Multi-level modelling of ‘country effects’: a cautionary tale

Mark Bryan (ISER) and

Stephen P. Jenkins (LSE, ISER, IZA)

Research undertaken for and part-supported by the Analysis of Life Chances in Europe (ALICE) project, funded by the UK Economic and Social Research Council (grant RES-062-23-1455). We also acknowledge support from the ESRC (grant RES-518-28-001) and the University of Essex through the Research Centre on Micro-Social Change

How robust are the estimates from a regression with 25 observations?

– Are parameter estimates biased?
– Are reported SEs and CIs reliable?

The background

- There is much regression-based analysis of harmonised individual-level data from multiple countries
  - *Multilevel (a.k.a. hierarchical or mixed) models*
  - Linear and non-linear (binary logit) models
  - Outcome modelled as function of individual-level and country-level variables (including unobserved country-level variables)
- Many social science researchers aim to quantify 'country effects' a.k.a. 'contextual effects':
  - *regression coefficients on level-2 (country-level) predictors*: extent to which differences in outcomes reflect differences in country-specific features of demographic structure, labour markets, tax-benefit systems etc, as distinct from the differences in outcomes associated with variations in characteristics of individuals
  - *level-2 variances (and Intra-Cluster Correlation, ICC*): importance of 'country effects' also summarised in terms of variance of unobserved country-level factors (relative to the variance of unobserved individual-level factors)

Many multi-country datasets, much-used: small # countries, large # respondents/country

Data sources (in alphabetical order)	Number of countries per round (approx.)
Eurobarometer	27
European Community Household Panel (ECHP)	15
European Quality of Life Survey (EQLS)	31
European Social Survey (ESS)	30
EU Statistics on Income and Living Conditions (EU-SILC)	27
European Values Study (EVS)	45
International Social Survey Program (ISSP)	36
Luxembourg Income Study (LIS)	32
Survey of Health, Ageing and Retirement in Europe (SHARE)	14

Notes: All datasets are based on cross-sectional surveys with the exception of ECHP and SHARE which are panel surveys.

NB Number of countries used in empirical studies is often smaller than the maximum possible

Many publications on wide range of topics using multi-country datasets

- Topics range from labour force participation and wages, to political and civic participation rates, and social and political attitudes:
- Many published papers:
  - Of 340 articles published in *European Sociological Review* between 2005 and 2012, 75 used regression-based analysis of multi-country data, of which 43 use multilevel modelling methods (13% of all published articles)
  - Significant number also in *Journal of European Social Policy* (14/111 between 2005 and 2009)
  - And, of course, publications elsewhere as well
- Our project's question: are estimates of country effects reliable given the nature of the datasets?
  - Many applied social science researchers appear unaware of the issue …

Project Output #1: Bryan and Jenkins ‘Regression analysis of country effects using multilevel data: a cautionary tale’

ISER Working Paper 2013-14

https://www.iser.essex.ac.uk/publications/working-papers/iser/2013-14

- Review of the several regression approaches used to analyse multi-country data
- Review of existing evidence, including Monte-Carlo simulation studies, of the ‘small number of countries’ issue
- New Monte-Carlo analysis of linear and binary logit mixed models with 2 specifications for each:
  - Basic: random country intercept and a country-level predictor;
  - Extended: as (i), plus 2 random slopes and cross-level interaction

This talk focuses on our analysis of MLMs

[See WP for gory details and other stuff]

Headline messages

- MLMs are not the only way to analyse multi-country data
  - We review MLM and other approaches
- Social science MLM user community should be more aware of potential problems of inferring country effects when there is only a small number of countries in the data set
- What is a ‘small’ number of countries?
  - Close to or larger than the number in the analysis datasets that many researchers have used when fitting MLMs!

Prototypic linear model

$$y_{ic} = \boldsymbol{X}_{ic}\boldsymbol{\beta} + \boldsymbol{Z}_{c}\boldsymbol{\gamma} + u_c + \varepsilon_{ic}$$

- Individuals within countries: $i = 1, \ldots, N_c$ (N_c large)
- Countries: $c = 1, \ldots, C$ (C small, often c. 25)

- y_{ic} : individual-level outcome (e.g. 'well-being')
- $\boldsymbol{X}_{ic}$: individual-level characteristics (e.g. age, sex, education, or marital status)
- $\boldsymbol{Z}_c$: country-level features such as socio-economic institutions or labour markets (e.g. public health expenditure)
- u_c : unobserved country effects shared by all people within the same country (e.g. unmeasured quality of public services)
- ε_{ic}: unobserved individual effects
- $\varepsilon_{ic} \sim N(0, \sigma_\varepsilon^2)$ and $u_c \sim N(0, \sigma_u^2)$, assumed to be uncorrelated with each other, and with $\boldsymbol{X}_{ic}$ and $\boldsymbol{Z}_c$
- Model with "random effects" ("random intercept"); "Multi-Level Model"

Four regression modelling approaches commonly applied to multilevel country datasets

	Approach	Remarks about how country effects are specified
1.	Common model for all countries, pooled data, country-specific clustered standard errors	Country effects controlled for, not modelled (number of clusters shouldn't be small)
2.	Separate model fitted to the data for each country	Country effects not separately identified (absorbed into the intercept of each country's model)
3.	Common model applied to pooled data (as in approach 1), except that model has country fixed effects	All country-level factors are absorbed into the country fixed effect; estimates refer to specific sample of countries
4.	Common model applied to pooled data (as in approach 1), except that model has country random effects (multilevel model)	Country effects can be specified in terms of a country error variance and fixed effects of country-level predictors; 'exchangeable' estimates

- Researchers primarily interested in β favour approaches 1–3
- Researchers interested in γ and σ_u^2 favour approach 4 (MLM)
- Bayesian MLM approaches not discussed in detail here

A fifth **two-step approach** is also instructive and highlights the ‘small number of countries’ issue

- Many authors have noted or rediscovered that two-level models can be estimated using a two-step method
  - Hanushek (1974), Saxonhouse (1976), Card (1995), papers in 2005 *Political Analysis* special issue, Gelman and Hill (2007), Donald and Lang (2007)
- A two-step approach illustrates the issues clearly, and can also be a viable estimating strategy
  - Under certain conditions, it is identical to GLS
- To see the intuition, re-write the prototypic model in terms of two equations, one for each level (as in some MLM literature and in HLM software):
  - Level 1: $y_{ic} = \boldsymbol{X}_{ic}\boldsymbol{\beta} + v_c + \varepsilon_{ic}$
  - Level 2: $v_c = \boldsymbol{Z}_c\gamma + u_c$

Two-step approach: properties of estimators

1. Estimate Level 1 equation by OLS using fixed (sic) country intercepts v_c (country indicators):

   $y_{ic} = \boldsymbol{X}_{ic}\boldsymbol{\beta} + v_c + \varepsilon_{ic}$ (regression with large # obs)

2. Estimate Level 2 equation by, e.g. GLS, using estimates of country intercepts (coefficients on country indicators) as the dependent variable:

   $\hat{v}_c = \alpha + Z_c\boldsymbol{\gamma} + \eta_c$ (regression with small # obs)

- If Step 2 is estimated by feasible GLS, estimates of $\boldsymbol{\gamma}$ are ‘numerically identical’ to FGLS estimates of the complete model in one step (Donald and Lang 2007)

Lessons from a two-step approach (1)

- Step 1: large N_c obs per country → good estimates of fixed coefficients $\boldsymbol{\beta}$ on individual characteristics and country intercepts v_c (but NB v_c includes $\boldsymbol{Z}_c$)
- Step 2: a regression with only small number C obs (e.g. 25) → estimates of country effects (fixed coefficients $\boldsymbol{\gamma}$ on country-level variables and intercept variance σ_u^2) may be of doubtful reliability
- Knowing the distribution of country intercepts (η_c) is vital for calculating confidence intervals and doing hypothesis tests
  - Cannot rely on having 'large' C
- Even if we can credibly assume country intercepts are normally distributed (standard in MLM literature), we need to use t not z statistics due to small C (Donald and Lang 2007)
  - If we naïvely use z stats from software output, will find too many significant coefficients and CIs will be too narrow

Lessons from a two-step approach (2)

- Can be extended to accommodate random $\boldsymbol{\beta}$ coefficients
- Can also be applied to non-linear models (logit, probit, etc.)

- The two-step method leads naturally to exploratory data analysis: graphical summary of country-level variations in outcomes in which one plots the country intercepts fitted at step 1 against elements of $\boldsymbol{Z}_c$
  - Kedar and Shively (2005), Bowers and Drake (2005)
  - "Showing what is in the data and not more"

How many countries required for reliable estimates of country effects? Lit. review

- If *C* and N_C both ‘large’, parameter estimates have ‘nice’ properties (asymptotic results):
  - Consistent (converge to true values) and normally distributed
- When *C not* ‘large’, there are some statements but apparently not well known in applied social science:
  - Small # level-2 units ⇒ imprecise estimates of level-2 variance and likely biased downwards (Hox 2010, Raudenbush and Bryk 2002)
  - Estimates of fixed parameters also affected by uncertainty in variance estimates: SEs biased downwards and distribution of test statistics unknown (Raudenbush and Bryk 2002)

Little *concrete* guidance about the number of groups required to avoid problems

- Centre for Multilevel Modelling website: “Rules of thumb such as only doing multilevel modelling with 15 or 30 or 50 level 2 units can be found and are often personal opinions based on personal experience and varying reasons e.g. getting a non zero variance, being able to check the normality assumption etc.” http://www.bristol.ac.uk/cmm/learning/multilevel-models/samples.html
- Most MLM textbooks mention the issue and sometimes cite rules of thumb, recommending anywhere between 10 and 50 groups as a minimum
  - But stress that the minimum number depends on (a) application-specific factors like the number of group-level predictors (Raudenbush and Bryk 2002) and (b) whether interest is focussed on the coefficients on the fixed regression predictors or the parameters describing the distribution of the random effects (Hox 2010)
- Advice about sample size often bound up with considerations of the cost of primary data collection (e.g. Snijders and Bosker 1999)
  - Irrelevant for multilevel country datasets already in existence

How large does *C* need to be? Guidance from previous Monte-Carlo analysis

- Most evidence refers to basic linear model (as above)
- Estimates of fixed parameters (β and γ) unbiased even if *C* is as small as 10 (Hox 2010, Maas & Hox 2004, 2005):
- Level-2 (country) variance, σ_u^2, increasingly under-estimated as *C* gets smaller
- SEs for both β and γ and especially σ_u^2 biased downwards with small *C*

Maas and Hox (2004) rules of thumb: $C \geq 10$ for unbiased estimates of β and γ ; $C \geq 50$ for accurate SE estimates, especially of level-2 variances

- The few studies of non-linear (binary logit/probit) models suggest similar conclusions to those for linear models
  - Stegmueller (2013), Moineddin (2007)

How large does *C* need to be? The case for more Monte-Carlo analysis

Previous analysis does not necessarily translate to typical multi-country dataset applications

- Considered different data structures (education and health research) in which moderate numbers of both level-1 and level-2 observations
- And/or don’t consider values of *C* sufficiently small
  - E.g. Maas and Hox (2004, 2005): $C = 30, 50, 100$ and $N_C = 5, 30, 50$
- Focus on linear models mostly
- Focus on models typically including only 2 or 3 regressors, and assumed standard normal

