Small Samples and Low Quality Data
Some Methodological Issues

Peter Kugler
WWZ/Universität Basel

10th Annual NBP-SNB Joint Seminar on Monetary Transmission Mechanism in Transition Countries, Zürich June 2-4 2013

Outline

- Introduction
- Pooling of time series and cross section data
- Asymptotic distributions with small samples and non-normal distributions: bootstrapping
- Outliers: robust estimation and qualitative variables
- Conclusion

Introduction

- Econometric analyses for transition countries are often confronted with a lack of long and reliable data series.
- Optimal estimation for most standard econometric models is based on normal distribution and we only know the asymptotic distribution of the estimates we get in most applications.
- Short time series and non-normal distributions (outliers) we often see with transition economy data question the reliability of the estimation and test results.
- Three approaches to improve the quality of econometric analysis:
  - Pooling of time series and cross section data
  - Bootstrapping
  - Robust estimation and qualitative variables

Pooling of time series and cross section data

- Pooling of cross section units (countries, industries, firms) with time series data should result in more precise estimates.
- This is even the case if do not impose the strict panel assumption of constant slope coefficient over all cross section units.
- In order to illustrate the option we have let us consider a simple dynamic model fort two cross section units:

$$y_{1t} = \alpha_1 + \beta_1 x_{1t} + \lambda_1 y_{1t-1} + \varepsilon_{1t}$$
$$y_{2t} = \alpha_2 + \beta_2 x_{2t} + \lambda_2 y_{2t-1} + \varepsilon_{2t}$$
$$t = 1,2,\ldots T$$

- Even if we assume that all the regression coefficients are different across the two units we gain efficiency in joint estimation (Seemingly Unrelated Regression) if the two error terms are correlated. This condition is often fulfilled.
- If we reasonably can adopt the panel assumption that the slope coefficient are the same in both regression we get an additional gain in estimation efficiency.

- However, we do not need to make one of these two extreme assumptions: we may only restrict some of the slope coefficients to be equal across equations.
- In our context the restriction of the same long run effect of *x* on *y,*

$$\gamma = \beta_i /(1-\lambda_i) \Rightarrow \beta_i = \gamma(1-\lambda_i),\ i = 1,2$$

may be a reasonable assumption. To this end we re-formulate the model

$$y_{1t} = \alpha_1 + \gamma(1-\lambda_1)x_{1t} + \lambda_1 y_{1t-1} + \varepsilon_{1t}$$
$$y_{2t} = \alpha_2 + \gamma(1-\lambda_2)x_{2t} + \lambda_2 y_{2t-1} + \varepsilon_{2t}$$
$$t = 1,2,\ldots T$$

and estimate only a reduced number of parameters in a model allowing different short run dynamics.

- Note that such a system with cross equation restrictions is easily estimated using standard software packages as EVIEWS. Moreover, we can test the appropriateness of our restriction by applying a Wald test to the unrestricted system.

- The same approach is feasible for an Error-Correction model:

$$\Delta y_{1t} = \alpha_1 + \lambda_1 (y_{1t-1} - \gamma x_{1t-1}) + \varepsilon_{1t}$$
$$\Delta y_{2t} = \alpha_2 + \lambda_2 (y_{1t-1} - \gamma x_{2t-1}) + \varepsilon_{2t}$$
$$t = 1,2,\ldots T$$

