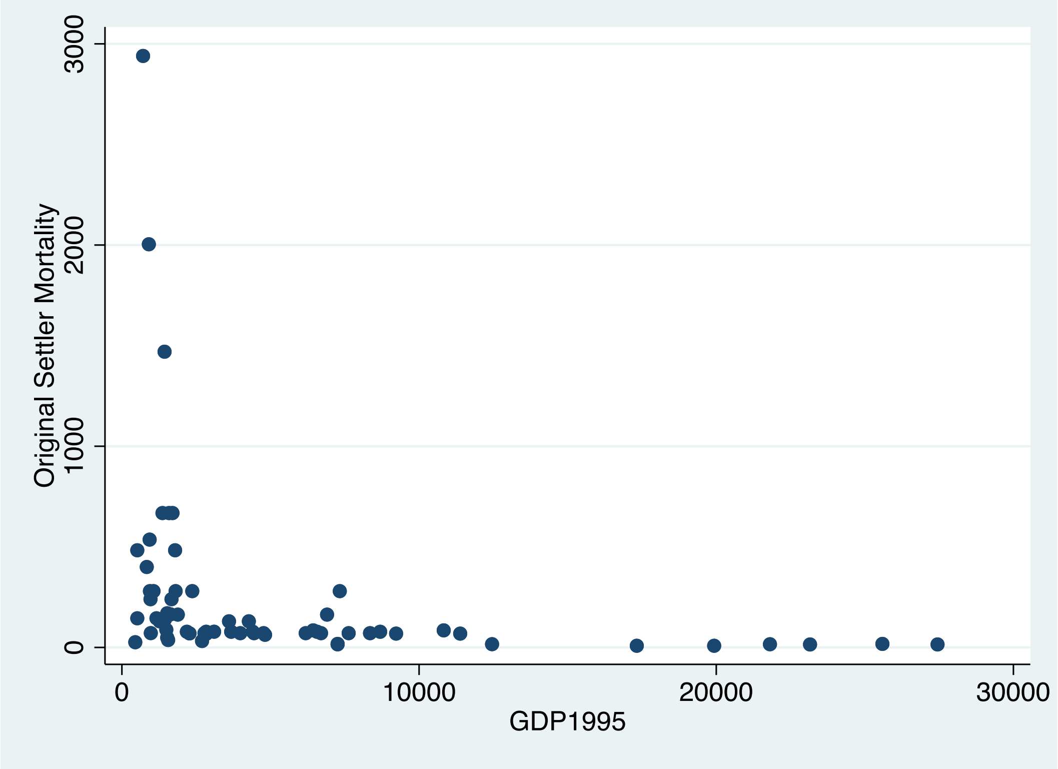


Figure 2: Graph using nature log

(b) Create a scatter plot of mortality rates (y-axis) against GDP per capita in 1995 (x-axis). Do the same with log mortality rate and log GDP. Compare the two. What do you observe?
Answer: Due to outliers the table without log values stacked in one corner of the graph.

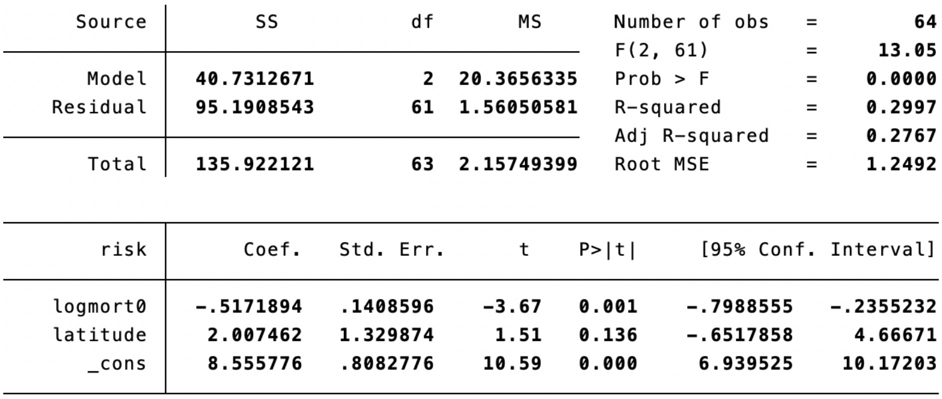
(c) What is the benefit of using log values in tables?
Answer: Many variables grows exponentially, therefore sometimes there are outliners in the graph which makes the graph unreadable, therefore we need log to concentrate the data points on the graph.