How large does *C* need to be? More Monte-Carlo analysis: this paper

- Linear and non-linear models in 2 versions
  - Basic: random intercept model
  - Extended: adds two random coefficients (level-2 variances) and a cross-level interaction
- Systematic examination of variation in *C*:
  - $C = 5(5)50\ 100$, with $N_C = 1000$
- Regressors: continuous, binary, and categorical
- Model specifications inspired by a real-life multi-country dataset application
  - EU-SILC 2007 data for 26 countries
  - Women aged 18–64
  - Two outcome variables:
    - Hours or work (linear model) and Labour force participation (binary logit model)

DGPs reflect EU-SILC illustration: basic models

Regressors	Linear model (‘hours’)		Logit model (‘participation’)	
	Parameter value	Mean of regressor	Parameter value	Mean of regressor
Fixed effects				
$constant$	22	1	–9.1	1
age_{ic}	0.8	41.6	0.5	41.0
$(age_{ic})^2$	–0.01	1832.5	–0.006	1862.4
$cohab_{ic}$	–1	0.725	0.02	0.658
$nownch_{ic}$	–1.2	1.110	–0.27	0.911
$isced3_{ic}$	0.7	0.446	0.7	0.449
$isced4_{ic}$	1.4	0.058	0.9	0.052
$isced56_{ic}$	1.6	0.328	1.4	0.243
$chexp_c$	–0.23	0.535	0.98	0.586
Random effects				
σ_e *(sig_e)*	9.5		$\pi/\sqrt{3}$	
σ_u *(sig_u)*	3.5		0.275	
ICC	0.120		0.022	

The random effects are: an individual-specific error $e_{ic} \sim N(0, \sigma_e^2)$; a random country intercept $u_c \sim N(0, \sigma_u^2)$. The ICC is implied by the error variance values. The country-level regressor is $chexp_c$. The omitted ‘education’ category is $isced12_{ic}$. The mean value of the outcome is 35.7 in the linear Model, 0.78 in the logit model. All means refer to the dataset associated with the case in which $C = 25$.

Some further details re DGPs:

- *Model parameters* derived from preliminary estimates of linear and binary logit models derived from EU-SILC 2007 data
- *Joint distribution(s) of the regressors* derived using a cell-based approach
  - Combinations of regressors define cells; Pr(individual in cell) derived from empirical frequency distribution in EU-SILC estimation samples
  - Age distribution fitted as Singh-Maddala for model (i), and uniform for model (ii) in EU-SILC data. Parameters used to generate age values that were then grouped into 5 classes in order to construct the cells
- DGP is same for each model examined; MC design varies *C*

Some further details re MC analysis

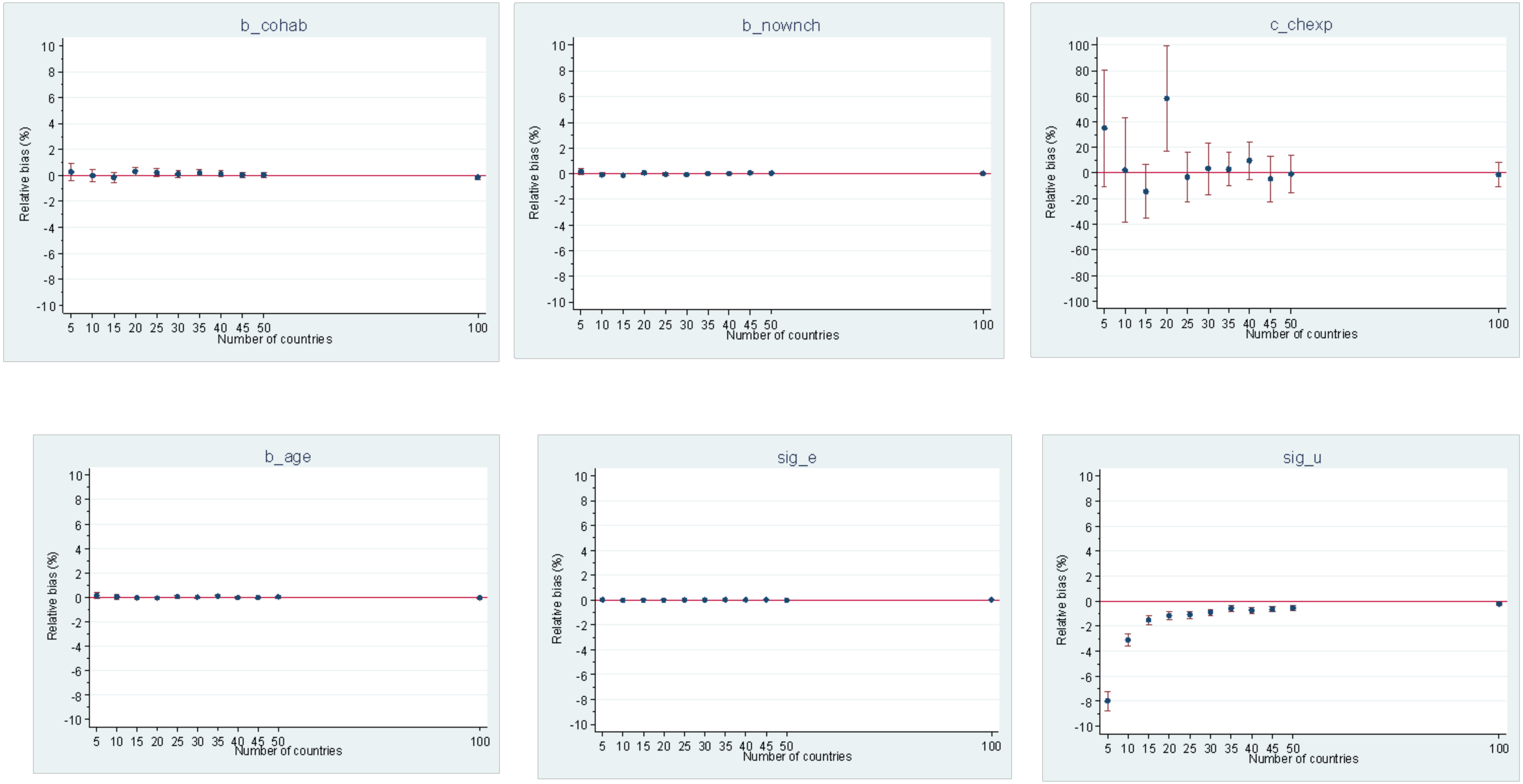
- Simulation and estimation: Stata (version 11)
- Estimation used software command defaults – as most users would
- Fitting of linear models: xtmixed command's REML estimator
- Fitting of non-linear models: xtmelogit command's adaptive quadrature estimator with 7 integration points
  - Also have results from MLwiN's PQL2 estimator (via runmlwin)
- Number of replications, *R*, as large as possible in order to reduce simulation variability, subject to constraints on estimation time
  - Linear model: 10,000 (basic) and 5,000 (extended)
  - Binary logit model: 5,000 (basic) and 1,000 (extended)

Summarizing MC analysis results

1. *Relative parameter bias*
   - Percentage difference between each estimated parameter and true value at each replication, averaged over *R* replications
   - Ideal reference point: 0% for each parameter
2. *Relative standard error bias*
   - Percentage difference between 'analytical' and 'empirical' SEs
     - Analytical SE: Stata-reported SE, averaged over *R* replications
     - Empirical SE: standard deviation of estimated parameter from same *R* replications
   - Ideal reference point: 0% for each SE
3. *Non-coverage rate for 95% CI*
   - At each replication, calculate (i) 95% CI for each parameter given estimated SE and assuming normality, and (ii) binary non-coverage indicator = 0 if CI includes true parameter and = 1 if did not. (iii) non-coverage rate is average of indicator over *R* replications
   - Ideal reference point: 0.05
   - Rates larger than 0.05: estimated CI too narrow (reflects 1 and/or 2)

Results from analysis of ‘basic’ models: linear and logit

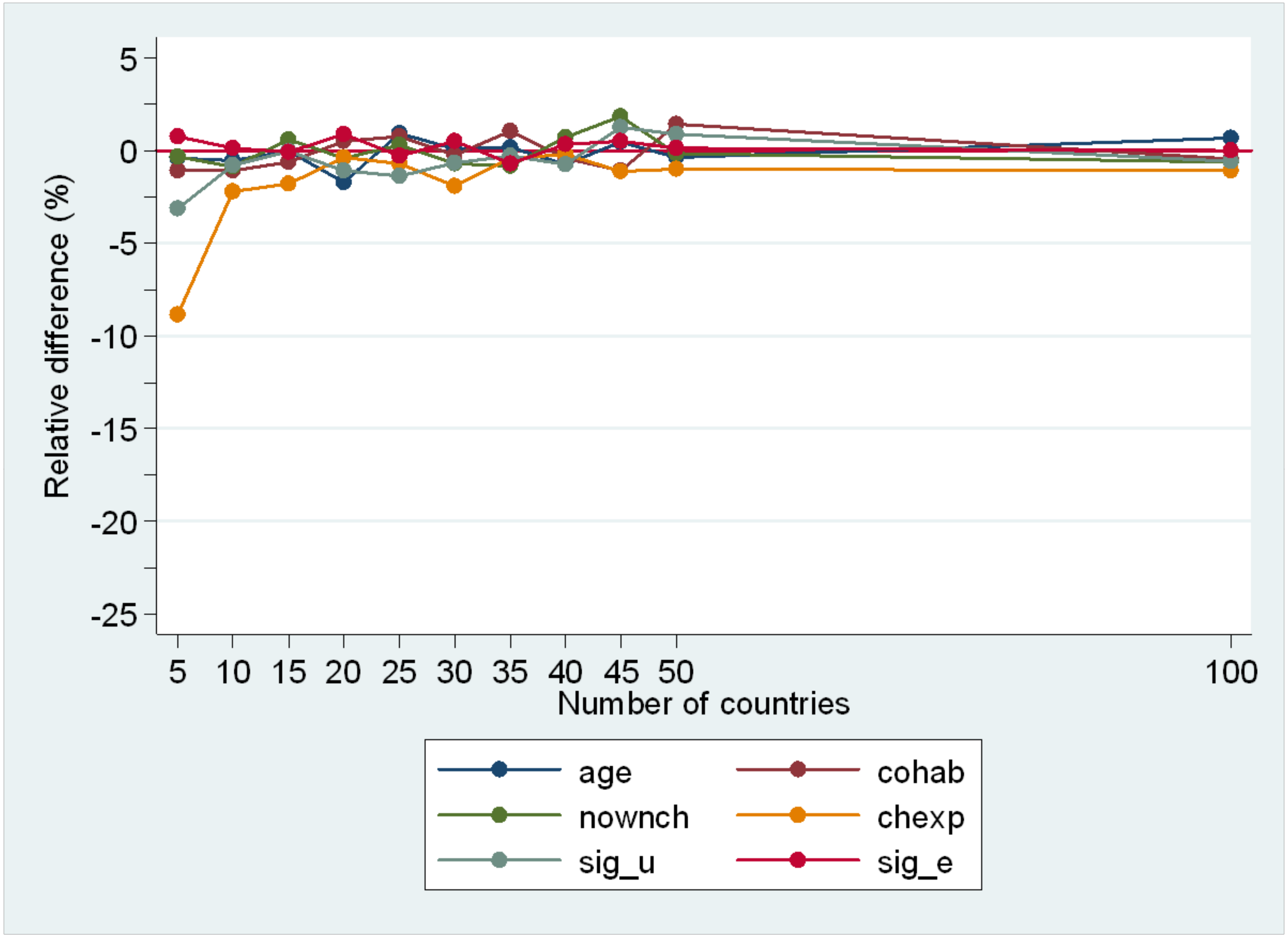
Basic linear model (‘work hours’): relative parameter bias