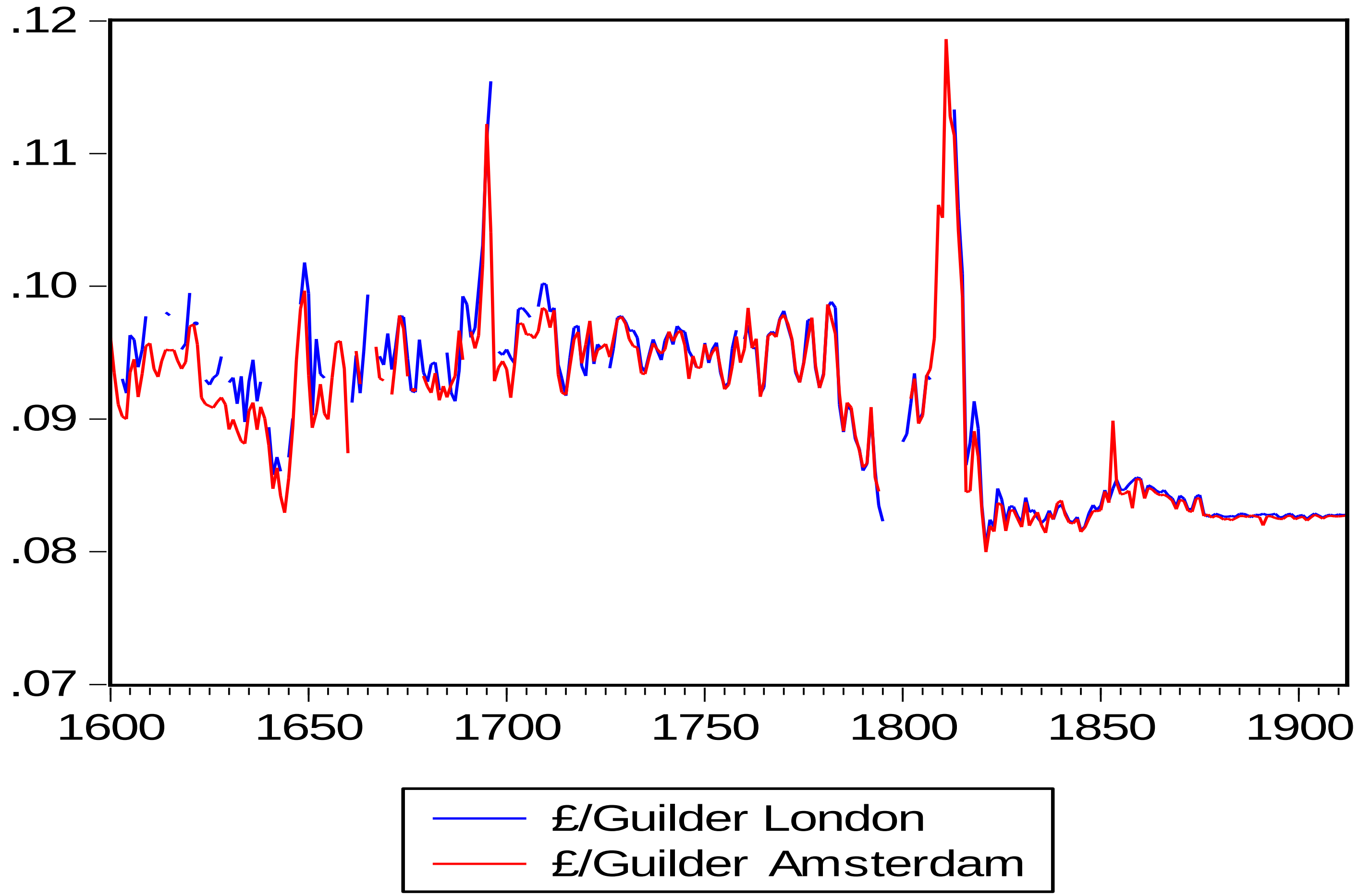
- This approach results in larger efficiency gains when it is applied to more than two reasonably similar cross section units.
- This framework is easily extended to models with more than one lag.

Distributions of estimates with small samples and non-normal distributions: bootstrapping

- The reliability of asymptotic distribution of estimates obtained with small samples and non-normal data is questionable.
- Bootstrapping allows to explore the properties of estimated parameters using the actual distribution of the model residuals, i.e. we do not assume a given form of the distribution as with Monte Carlo replications.
- To this end the estimated model residual are re-sample (with replacement) and artificial time series are created for the model’s endogenous variables. These are used to estimate the model and the replication of this procedure generates an empirical distribution of the parameter estimates.
- The implementation of this approach is illustrated for a simple two equations EC model for historical exchange rate data in EVIEWS.