Correction: With log the growth is multiplicative not additive. This makes tables containing variables with a large range of values much more readable. This is especially useful when the variable grows exponentially (e.g. gdp growth).

3. Let us now reproduce the results of Acemoglu et al. (2001).

   (a) Write down the two equations you will use in your 2 Stage Least Square (2SLS) regression (include a single control for Latitude).
   Answer:
   1st stage: $exp\_risk = b_0 + b_1 * logmort0 + b_2 * Latitude + \epsilon_i$
   2st stage: $y_i = a_0 + a_1 * \widehat{exp\_risk} + a_2 * Latitude + \epsilon_2$
   PIC $\widehat{exp\_risk}$ being the predicted value of risk derived from the first stage.

   (b) Run the first stage equation. Does the instrument seem valid?
   Answer: Since we can tell from the p-value of risk against logmort0 is 0.001 which is less then 1% , which implies that IV is valid. Here is some explanation about p-value.



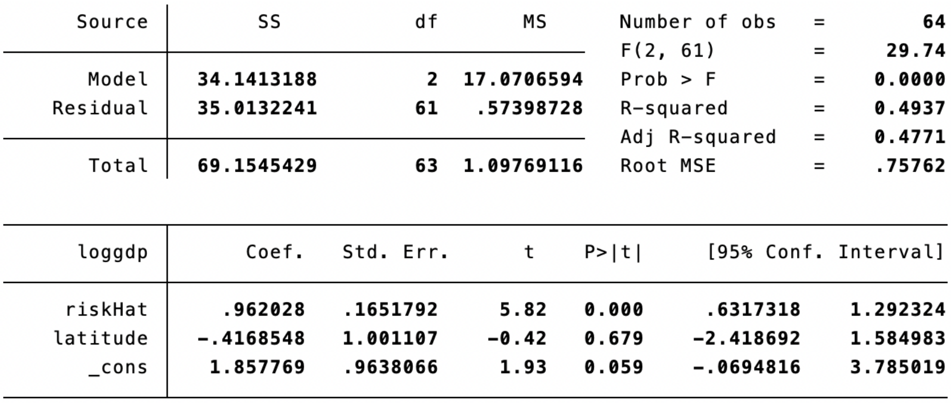
```
      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(2, 61)        =     13.05
       Model |  40.7312671         2  20.3656335   Prob > F        =    0.0000
    Residual |  95.1908543        61  1.56050581   R-squared       =    0.2997
-------------+----------------------------------   Adj R-squared   =    0.2767
       Total |  135.922121        63  2.15749399   Root MSE        =    1.2492

------------------------------------------------------------------------------
        risk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    logmort0 |  -.5171894    .1408596    -3.67   0.001    -.7988555   -.2355232
    latitude |   2.007462    1.329874     1.51   0.136    -.6517858     4.66671
       _cons |   8.555776    .8082776    10.59   0.000     6.939525    10.17203
------------------------------------------------------------------------------
```

Figure 3: regression of risk against logmort0

   (c) Using the command **predict**, generate a variable riskhat containing the fitted values of this first stage.

   (d) Run the second stage regression. Comment on the results. Are the standard errors correct here?
   Correction: Not the standard error are wrong because they don't take into account that riskHat is a predicted value. Since stata can't really tell the difference between riskHat and really value