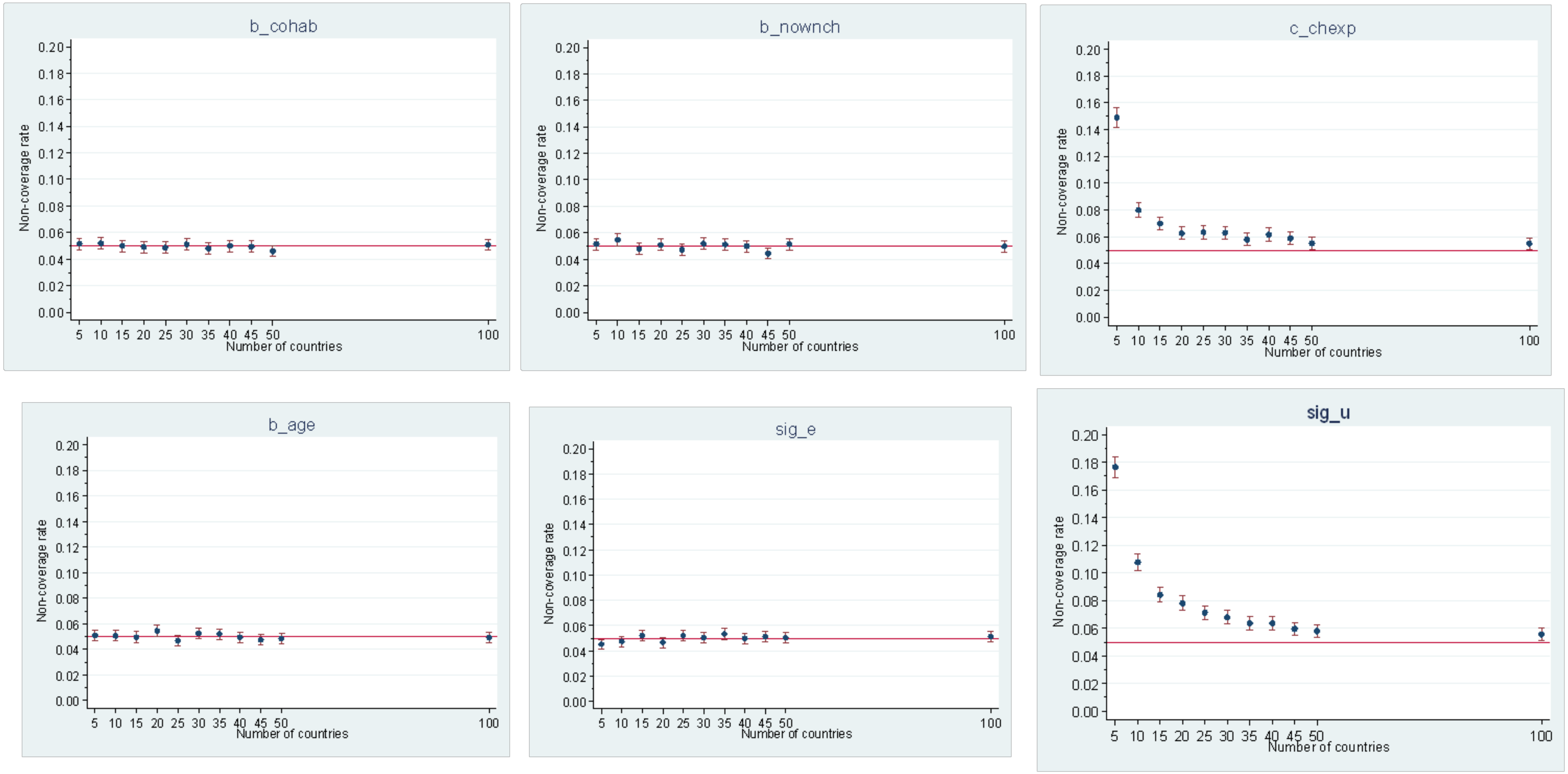
Relative parameter bias = percentage difference between estimated parameter and the true parameter, averaged over *R* = 10,000 replications (ideally 0%)

Vertical lines show normal 95% CI of mean of *R* estimates

Basic linear model (‘work hours’): relative SE bias

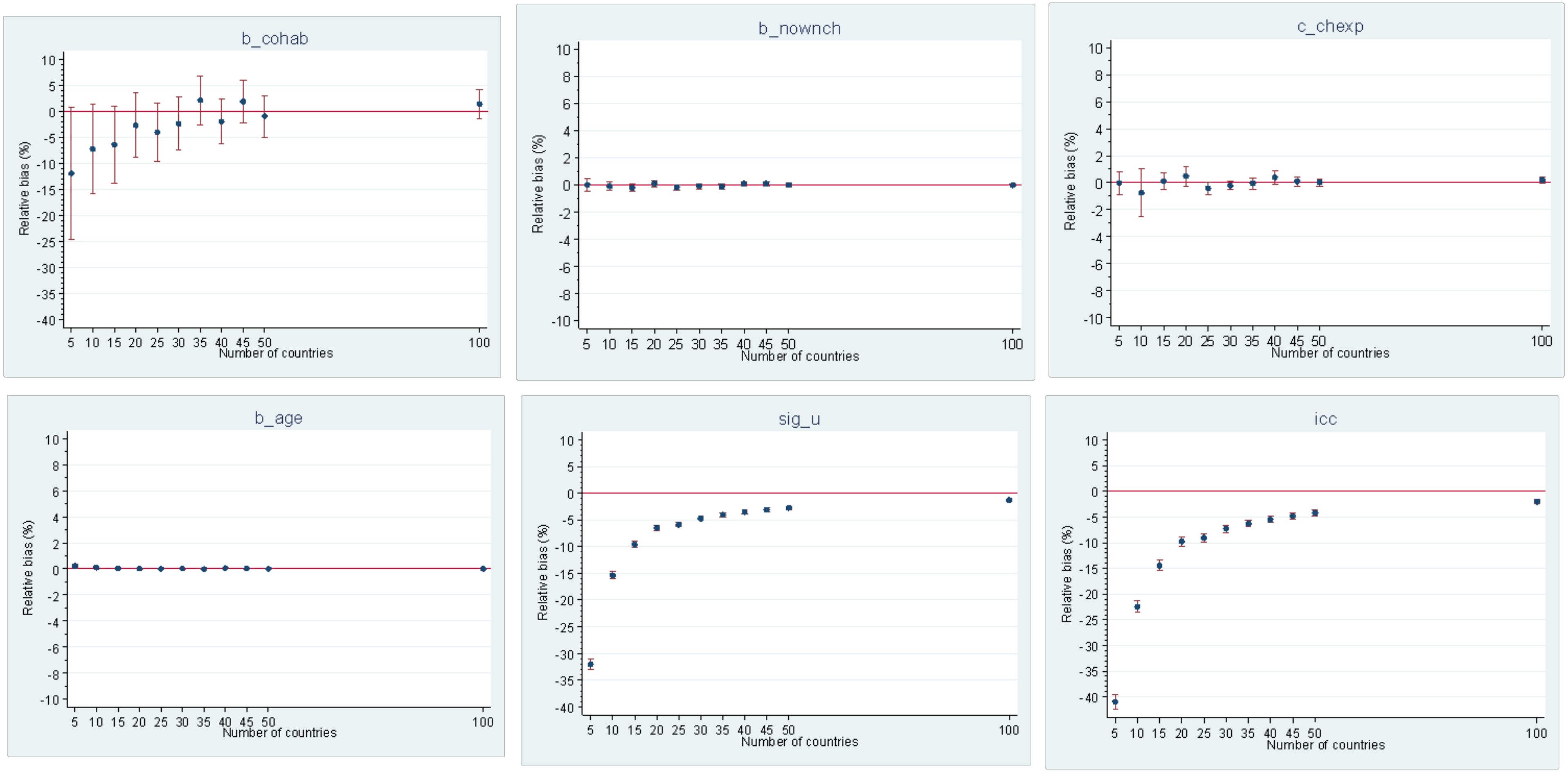


Basic linear model (‘work hours’): non-coverage rate



Non-coverage rate is average over R = 10,000 replications of indicator of whether estimated 95% confidence interval does not contain true parameter value (rate is ideally 0.05). Rates > 0.05 indicate estimated SEs are too small and so estimated CIs are too short

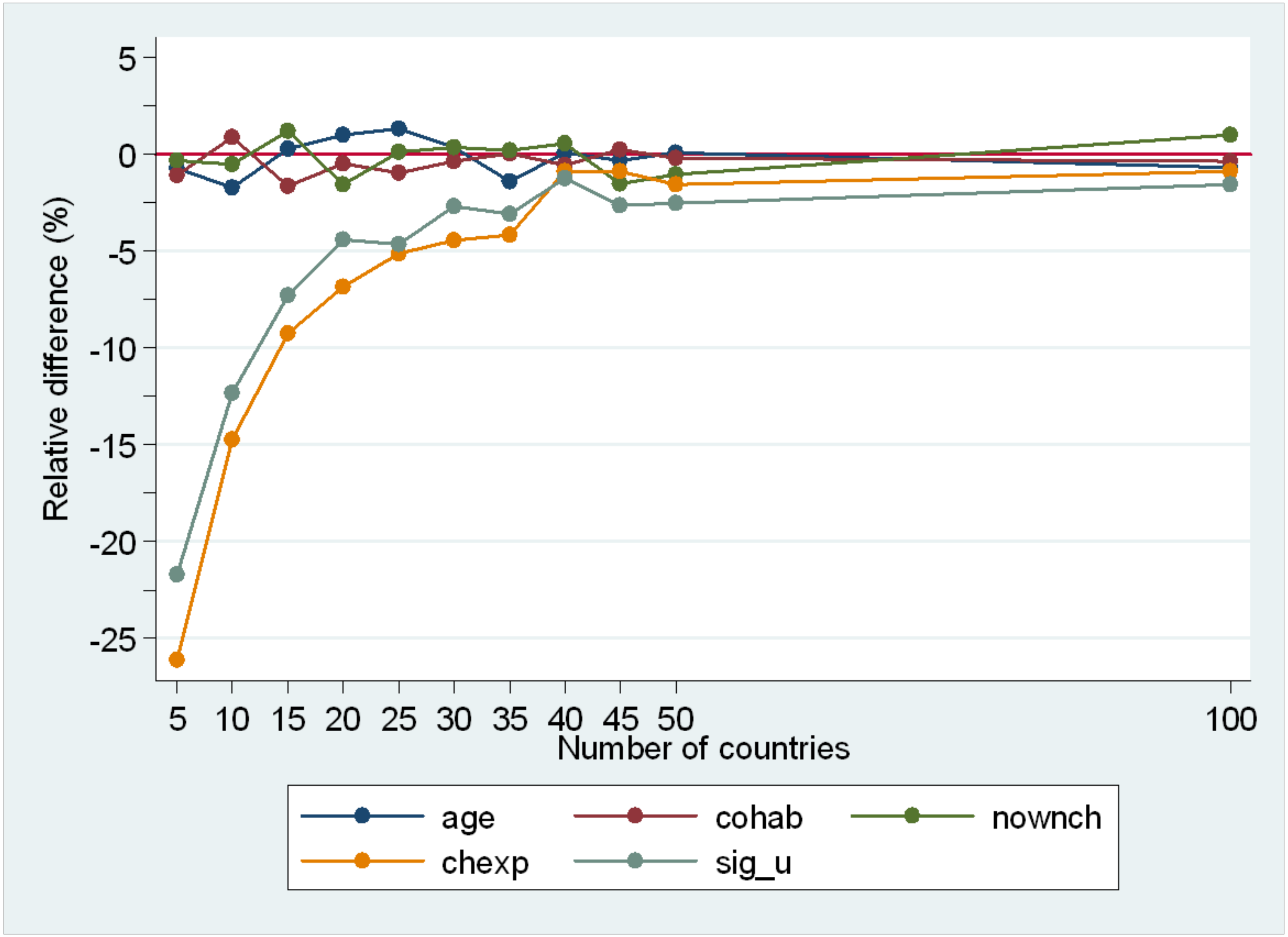
Basic logit model (‘participation’): relative parameter bias