Figure 1: £/Dutch Guilder Exchange Rate in Amsterdam and London 1600 1912, Guilder: silver (1600-1838), bimetallic (1839-1875), gold (1875-1914)

£: bimetallic (1600-1717), bimetallic (1718-97), paper (1797-1819), gold (1819-1914)



Estimated system (SYS01) and bootstrap system (SYS02)

- SYS01:

dlog(pdglo) = c(1)+c(2)*(log(pdglo(-1)/pdgam(-1)))

dlog(pdgam) = c(3)+c(4)*(log(pdglo(-1)/pdgam(-1)))

1600 - 1700 dlog(pdglo) = 0.019 - 0.821*(log(pdgam(-1)/pdglo(-1)))

(0.0048) (0.199)

$$R^2 = 0.233 \ \ \mathrm{DW} = 1.96$$

dlog(pdgam) = 0.0051 – 0.135*(log(pdgam(-1)/pdglo(-1)))

(0.0055) (0.203)

$$R^2 = 0.233 \ \ \mathrm{DW} = 1.96$$

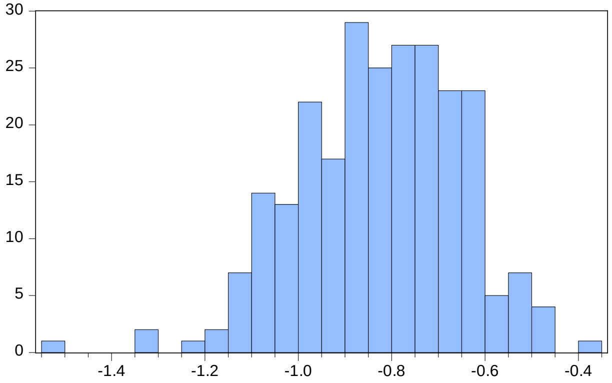
- SYS02:

dx = c(11)+c(12)*(x(-1)-y(-1))

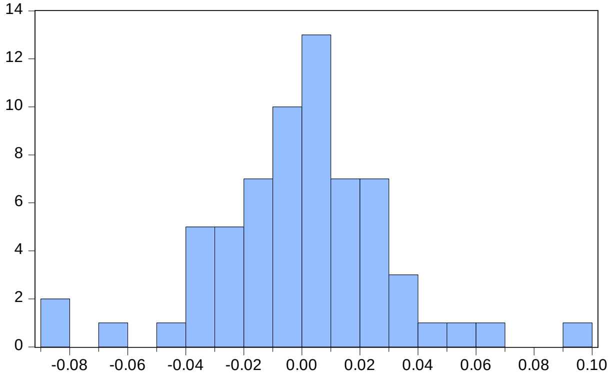
dy = c(13)+c(14)*(x(-1)-y(-1))

```
vector(250) c1
vector(250) c2
vector(250) c3
vector(250) c4
smpl 1590 1912
genr x = lpdglo
genr y = lpdgam
smpl 1600 1700
sys01.ls
sys01.makeresids e1 e2
group g1 e1 e2
for !id=1 to 250
g1.resample(outsmpl="1600 1700",dropna)
for !i=1 to 58
smpl 1600+!i 1600+!i
genr dx =c(1)+c(2)*(x(-1)-y(-1))+ E1_B
genr dy=c(3)+c(4)*(x(-1)-y(-1))+ E2_B
genr x = x(-1) + dx
genr y = y(-1) + dy
next
smpl 1600 1600+58
sys02.ls
c1(!id) = c(11)
c2(!id) = c(12)
c3(!id) = c(13)
c4(!id) = c(14)
next
smpl 1590 1840
mtos(c1,c1s)
show @quantile(c1s,0.05)
show @quantile(c1s,0.95)
mtos(c2,c2s)
show @quantile(c2s,0.05)
show @quantile(c2s,0.95)
mtos(c3,c3s)
show @quantile(c3s,0.05)
show @quantile(c3s,0.95)
mtos(c4,c4s)
show @quantile(c4s,0.05)
show @quantile(c4s,0.95)
```

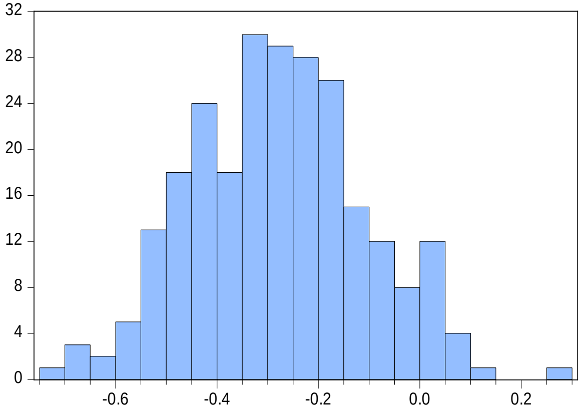
Empirical distribution of residuals and bootstrap estimates