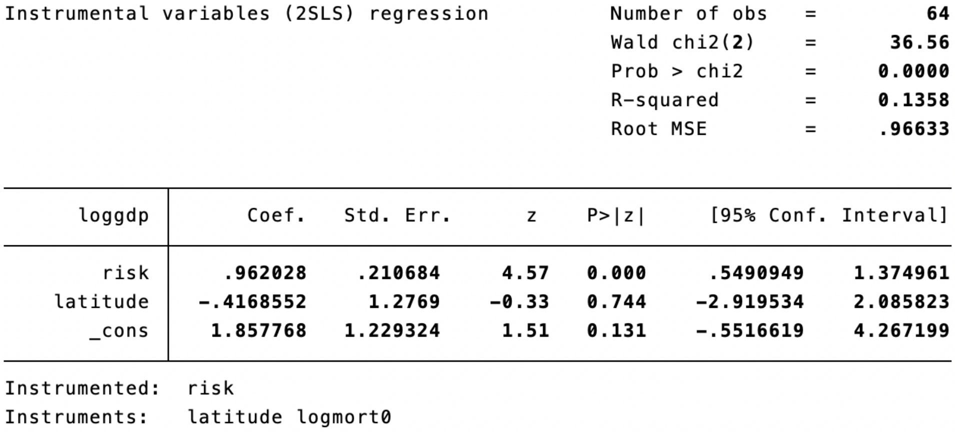


```
      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(2, 61)        =     29.74
       Model |  34.1413188         2  17.0706594   Prob > F        =    0.0000
    Residual |  35.0132241        61   .57398728   R-squared       =    0.4937
-------------+----------------------------------   Adj R-squared   =    0.4771
       Total |  69.1545429        63  1.09769116   Root MSE        =    .75762

------------------------------------------------------------------------------
      loggdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     riskHat |    .962028    .1651792     5.82   0.000     .6317318    1.292324
    latitude |  -.4168548    1.001107    -0.42   0.679    -2.418692    1.584983
       _cons |   1.857769    .9638066     1.93   0.059    -.0694816    3.785019
------------------------------------------------------------------------------
```

Figure 4: regression of loggdp against riskHat

(e) Run the IV regression using the **ivregress** command this time. Compare with your previous results.
Correction: Yeah, same point estimated but with corrent standard error this time. Here is code method of ivregress.



```
Instrumental variables (2SLS) regression          Number of obs     =         64
                                                  Wald chi2(2)      =      36.56
                                                  Prob > chi2       =     0.0000
                                                  R-squared         =     0.1358
                                                  Root MSE          =     .96633

------------------------------------------------------------------------------
      loggdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        risk |    .962028    .210684     4.57   0.000     .5490949    1.374961
    latitude |  -.4168552     1.2769    -0.33   0.744    -2.919534    2.085823
       _cons |   1.857768   1.229324     1.51   0.131    -.5516619    4.267199
------------------------------------------------------------------------------
Instrumented:  risk
Instruments:   latitude logmort0
```

Figure 5: IV regression

4. We can now turn to Albouy (2012)'s critic. Albouy's first critic concerns Acemoglu et al. (2001)' standard errors.

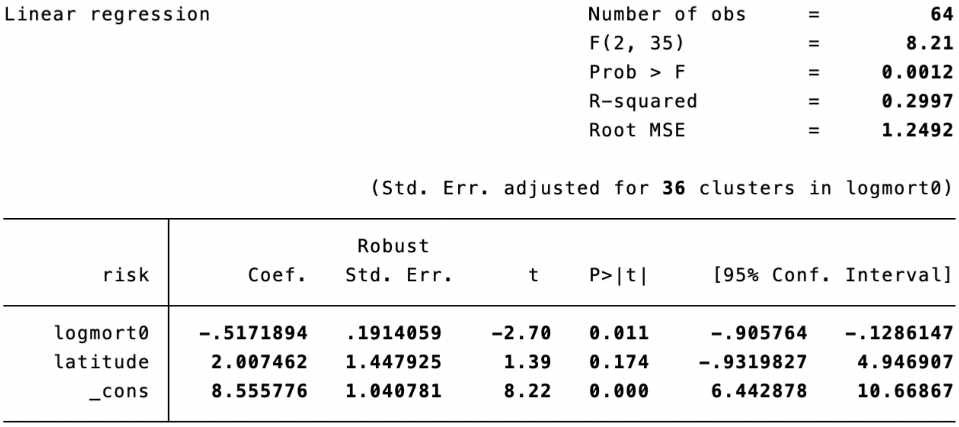
   (a) Albouy notices that our authors make conjectures about mortality rates for some countries. Notably, using mortality rate for some countries they extrapolate the mortality rate of neighbouring countries - in fact they even reuse the same rates sometimes. This violates one of the basic assumptions of OLS regression. Which one?

   Correction: This violates the iid assumption since several countries are not independent anymore.
   I.I.D means independent and identically distributed, a collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent.

   (b) To make up for this, it is possible to cluster data, that is to consider that different groups of data (say continents) are independent but the observations composing them (countries here) are not. Albouy also suggest to run an regression robust to heteroskedasticity. Do you know what this is? Can you give an example of what it corresponds to?

   Correction:
   Heteroskedasticity is when the variance of the error term varies between observations (this is common).
   An example: The variance of wage for a given education level is higher for higher levels of education than for lower levels.