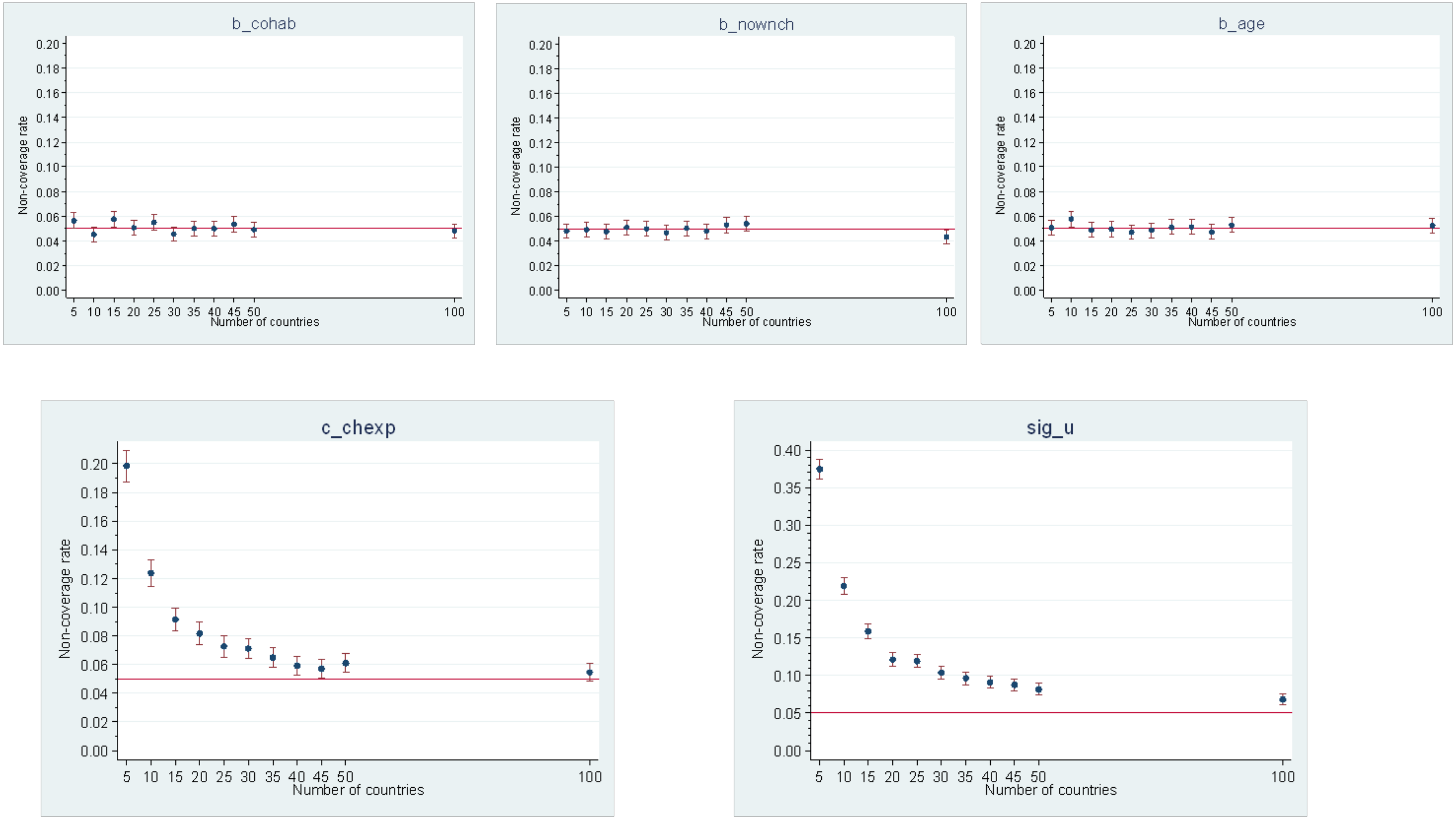
Relative parameter bias = percentage difference between estimated parameter and the true parameter, averaged over R = 5,000 replications (ideally 0%)

Vertical lines show normal 95% CI of mean of R estimates

Basic logit model (‘participation’): relative SE bias



Basic logit model (‘participation’): non-coverage rate



NB vertical scales differ across graphs

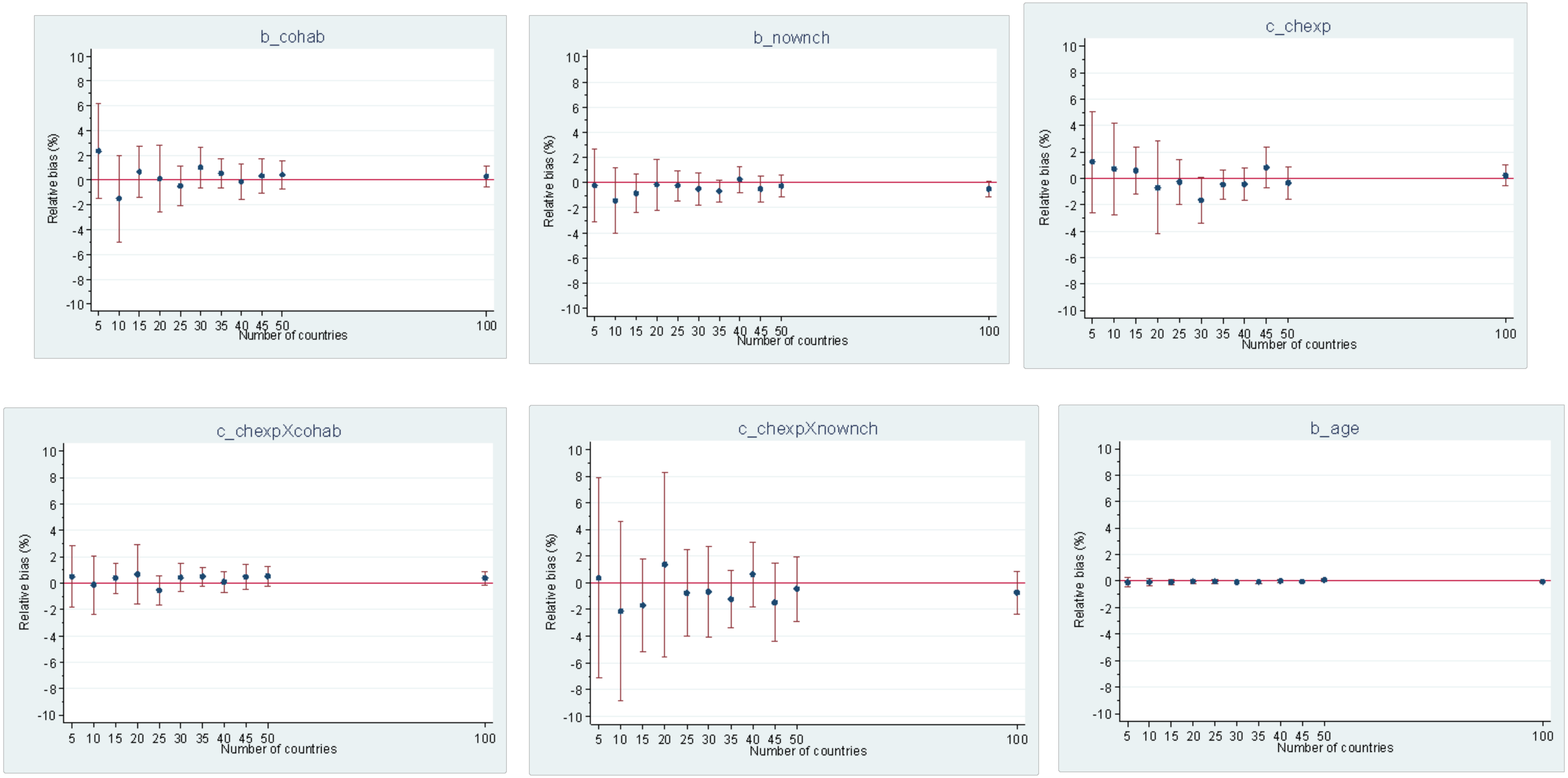
Results from analysis of ‘extended’ models: linear and logit

Adding 2 random slopes and
a cross-level interaction

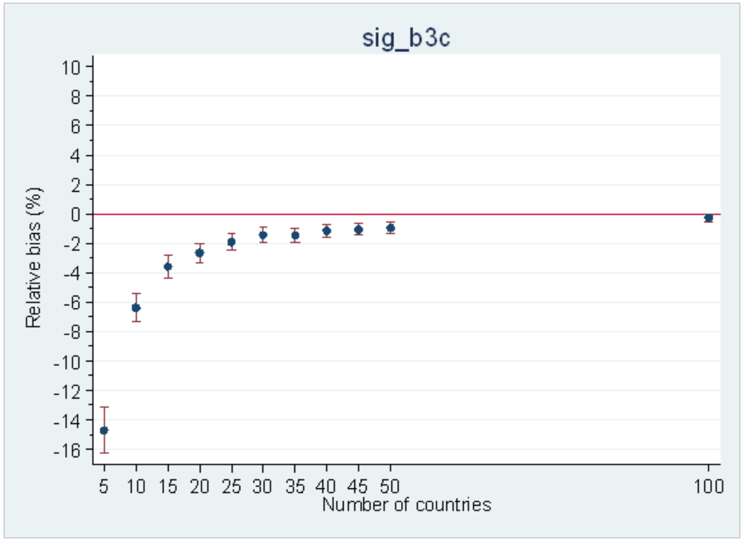
Linear model ('hours') with random intercept, 2 random slopes, country-level regressor & cross-level interaction