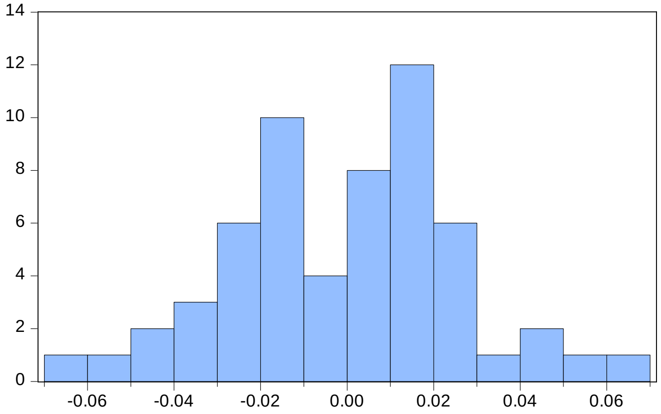
```
Series: C2S
Sample 1590 1840
Observations 250

Mean          -0.826512
Median        -0.820842
Maximum       -0.391398
Minimum       -1.502977
Std. Dev.      0.174234
Skewness      -0.339359
Kurtosis       3.324714

Jarque-Bera    5.896861
Probability    0.052422
```



```
Series: E1
Sample 1590 1840
Observations 65

Mean          -5.34e-19
Median         0.003451
Maximum        0.098383
Minimum       -0.086661
Std. Dev.      0.030905
Skewness      -0.134881
Kurtosis       4.618930

Jarque-Bera    7.295455
Probability    0.026050
```



```
Series: C4S
Sample 1590 1840
Observations 250

Mean          -0.277133
Median        -0.274771
Maximum        0.282858
Minimum       -0.728939
Std. Dev.      0.170992
Skewness       0.120759
Kurtosis       2.900379

Jarque-Bera    0.710993
Probability    0.700826
```



```
Series: E2
Sample 1590 1840
Observations 58

Mean          -3.41e-18
Median         0.001825
Maximum        0.066784
Minimum       -0.062903
Std. Dev.      0.026025
Skewness      -0.020990
Kurtosis       2.994697

Jarque-Bera    0.004327
Probability    0.997839
```

- Bootstrap distribution of the two EC-coefficients are indicated to be normal or close to normal and the parameters of the bootstrap distribution are close to the asymptotic OLS-estimates: asymptotic approximation is ok.
- However, if we estimate a lot of parameters with a small number of observations we expect large differences between bootstrap and asymptotic distributions.
- Bootstrap replications are often used when no asymtotic distribution is available, for instance if some parameters are estimated by grid search.

Outliers: robust estimation and qualitative variables

- Outliers have a strong influence on least squares based estimation procedures which have their strongest justification with normal distributions.
- Non-normal distributions (in particular fat tail distributions) may lead to a poor performance of least squares based procedures.
- As an alternative robust procedures can be used. A basic method is “Least Absolute Deviation” implemented in EVIEWS which corresponds to the calculation of the median for a single series of observations. This procedure is applied to our first EC equation and we obtain nearly the same coefficient estimates as by LS which points to no severe outlier problems.
- An other way to deal with strong outliers consist of using a qualitative but ordered variable (for instance three categories as “high”, “moderate” and “low”) in an ordered choice framework (Probit, Logit), An example involving hyperinflation data is taken from Bernholz/Kugler (German Economic Review, 10(2), 2009, 165-175,

LAD estimates of EC model

```
Dependent Variable: DLOG(PDGLO)
Method: Quantile Regression (Median)
Sample (adjusted): 1600 1690
Included observations: 51 after adjustments
Estimation successfully identifies unique optimal solution

DLOG(PDGLO) =C(1)+ C(2)*(LOG(PDGLO(-1)/PDGAM(-1)))
            Coefficient Std. Error  t-Statistic  Prob.
C(1)        0.019871    0.007598    2.615385     0.0118
C(2)        -0.807080   0.281948    -2.862510    0.0062

Pseudo R-squared      0.122498    Mean dependent var                     0.001933
Adjusted R-squared    0.104590    S.D. dependent var                     0.029159
S.E. of regression    0.025366    Objective                              0.508848
Quantile dependent var            0.001696     Restr. objective          0.579882
 Quasi-LR statistic   7.674211                 Prob(Quasi-LR stat)       0.005602
```