(c) Rerun the first stage equation with standard errors (SE) that are robust (*i.e.* allowing for heteroskedasticity) and clustered at the mortality rate level. How does it change?



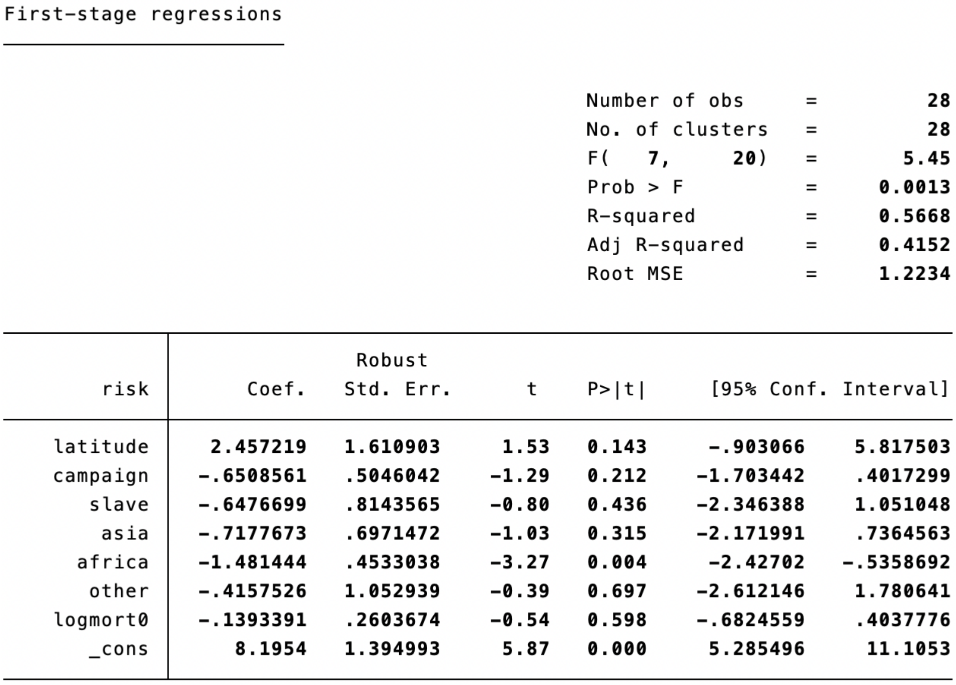
```
Linear regression                               Number of obs     =         64
                                                F(2, 35)          =       8.21
                                                Prob > F          =     0.0012
                                                R-squared         =     0.2997
                                                Root MSE          =     1.2492

                                   (Std. Err. adjusted for 36 clusters in logmort0)
------------------------------------------------------------------------------
             |               Robust
        risk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    logmort0 |  -.5171894   .1914059    -2.70   0.011     -.905764   -.1286147
    latitude |   2.007462   1.447925     1.39   0.174    -.9319827    4.946907
       _cons |   8.555776   1.040781     8.22   0.000     6.442878    10.66867
------------------------------------------------------------------------------
```

Figure 6: IV regression

5. Albouy's second critic concern the validity of the data for some countries. Not only is the data for some countries simply extrapolated from neighbouring countries but also all data sources are not comparable, some rates concerning soldiers living in barracks, some concerning soldiers during campaign and some concerning forced labor.

   (a) Retaining only countries for which the mortality rate is not extrapolated and including dummies for data sources (campaign and slave variables) and continents rerun a 2SLS regression. Comment on the first stage. What value do you find for the expropriation risk coefficient?