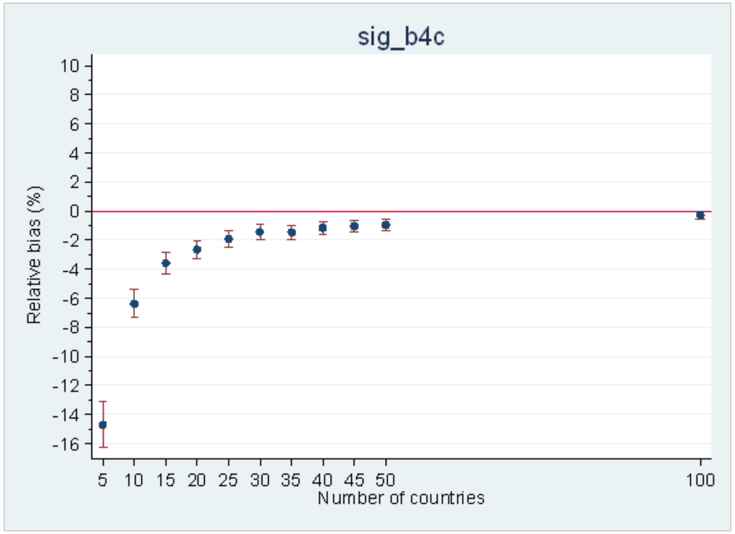
Hours_ic = b0		b0 = 22
+ b1 * age_ic		b1 = 0.8
+ b2 * age-squared_ic		b2 = – 0.01
+ b3 * cohab_ic + b3c * cohab_ic	← random slope	b3 = –1
+ b4 * nownch_ic + b4c * nownch_ic	← random slope	b4 = –1.2
+ b5 * isced3_ic		b5 = 0.7
+ b6 * isced4_ic		b6 = 1.4
+ b7 * isced56_ic		b7 = 1.6
+ c1 * chexp_c	← country-level	c1 = –2.7
+ c2 * (chexp_c X cohab_ic)	← country-individual interaction	c2 = 2.4
+ c3 * (chexp_c X nownch_ic)	← country-individual interaction	c3 = 0.7
+ u_c		
+ e_ic		
u_c ~ N(0, sig_u^2)		sig_u = 2.4
e_ic ~ N(0, sig_e^2)		sig_e = 9.4
cov(u_c, e_ic) = 0		⇒ ICC ≈ 0.06
b3c ~ N(0, sig_b3c^2)	← variance of random slope	sig_b3c = 1.2
b4c ~ N(0, sig_b4c^2)	← variance of random slope	sig_b4c = 1.2

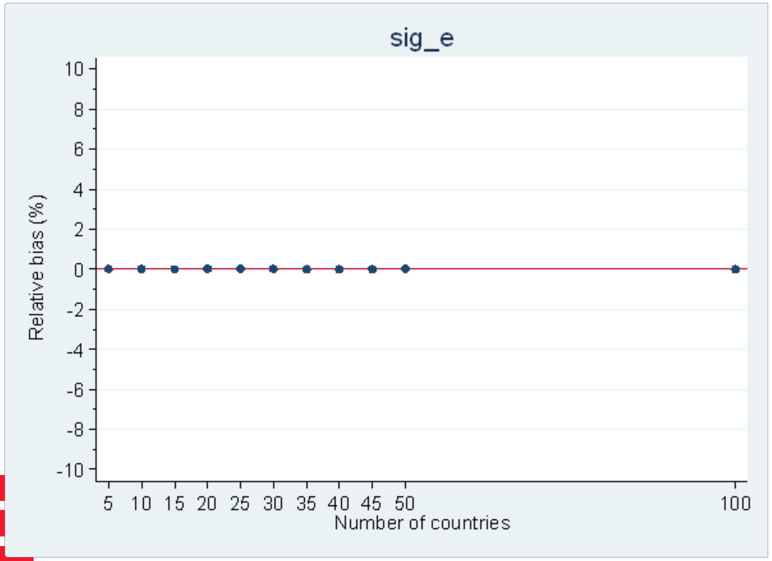
Extended linear model (‘hours’): relative parameter bias

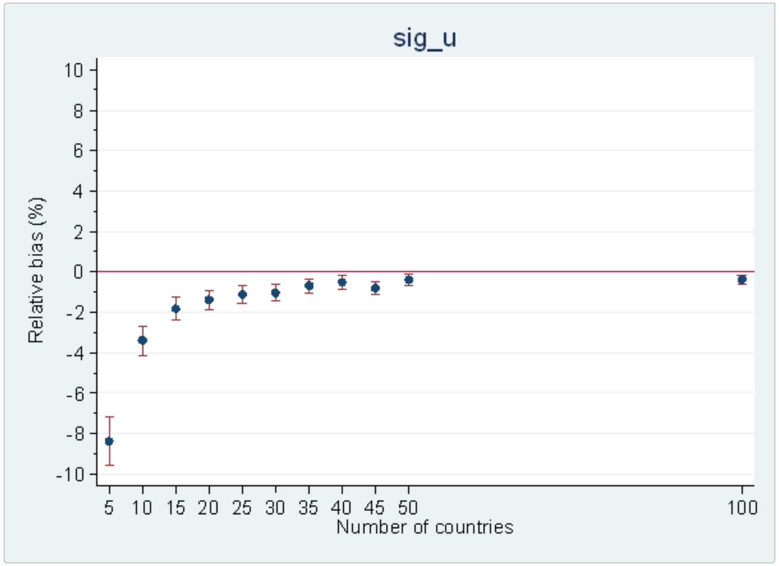


Extended linear model (‘hours’): relative parameter bias, ctd.

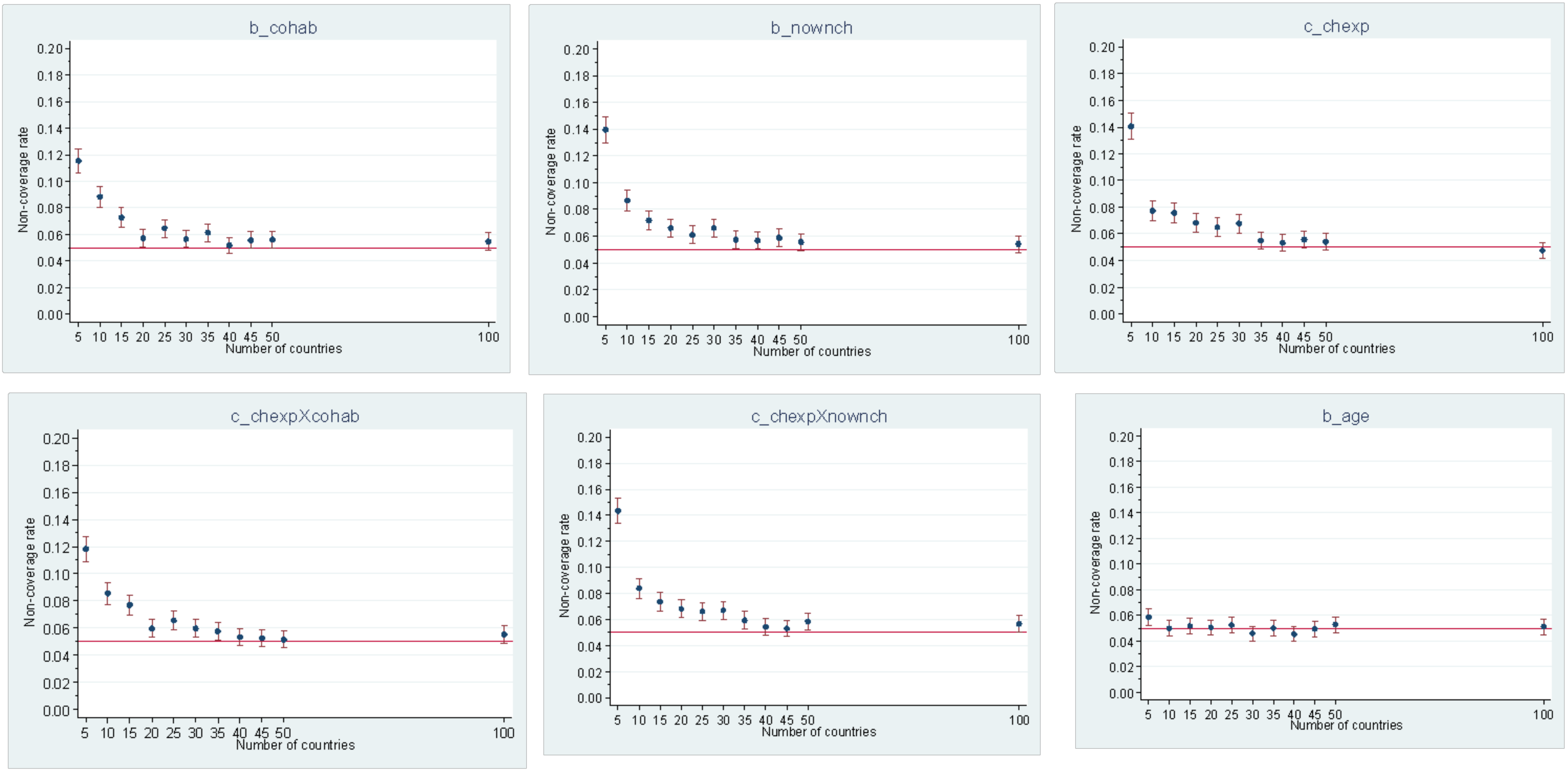




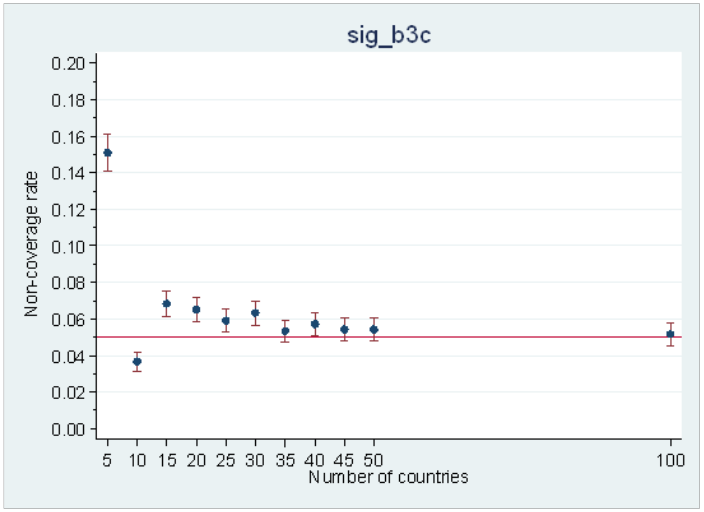


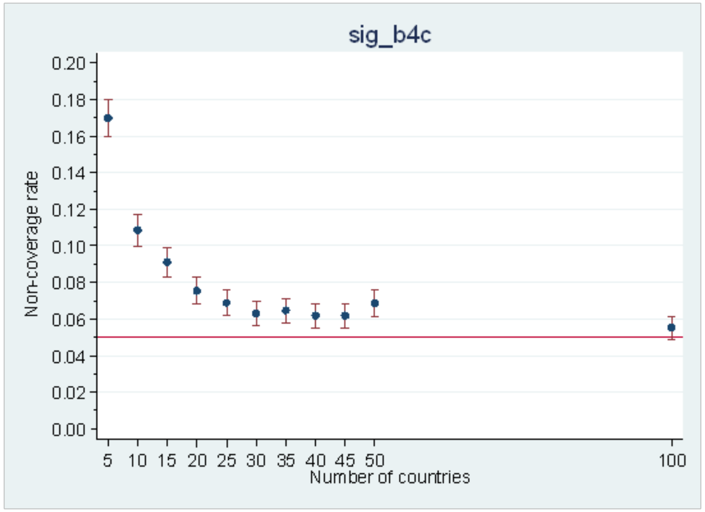


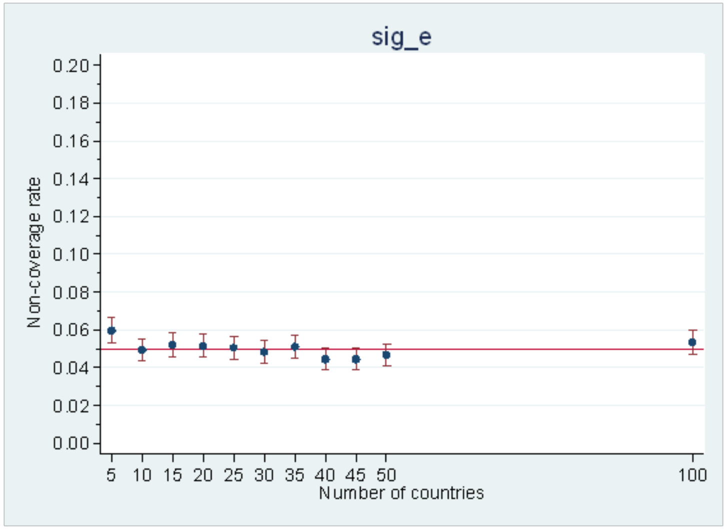
Extended linear model (‘hours’): non-coverage rate