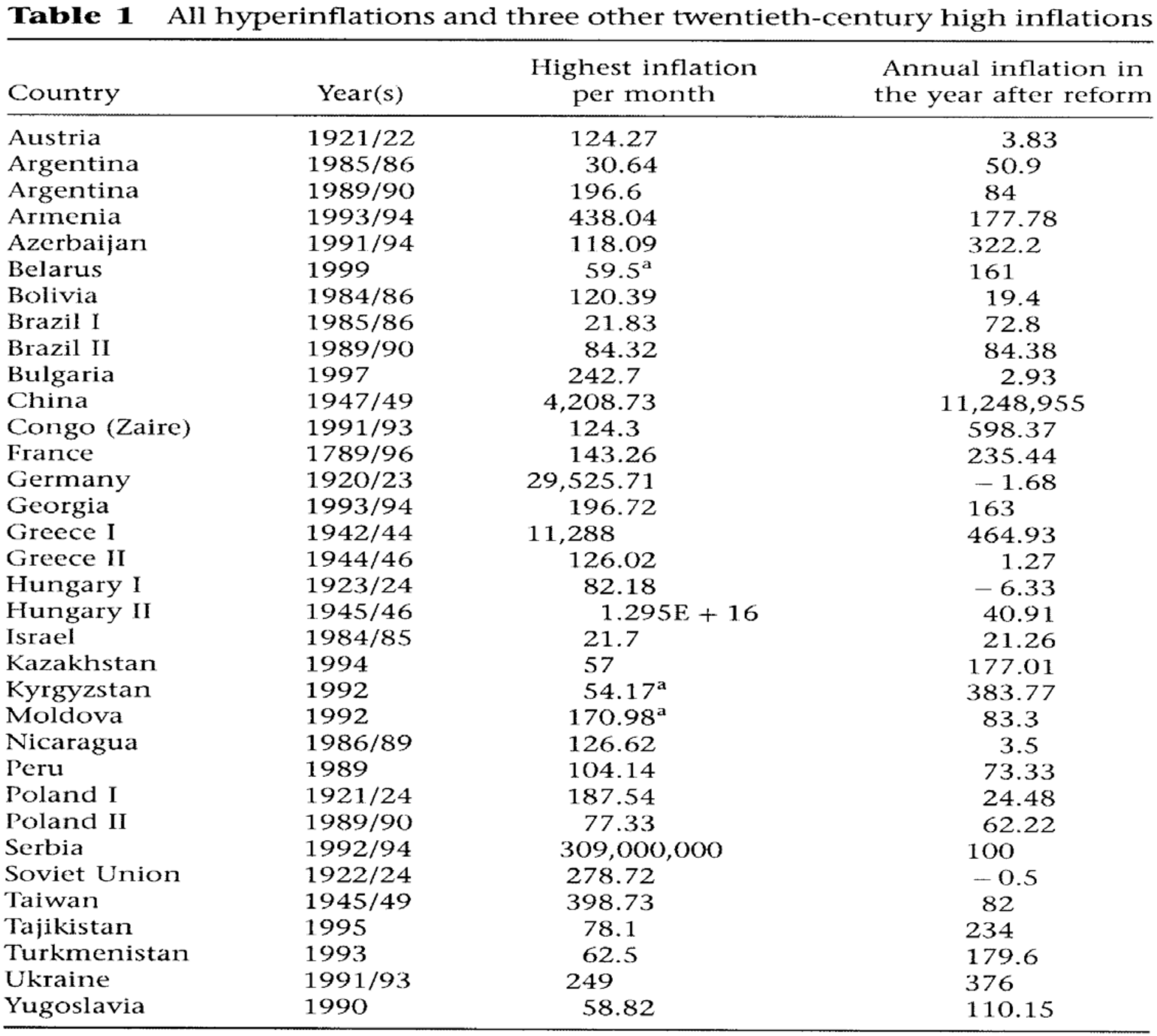


Table 1 All hyperinflations and three other twentieth-century high inflations

Country	Year(s)	Highest inflation per month	Annual inflation in the year after reform
Austria	1921/22	124.27	3.83
Argentina	1985/86	30.64	50.9
Argentina	1989/90	196.6	84
Armenia	1993/94	438.04	177.78
Azerbaijan	1991/94	118.09	322.2
Belarus	1999	59.5[a]	161
Bolivia	1984/86	120.39	19.4
Brazil I	1985/86	21.83	72.8
Brazil II	1989/90	84.32	84.38
Bulgaria	1997	242.7	2.93
China	1947/49	4,208.73	11,248,955
Congo (Zaire)	1991/93	124.3	598.37
France	1789/96	143.26	235.44
Germany	1920/23	29,525.71	– 1.68
Georgia	1993/94	196.72	163
Greece I	1942/44	11,288	464.93
Greece II	1944/46	126.02	1.27
Hungary I	1923/24	82.18	– 6.33
Hungary II	1945/46	1.295E + 16	40.91
Israel	1984/85	21.7	21.26
Kazakhstan	1994	57	177.01
Kyrgyzstan	1992	54.17[a]	383.77
Moldova	1992	170.98[a]	83.3
Nicaragua	1986/89	126.62	3.5
Peru	1989	104.14	73.33
Poland I	1921/24	187.54	24.48
Poland II	1989/90	77.33	62.22
Serbia	1992/94	309,000,000	100
Soviet Union	1922/24	278.72	– 0.5
Taiwan	1945/49	398.73	82
Tajikistan	1995	78.1	234
Turkmenistan	1993	62.5	179.6
Ukraine	1991/93	249	376
Yugoslavia	1990	58.82	110.15

