How and how much does expert error matter? ReadMe/Codebook

Kyle L. Marquardt*

November 4, 2020

This file contains replication materials for the analyses of actual data discussed in Online Appendix F of the article, "How and how much does expert error matter" (forthcoming in the *Journal of Peace Research*), which are the analyses most relevant to researchers interested in applying the methods discussed in the paper to their own work. Complete replication materials, including all simulated data, will be available at http://www.prio.org/jpr/datasets.

This ReadMe contains first a discussion of how to replicate the analyses and apply the models to different data; followed by a description of the the replication code, model files and data files necessary to replicate these analyses. All analyses use R; some analyses require the packages **coda**, **boot**, **ggplot2**, **gridExtra**, **lme4**, **parallel**, **rstan**, **tidyverse**, **zoo**.

In addition to the article, all analyses using the country-year UCDP data should cite Gleditsch et al. (2002); Pettersson & Wallensteen (2015); Girardin et al. (2015). All analyses using the V–Dem data should cite Coppedge et al. (2018); Pemstein et al. (2018).

1 Replication instructions

1.1 Structural equation models

In order to replicate the Bayesian structural equation models, run the file **sem.R**. This file requires 1) the country-year dataset with all relevant regression variables and identifiers linking (reduced) coder-level data with these data (**uv.Rdata**), 2) the reduced coder-level data (**v2clsocgrp.rds**), and 3) Stan model files (**jpr_le.stan** for the IRT model and **jpr_am.stan** for the A–M model).

In order to pursue a similar modeling approach using other data, it is necessary to create a dataset by merging the relevant V–Dem coder-level data identifiers with the additional data. The file **dataset.R** provides a template for doing so. It is also necessary to edit the regression equation in the Stan files to reflect the new data, and relabel variables in the R code; the latent variable portion of the analysis should require minimal alterations for use with other V–Dem data.

*School of Politics and Governance and International Center for the Study of Institutions and Development, National Research University Higher School of Economics

1.2 Point estimate models

In order to replicate the Bayesian regression analyses, run the file **bayesian_point_estimates.R**. This file requires the country-year dataset with all relevant regression variables (**uv.Rdata**), the latent point estimates already having been merged with the UCDP data; and the model file **jpr_nE.stan**. While the mean and median over expert codings can be estimated directly using the coder-level data **v2clsocgrp.rds**, estimating the posterior median from the IRT and A–M models requires the relevant posterior distributions (**v2clsocgrp.10000.beta.RData** and **v2clsocgrp.10000.am.RData**, respectively; summarized in **zmedian.RData** and **amedian.RData**), which in turn were estimated using **latent_models.R** and the Stan code **beta.stan** (IRT) and **ham.stan** (A–M).

In order to pursue a similar modeling approach using other data, it is necessary to edit the model Stan code to reflect the new model, and create a dataset merging the relevant coder-level point estimates with the other relevant data. The file **dataset.R** provides a template for performing the latter task. Note that the V–Dem dataset already provides estimates of the mean and posterior median for most expert-coded variables (though the latent variable models in the V–Dem dataset use empirical priors, whereas the analyses here use vague priors centered at zero), and it is likely easier to simply use these data than building a dataset from the bottom-up. That said, the Stan and R latent variable modeling code included in this Dataverse are readily applicable to all V–Dem variables with only minor alterations.

1.3 Analyses using multiple imputation

In order to replicate the Bayesian multiple imputation model, run the file **mi.R**. This file requires 1) the country-year dataset with all relevant regression variables and identifiers linking (reduced) coder-level data with these data (**uv.Rdata**), and 2) the posterior distribution from the IRT model (**v2clsocgrp.10000.beta.RData**).

In order to pursue a similar modeling approach using other data, it is necessary to create a dataset by merging the relevant V–Dem coder-level data identifiers with the additional data. The file **dataset.R** provides a template for doing so. It is also necessary to have a posterior distribution for the latent variable, which can be estimated using as templates the file **latent_models.R** and the Stan code **beta.stan**. Finally, it is also necessary to edit the regression equation in the R Stan code to reflect the new data, and relabel variables in the R code.

Note that I only analyzed IRT—not A–M—multiple imputation in the context of the actual data, and therefore only include data for the IRT-based model here; however, running the model with A–M data would require only replacing the IRT posterior distribution with the A–M posterior distribution in the R code.

1.4 Estimating posterior-predicted effects

The file **bayesian_actual.R** provides code to illustrate the predicted effects using the posterior distribution from the aforementioned regression models. In addition to the relevant posterior distributions, many analyses require data from the dataset (**uv.Rdata**) and thresholds from the latent variable posterior distribution (**gamma.RData**).

To apply this code to other models, the prediction equation must be altered and variables relabeled as appropriate.

1.5 Bootstrap analyses

The file **bootstrap.R** provides code to estimated bootstrapped effect estimates using the posterior distribution from the aforementioned regression models. These analyses require the dataset (**uv.Rdata**) and thresholds from the latent variable posterior distribution (**gamma.RData**). The files **re_graphics.R** and **fe_graphics.R** provide code to create graphics for the random and fixed effects analyses, respectively.

To apply this code to other models, the estimation and prediction equations must be altered and variables relabeled as appropriate.