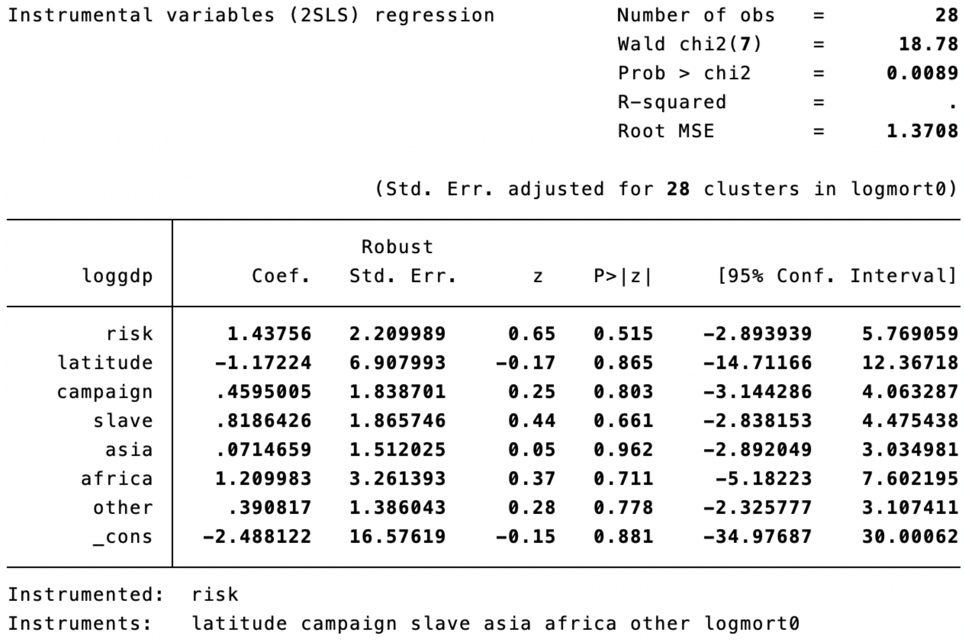


```
First-stage regressions
-----------------------

                                                Number of obs     =         28
                                                No. of clusters   =         28
                                                F(   7,      20)  =       5.45
                                                Prob > F          =     0.0013
                                                R-squared         =     0.5668
                                                Adj R-squared     =     0.4152
                                                Root MSE          =     1.2234

------------------------------------------------------------------------------
             |               Robust
        risk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    latitude |   2.457219   1.610903     1.53   0.143     -.903066    5.817503
    campaign |  -.6508561   .5046042    -1.29   0.212    -1.703442    .4017299
       slave |  -.6476699   .8143565    -0.80   0.436    -2.346388    1.051048
        asia |  -.7177673   .6971472    -1.03   0.315    -2.171991    .7364563
      africa |  -1.481444   .4533038    -3.27   0.004     -2.42702   -.5358692
       other |  -.4157526   1.052939    -0.39   0.697    -2.612146    1.780641
    logmort0 |  -.1393391   .2603674    -0.54   0.598    -.6824559    .4037776
       _cons |     8.1954   1.394993     5.87   0.000     5.285496     11.1053
------------------------------------------------------------------------------
```

Figure 7: IV regression



```
Instrumental variables (2SLS) regression          Number of obs     =         28
                                                  Wald chi2(7)      =      18.78
                                                  Prob > chi2       =     0.0089
                                                  R-squared         =          .
                                                  Root MSE          =     1.3708

                                (Std. Err. adjusted for 28 clusters in logmort0)
------------------------------------------------------------------------------
             |               Robust
      loggdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        risk |    1.43756   2.209989     0.65   0.515    -2.893939    5.769059
    latitude |   -1.17224   6.907993    -0.17   0.865    -14.71166    12.36718
    campaign |   .4595005   1.838701     0.25   0.803    -3.144286    4.063287
       slave |   .8186426   1.865746     0.44   0.661    -2.838153    4.475438
        asia |   .0714659   1.512025     0.05   0.962    -2.892049    3.034981
      africa |   1.209983   3.261393     0.37   0.711     -5.18223    7.602195
       other |    .390817   1.386043     0.28   0.778    -2.325777    3.107411
       _cons |  -2.488122   16.57619    -0.15   0.881    -34.97687    30.00062
------------------------------------------------------------------------------
Instrumented:  risk
Instruments:   latitude campaign slave asia africa other logmort0
```

Figure 8: IV regression

(b) Compute the GDP ratio of Mexico on the US. Now using the previous coefficient what would be the new value of this ratio were Mexico to have the same property right as the US? Comment on the result.

Correction:
Since we have the coefficient for the expropriation risk $= 1.44$ the expropriation risk for USA is 10 and for Mexico is 7.5, therefore the difference of risk between this two countries is 2.5, then the increase of log GDP is $1.44 * 2.5 = 3.6$
Current ratio is $e^{log(Mexico)-log(USA)} = 0.28$.
After Mexico has the same expropriation risk as USA, the the ration will be $e^{[log(Mexico)+3.6]-log(USA)} = 10.18$, which implies that the GDP of Mexico could be 10 times of USA.

Bibliography

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401.

Albouy, D. Y. (2012). The colonial origins of comparative development: an empirical investigation: comment. *American economic review*, 102(6):3059–76.