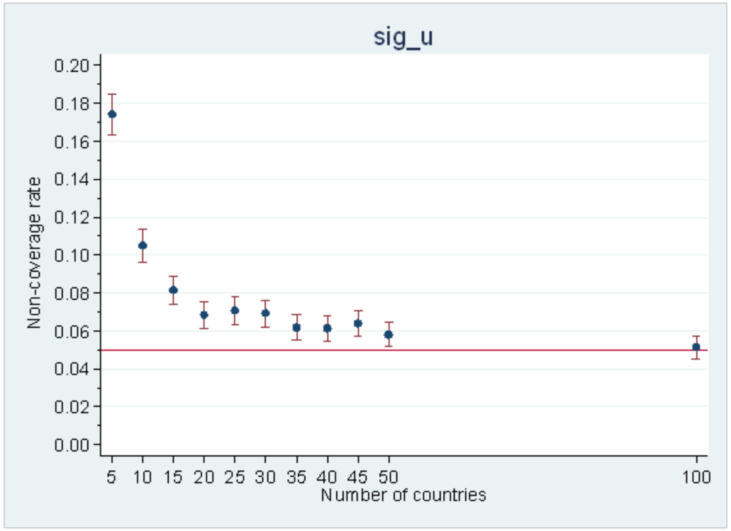


Extended linear model (‘hours’): non-coverage rate, ctd.





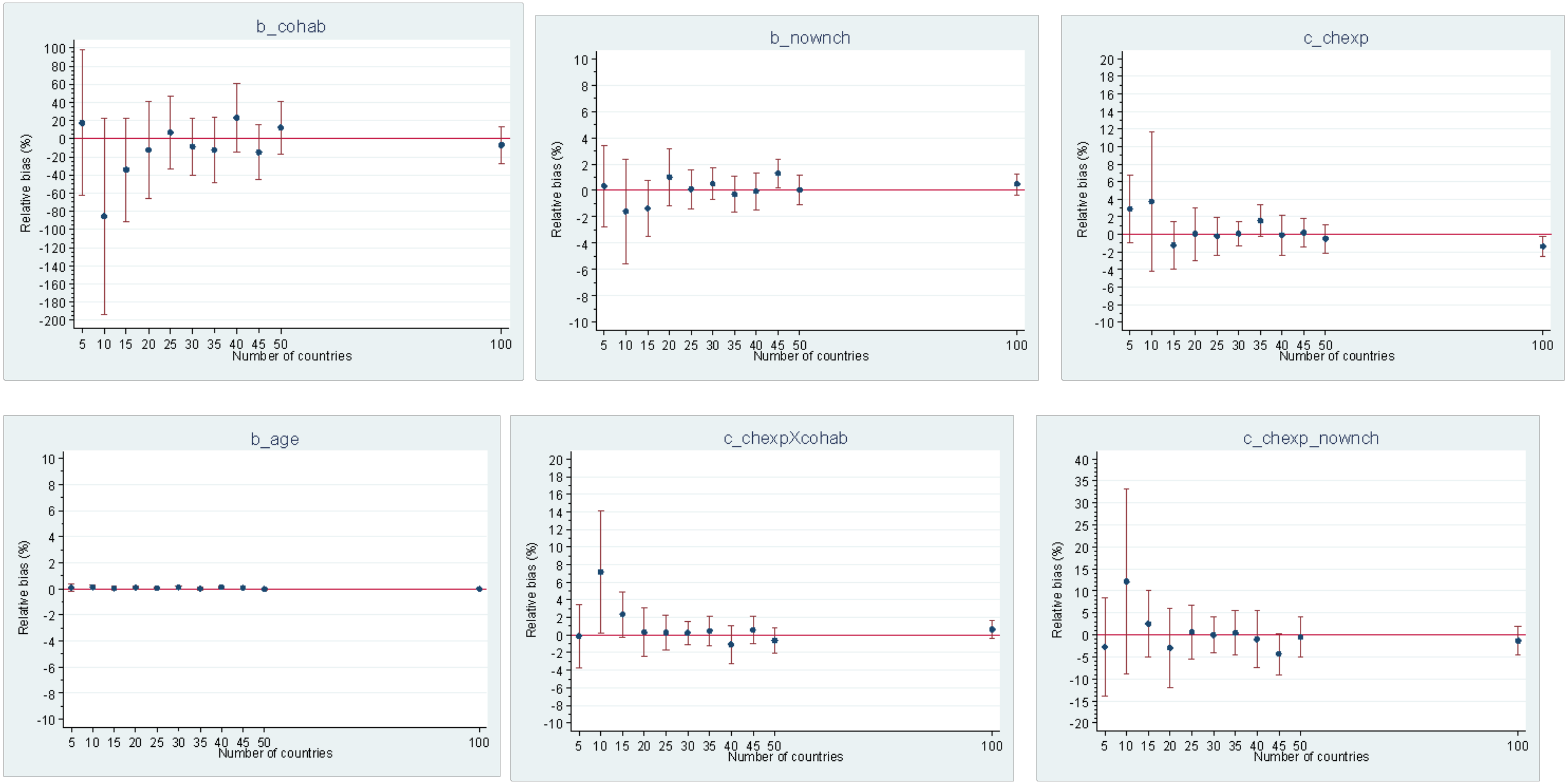




Binary logit model ('participation') with random intercept, 2 random slopes, country-level regressor, and cross-level interaction

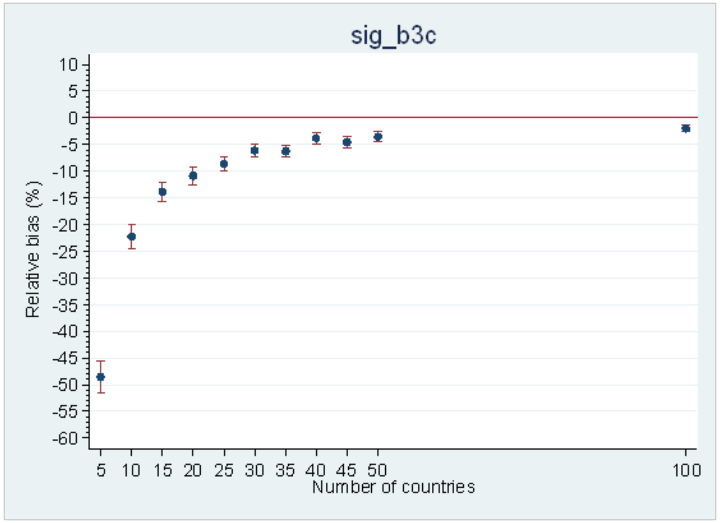
Hours_ic = b0		b0 = –9.1
+ b1 * age_ic		b1 = 0.5
+ b2 * age-squared_ic		b2 = – 0.006
+ b3 * cohab_ic + b3c * cohab_ic	← random slope	b3 = 0.02
+ b4 * nownch_ic + b4c * nownch_ic	← random slope	b4 = –0.27
+ b5 * isced3_ic		b5 = 0.7
+ b6 * isced4_ic		b6 = 0.9
+ b7 * isced56_ic		b7 = 1.4
+ c1 * chexp_c	← country=level	c1 = 0.7
+ c2 * (chexp_c X cohab_ic)	← country-individual interaction	c2 = 0.6
+ c3 * (chexp_c X nownch_ic)	← country-individual interaction	c3 = –0.1
+ u_c		
+ e_ic		
u_c ~ N(0, sig_u^2)		sig_u = 0.38
e_ic ~ cumlogit(0, sig_e^2)		sig_e = sqrt(_pi^2)/3
cov(u_c, e_ic) = 0		⇒ ICC ≈ 0.042
b3c ~ N(0, sig_b3c^2)	← variance of random slope	sig_b3c = 0.25
b4c ~ N(0, sig_b4c^2)	← variance of random slope	sig_b4c = 0.13

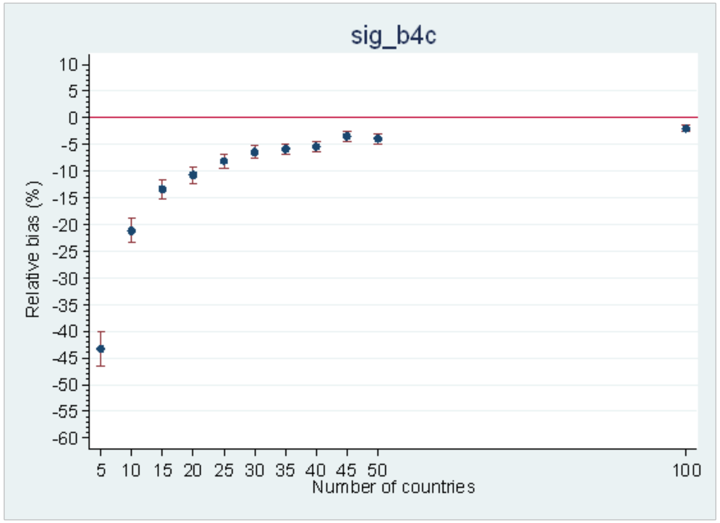
Extended logit model (‘participation’): relative parameter bias