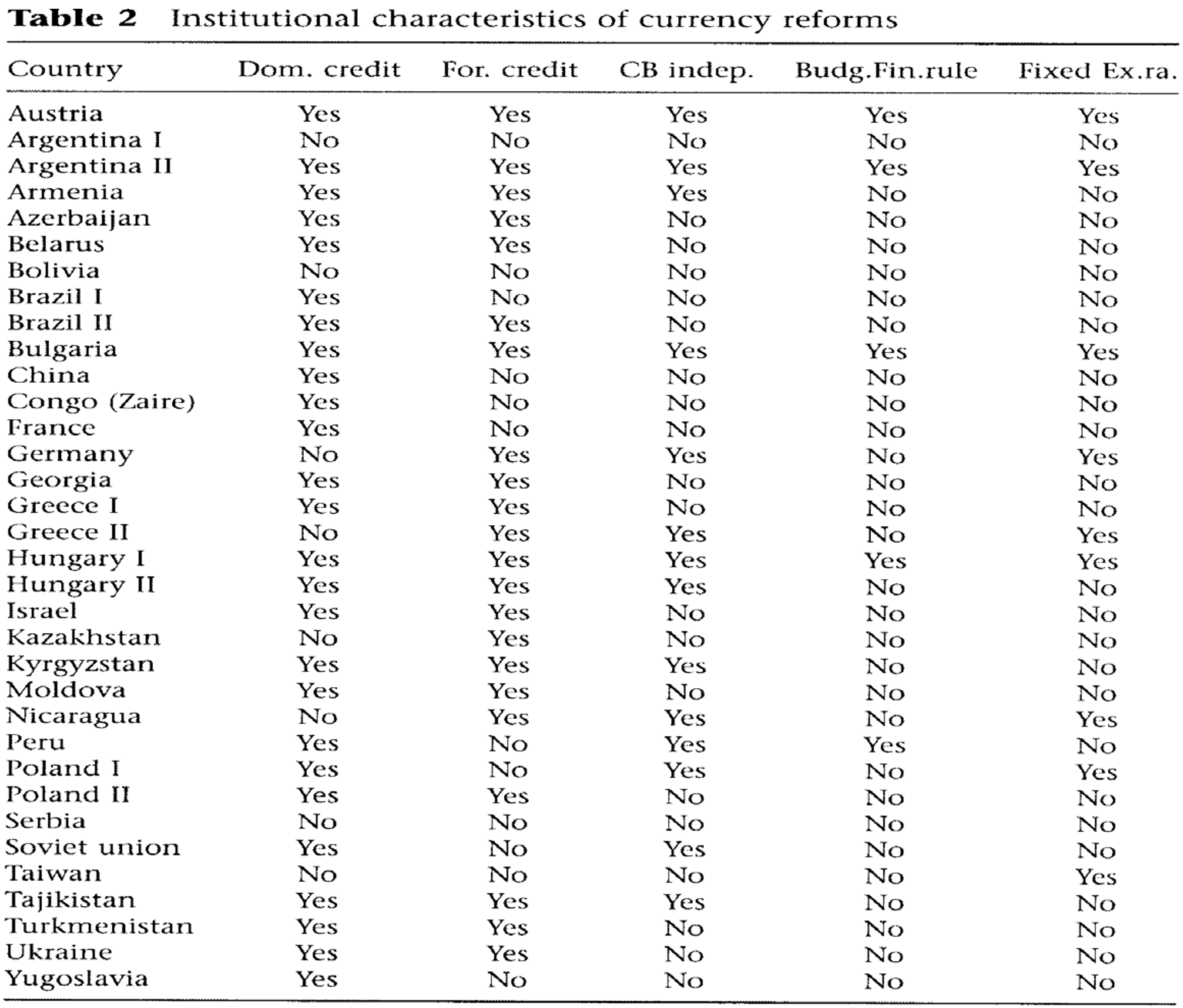


Table 2 Institutional characteristics of currency reforms

Country	Dom. credit	For. credit	CB indep.	Budg.Fin.rule	Fixed Ex.ra.
Austria	Yes	Yes	Yes	Yes	Yes
Argentina I	No	No	No	No	No
Argentina II	Yes	Yes	Yes	Yes	Yes
Armenia	Yes	Yes	Yes	No	No
Azerbaijan	Yes	Yes	No	No	No
Belarus	Yes	Yes	No	No	No
Bolivia	No	No	No	No	No
Brazil I	Yes	No	No	No	No
Brazil II	Yes	Yes	No	No	No
Bulgaria	Yes	Yes	Yes	Yes	Yes
China	Yes	No	No	No	No
Congo (Zaire)	Yes	No	No	No	No
France	Yes	No	No	No	No
Germany	No	Yes	Yes	No	Yes
Georgia	Yes	Yes	No	No	No
Greece I	Yes	Yes	No	No	No
Greece II	No	Yes	Yes	No	Yes
Hungary I	Yes	Yes	Yes	Yes	Yes
Hungary II	Yes	Yes	Yes	No	No
Israel	Yes	Yes	No	No	No
Kazakhstan	No	Yes	No	No	No
Kyrgyzstan	Yes	Yes	Yes	No	No
Moldova	Yes	Yes	No	No	No
Nicaragua	No	Yes	Yes	No	Yes
Peru	Yes	No	Yes	Yes	No
Poland I	Yes	No	Yes	No	Yes
Poland II	Yes	Yes	No	No	No
Serbia	No	No	No	No	No
Soviet union	Yes	No	Yes	No	No
Taiwan	No	No	No	No	Yes
Tajikistan	Yes	Yes	Yes	No	No
Turkmenistan	Yes	Yes	No	No	No
Ukraine	Yes	Yes	No	No	No
Yugoslavia	Yes	No	No	No	No