2 Detailed description of files

- Replication files
  - **latent_models.R**: Code to run, diagnose, and summarize both IRT and A–M models using actual data.
  - **dataset.R**: Code to 1) estimate country-year mean and median estimates, 2) interpolate reduced point estimates (mean, median, A–M and IRT posterior medians) over full time period, and 3) merge point estimates with UCDP data. Note: point estimates merged with one-year lag; cases of ongoing conflict removed; country-year ids in dataset necessary to running structural and multiple imputation models.
  - **bayesian_point_estimates.R**: Code to run Bayesian regression analyses of conflict onset using median, mean and posterior median estimates as right-hand side variables.
  - **bayesian_actual.R**: Code for generating Bayesian graphics in the Appendix. Figure F2
  - **mi.R**: Code to run multiple imputation model using latent variable model output.
  - **sem.R**: Code to run and diagnose structural equation model.
  - **bootstrap.R**: Code to run bootstrap analyses using latent variable model output.
  - **descriptive_stats.R**: Code for descriptive analyses in the Appendix. Generates Figures B1, B2, F1, F3.
  - **re_graphics.R**: Code for generating random-effect graphics in the Appendix. Figure N1.
  - **fe_graphics.R**: Code for generating fixed-effect graphics in the Appendix. Figure N8.
- Datasets
  - **uv.Rdata**: Dataset including country-year UCDP data and point estimates for inclusion. Point estimates and point-estimate regime identifiers lagged one year. Unique variables:
    - **onset_ko_flag**: Conflict onset indicator
    - **Z**: Posterior median inclusion estimate (IRT model), lagged by one year

    - **AM**: Posterior median inclusion estimate (A–M model), lagged by one year
    - **wmed**: Zero-centered median over-expert scores inclusion estimate, lagged by one year
    - **wavg**: Normalized average over-expert scores inclusion estimate, lagged by one year
    - **peaceyears**: Count of years since last conflict or dataset beginning
    - **peaceyears2**: Squared count of years since last conflict or dataset beginning
    - **peaceyears3**: Cubed count of years since last conflict or dataset beginning
    - **year**: Year identifier
    - **cid**: V–Dem country-text identifier
    - **rnames**: Regime identifiers, lagged by one year. Note that identifiers denote the first year of a regime, with regimes being defined as a period in which no coder for a country changes her score.
  - **data (2).csv**: Country-year UCDP dataset from GROW^{up}
  - **translation.csv**: Country identifiers to link V–Dem data with UCDP dataset

- Coder-level data
  - **v2clsocgrp.rds**: Reduced country-year data
  - **v2clsocgrpF.rds**: Complete country-year data for interpolation

- Point estimates
  - **amedian.RData**: Posterior median country-year estimates (A–M)
  - **zmedian.RData**: Posterior median country-year estimates (IRT)
  - **gamma.RData**: Posterior median universal threshold estimates (γ_μ, IRT)

- Latent variable posterior distributions
  - **v2clsocgrp.10000.am.RData**: Posterior parameter estimates for A–M latent variable model
  - **v2clsocgrp.10000.beta.RData**: Posterior parameter estimates for IRT latent variable model

- Bayesian regression analysis output
  - **jpr_am.10.post.RData**: Posterior parameter estimates for A–M SEM
  - **jprLE.10.post.RData**: Posterior parameter estimates for IRT SEM
  - **jprM.10.post.RData**: Posterior parameter estimates for regression using median over expert scores
  - **jprA.10.post.RData**: Posterior parameter estimates for regression using average over expert scores

  - **jpr.10.post.RData**: Posterior parameter estimates for regression using IRT posterior median
  - **jprME.5.post.RData**: Posterior parameter estimates for multiple imputation regression using IRT posterior draws

- Bootstrap regression analysis output
  - **bootZ.RData**: Bootstrap regression (random and fixed effect) estimates using IRT posterior median
  - **bootA.RData**: Bootstrap regression (random and fixed effect) estimates using average over expert scores
  - **bootM.RData**: Bootstrap regression (random and fixed effect) estimates using median over expert scores

- Model files
  - **jpr_am.stan**: A–M SEM
  - **jpr_lE.stan**: IRT SEM
  - **jpr_nE.stan**: Probit regression model
  - **beta.stan**: IRT model
  - **ham.stan**: A–M model

References

Coppedge, Michael; John Gerring, Carl Henrik Knutsen, Staffan I Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M Steven Fish, Agnes Cornell, Sirianne Dahlum, Haakon Gjerlow, Adam Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L Marquardt, Kelly M McMann, Valeriya Mechkova, Juraj Medzihorsky, Moa Olin et al. (2018) V–Dem Dataset v8. Technical report. Varieties of Democracy Project (dx.doi.org/10.2139/ssrn.3172819).

Girardin, Luc; Philipp Hunziker, Lars-Erik Cederman, Nils-Christian Bormann & Manuel Vogt (2015) Grow^{up}—Geographical Research On War, Unified Platform. Technical report. ETH Zurich (http://growup.ethz.ch/).

Gleditsch, Nils Petter; Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg & Håvard Strand (2002) Armed conflict 1946-2001: A new dataset. *Journal of Peace Research* 39(5): 615–637.

Pemstein, Daniel; Kyle L Marquardt, Eitan Tzelgov, Yi-Ting Wang, Joshua Krusell & Farhad Miri (2018) The V-Dem Measurement Model: Latent variable analysis for cross-national and cross-temporal expert-coded data. Working Paper 3rd Ed. Varieties of Democracy Project (dx.doi.org/10.2139/ssrn.3167764).

Pettersson, Therése & Peter Wallensteen (2015) Armed conflicts, 1946-2014. *Journal of Peace Research* 52(4): 536–550.