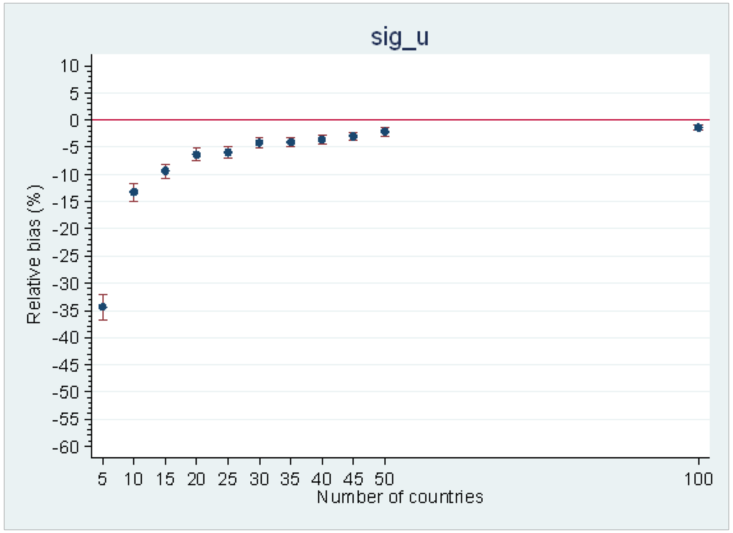


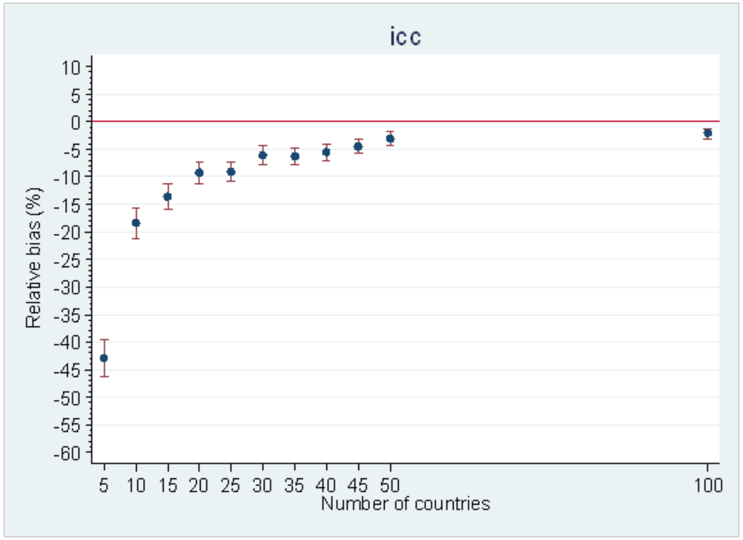
NB vertical scales differ across graphs

Extended logit model (‘participation’): relative parameter bias

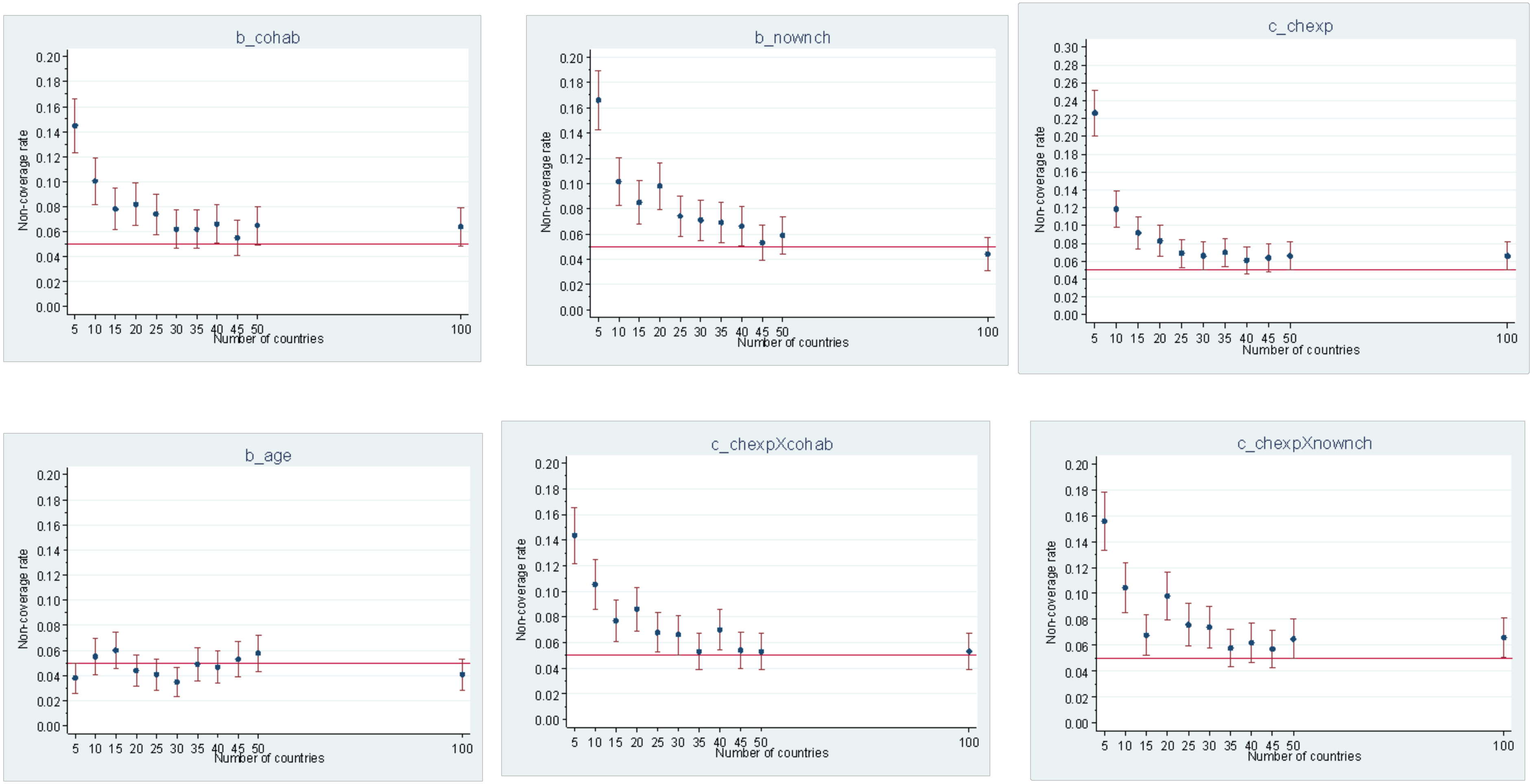




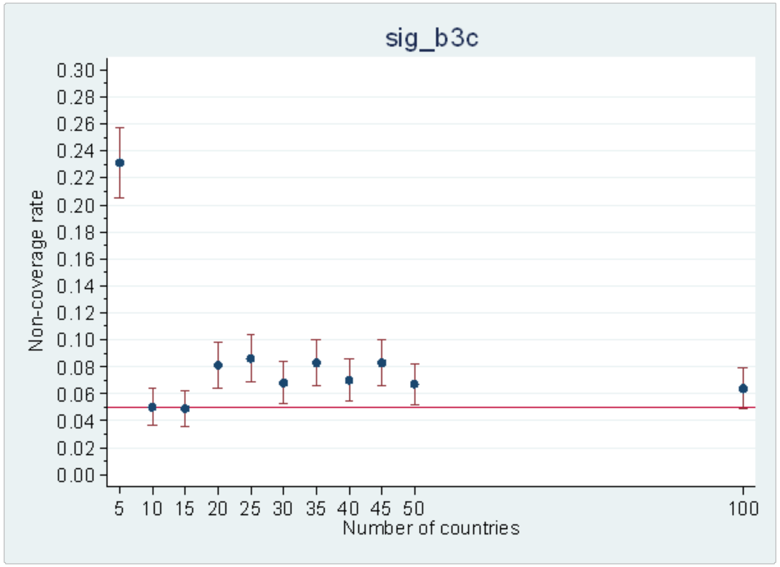


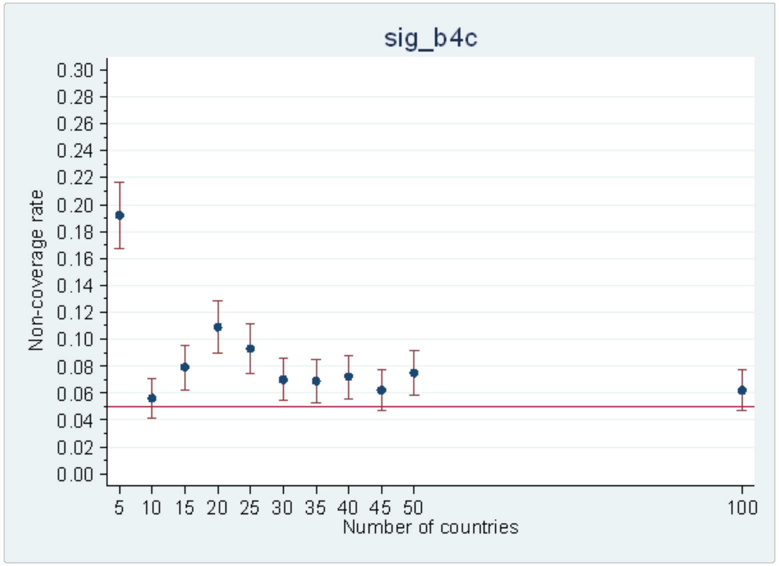


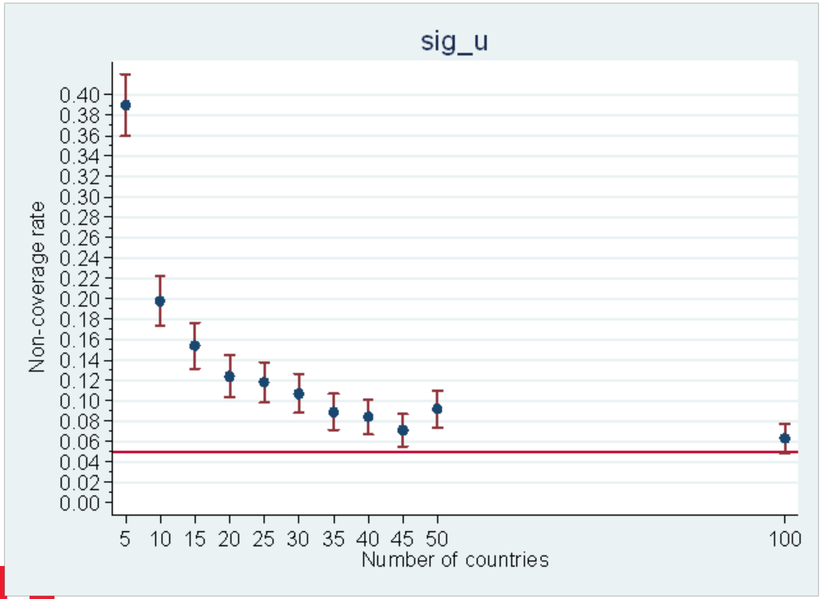
Extended logit model (‘participation’): non-coverage rate

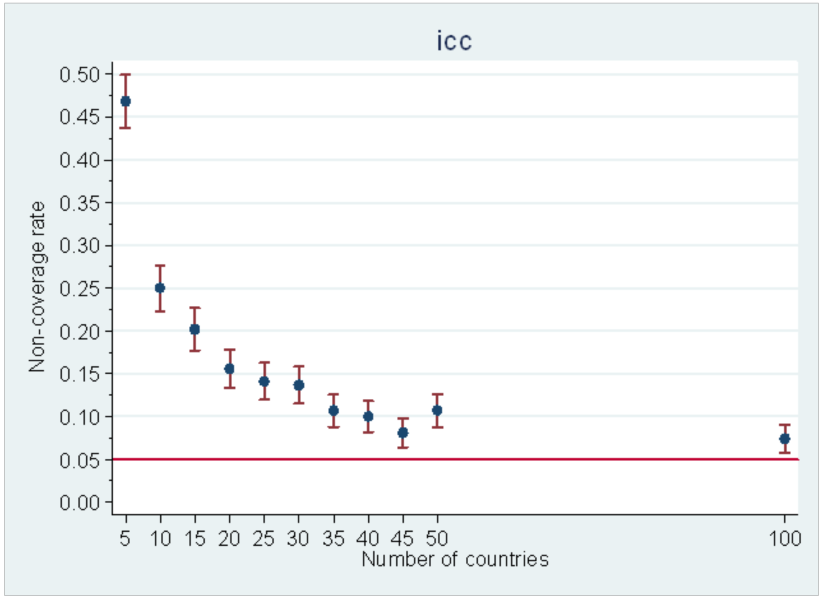


Extended logit model (‘participation’): non-coverage rate, ctd.