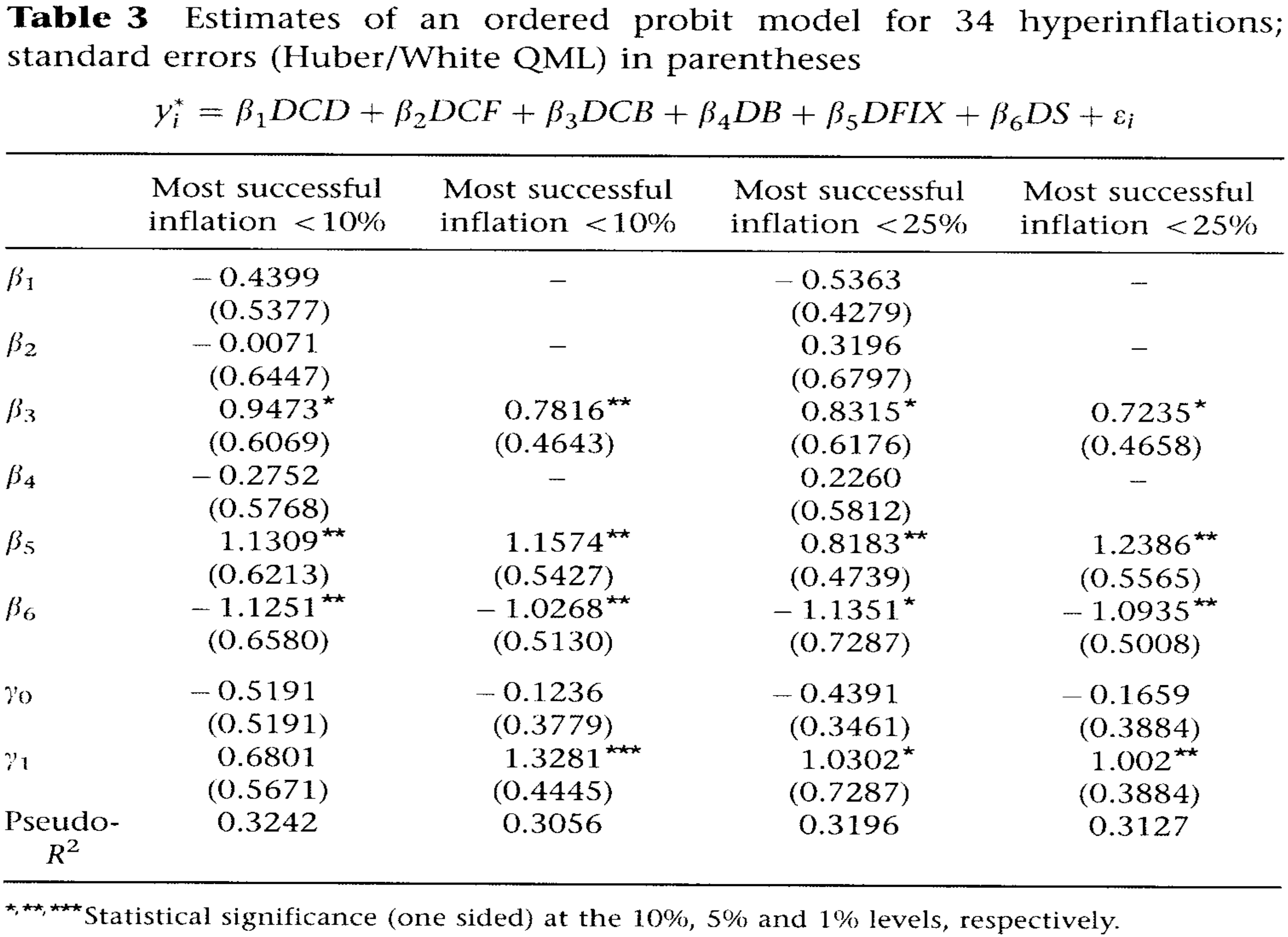


Table 3 Estimates of an ordered probit model for 34 hyperinflations; standard errors (Huber/White QML) in parentheses

$$y_i^* = \beta_1 DCD + \beta_2 DCF + \beta_3 DCB + \beta_4 DB + \beta_5 DFIX + \beta_6 DS + \varepsilon_i$$

	Most successful inflation <10%	Most successful inflation <10%	Most successful inflation <25%	Most successful inflation <25%
β_1	– 0.4399	–	– 0.5363	–
	(0.5377)		(0.4279)	
β_2	– 0.0071	–	0.3196	–
	(0.6447)		(0.6797)	
β_3	0.9473*	0.7816**	0.8315*	0.7235*
	(0.6069)	(0.4643)	(0.6176)	(0.4658)
β_4	– 0.2752	–	0.2260	–
	(0.5768)		(0.5812)	
β_5	1.1309**	1.1574**	0.8183**	1.2386**
	(0.6213)	(0.5427)	(0.4739)	(0.5565)
β_6	– 1.1251**	– 1.0268**	– 1.1351*	– 1.0935**
	(0.6580)	(0.5130)	(0.7287)	(0.5008)
γ_0	– 0.5191	– 0.1236	– 0.4391	– 0.1659
	(0.5191)	(0.3779)	(0.3461)	(0.3884)
γ_1	0.6801	1.3281***	1.0302*	1.002**
	(0.5671)	(0.4445)	(0.7287)	(0.3884)
Pseudo-R^2	0.3242	0.3056	0.3196	0.3127

*,**,***Statistical significance (one sided) at the 10%, 5% and 1% levels, respectively.

Conclusion

- The reliability of econometric results based on short time series often characterized by non-normality may be improved by three approaches :
  - Pooling of time series and cross section data
  - Bootstrapping
  - Robust estimation and qualitative variables
- These approaches are easily implemented with standard econometric packages as EVIEWS.
- More elaborate methods as the Bayesian approach combining a priory information with data are of course interesting in our context but the application needs more technical expertise as the simple approaches we considered.