NB vertical scales differ across graphs

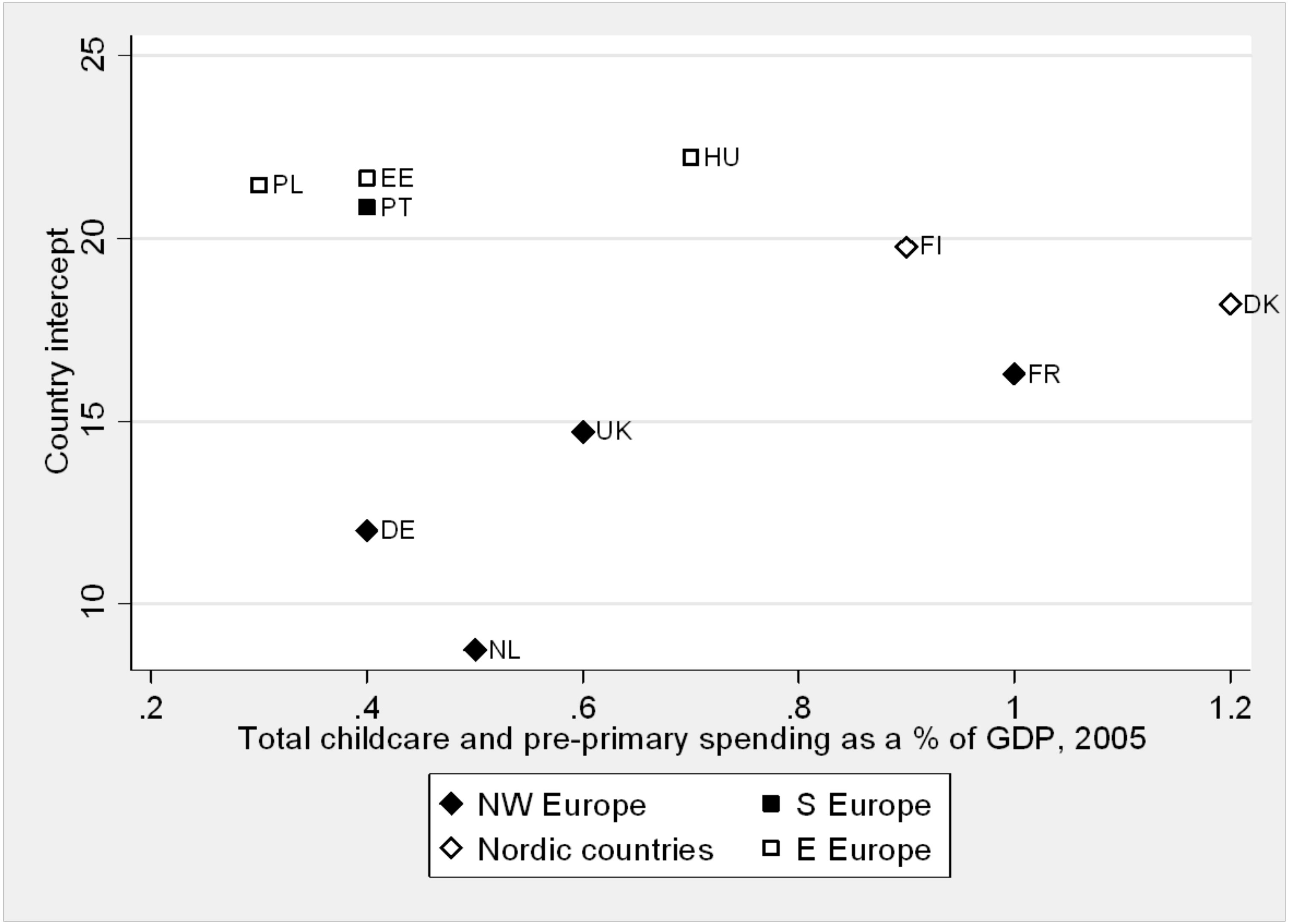
MC analysis: summary

- Fixed coefficients (for variables not interacted across levels) and level-1 variances are unbiased and have good coverage rates, even if C is very small
- Much simulation variability in estimates of relative bias or non-coverage rate for fixed country coefficient, and also cross-level interaction coefficients
- Appreciable downward bias in estimates of some country-level variances (and hence ICC), in both linear and binary logit models, even for $C \approx 20$
- Non-coverage problems, especially for country-level variances, even for $C \approx 25$
- Relative parameter and SE bias and non-coverage rates tend to be worse for binary logit models than for linear model
- Rule-of thumb? $C \geq 25$ for linear model, $C \geq 30$ for logit model

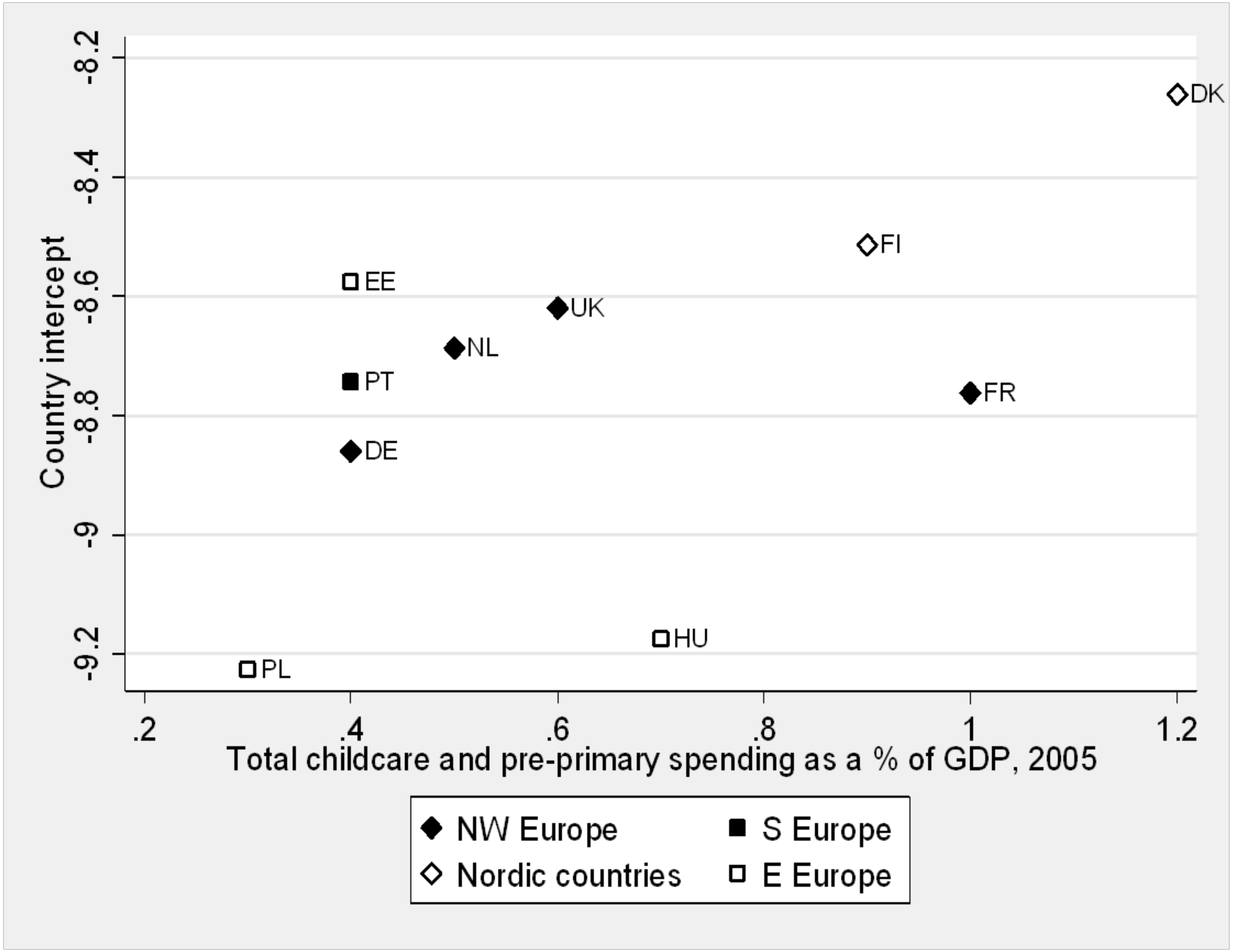
Two-step method and graphical summaries of country differences

- The two-step method leads naturally to exploratory data analysis: graphical summary of country-level variations in outcomes in which one plots the country intercepts fitted at step 1 against elements of $\boldsymbol{Z}_c$
  - Kedar and Shively (2005), Bowers and Drake (2005)
  - "Showing what is in the data and not more"
- Eyeball evidence about how country differences in women's work hours and labour force participation are related to differences in childcare/pre-primary spending (outliers? clusters?)
  - First step of two-step method provides country-specific intercepts; these estimates can be related to country-specific spending

Country-specific intercepts and childcare spending: womens' hours of work (EU-SILC data, 2007)



Country-specific intercepts and childcare spending: womens’ labour force participation (EU-SILC data, 2007)



Conclusions

- MLM users need to be cautious in claims made about 'country effects' in terms of coefficient point estimates and of their statistical significance when *C* is not 'large', especially in non-linear models
  - Problems can be apparent even for $C \approx 25$ (more countries than many use!)
  - Appropriate *C* depends on how much inaccuracy one is prepared to tolerate!
- [Alternative approach: suppose country effects incorporated in 'fixed' country intercepts]
- Consider non-statistical techniques to assess country effects when the number of countries is small:
  - Graphical and other exploratory data analysis techniques, e.g. plotting country intercepts and coefficients, with CIs, against key country-level variables
  - Richer description of country differences in terms of workings of national institutions (quantitative regression analysis can pull no rabbits out of the hat)
- Explore additional (less familiar) approaches to estimation and inference:
  - Bias adjustment for linear model fixed parameter SEs (SAS, R), nonparametric bootstrapping (MLwiN, SAS)
  - Bayesian approaches (MCMC in MLwiN, BUGS)
  - But implementation requires more statistical expertise among MLM users …

Further details of the study and results

- Bryan, M. L. and Jenkins, S. P. (2013). ‘Regression analysis of country effects using multilevel data: a cautionary tale’, Working Paper 2013-14**.** Colchester: ISER, University of Essex. https://www.iser.essex.ac.uk/publications/working-papers/iser/2013-14
- Jenkins, S. P. (2013). ‘A Monte-Carlo analysis of multilevel binary logit model estimator performance’, Presentation at the Stata User Group Meeting, London, 13 September 2013. http://repec.org/usug2013/jenkins.uk13.pdf
  - Compares Stata’s adaptive quadrature estimator with MLwiN’s PQL2 estimator (former is much slower but more accurate)