Advanced Quantitative Methods

Course for the master programme in peace and conflict studies,
Uppsala University March–May 2019

Håvard Hegre, Mihai Croicu and David Randahl

April 3, 2019

UPPSALA UNIVERSITET

Assignment 1: Interactions I

In this assignment, we want you to look at interaction terms. Section 1 in this text specifies the initial DGP. In section 2, we want you to formulate the interpretation of a model with interaction term, and to look at the correlation that emerges when two variables that have the same sign are interacted. In section 3, we want you show that omitting the interaction term from the initial DGP leads to omitted variable bias. In section 4, you should show that centering reduces collinearity problems that interaction models have given the initial DGP. In section 5, you should demonstrate that these collinearity problem increase when the two interaction variables are highly correlated.
Remember to install and load the `car,mvtnorm` packages

1 Assume that the true data-generating process has the form

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + u_i$$

where

- u_i has a normal distribution N(0,σ)
- X_1, X_2 are drawn from normal distributions $N(3, 1)$ (with mean = 3 and standard deviation = 1)
- $X_1 \times X_2$ is a multiplicative interaction term

UPPSALA UNIVERSITET

Assignment 1: Interactions II

- Set the true values for the β parameters to be an arbitrary combination of the values $(-1, -0.5, 0.5, 1)$, and the size of the sample to $N = 500$

2 Draw a realization from this DGP, fit two linear models – one including all the four terms and one omitting the interaction term

- Present the results in a joint regression table and interpret the results.
- Plot the predicted value for Y as a function of $X1$ for $X2 = 2$ for $X2 = 4$ and interpret the figures
- Present the joint distribution of X_1 and the multiplicative interaction term $X_1 \times X_2$ as three-dimensional density plots (one perspective plot and one contour plot). Describe and discuss.

3 Run a MC analysis to explore the consequences of omitting the interaction term from the model specification when fitting a regression model. Present results in terms of bias for the estimated parameter $\hat{\beta}_1$

UPPSALA UNIVERSITET

Assignment 1: Interactions III

4. Extend the MC analysis further by varying the correlation between X_1 and X_2 along the lines of the collinearity MC experiment from Chapter 5 in Carsey and Harden (2014). Again, focus on the fully specified model and concentrate on the collinearity issues in interaction models. For the centered and non-centered solution, report the following plots, discuss:
   1. The standard deviation of $\hat{\beta}_1$ as a function of `mc.level` – are estimates efficient?
   2. MSE for the β_1 estimates as a function of `mc.level` – are estimates biased?
   3. The mean variance inflation factor (VIF) for β_1 across simulations for each `mc.level`
   4. The estimated $\hat{\beta}_1$ against $\hat{\beta}_3$ for each of the repeated samples for `mc.level` = 0, 0.5 or .99. Choose either the centered or non-centered solution.
   5. Histograms of $\hat{\beta}_1$, $\hat{\beta}_3$, and the sum $\hat{\beta}_1 + \hat{\beta}_3$ for `mc.level` = 0, 0.5 or .99. Choose either the centered or non-centered solution. What happens here?

UPPSALA UNIVERSITET

Reformulating the linear model I

- The linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \epsilon \sim N(0, \sigma)$$

can also be written as Y following a normal distribution:

$$Y \sim N(\mu, \sigma), \mu = \beta_1 X_1 + \beta_2 X_2$$

- In the R script, this means that the formulation
`Y <- b0 + b1*X1 + b2*X2 + rnorm(n, 0, 1)`
is replaced with
`Y <- rnorm(n, b0 + b1*X1 + b2*X2, 1)`

UPPSALA UNIVERSITET

The logistic regression model I

- What if the dependent variable is dichotomous?
- The logistic regression model:

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1$$

  where $p = Pr(Y = 1)$
- Link function:
  the 'logit' $ln(\frac{p}{1-p})$ is the log odds of $Y = 1$
- Odds is the probability of something happening divided by the probability something else happens
- Inverting the logit:

$$Pr(Y = 1) = p = \frac{exp(Xb)}{1 + exp(Xb)}$$

UPPSALA UNIVERSITET

The logistic regression model II

- This gives

$$Pr(Y = 1) = \frac{exp(\beta_0 + \beta_1 X_1)}{1 + exp(\beta_0 + \beta_1 X_1)}$$

- Interpretation of β_1: How much the logit (log odds of $Y = 1$) increases when X increases by one unit
- Interpretation of $exp(\beta_1)$: Odds ratio: How much the odds of $Y = 1$ increases when X increases by one unit
- Example: Probability of armed conflict as a function of resource scarcity

Scarcity	No conflict	Conflict
No	0.9	0.1
Yes	0.8	0.2

UPPSALA UNIVERSITET

The logistic regression model III

- Odds of conflict if no scarcity: $o_1 = \frac{0.1}{0.9} = 0.011$
- Odds of conflict if scarcity: $o_1 = \frac{0.2}{0.8} = 0.025$
- Odds ratio: $OR = \frac{0.25}{0.11} = 2.25$
- In log odds form: $ln(o_1) = -2.197$, $ln(o_2) = -1.386$
- Log odds ratio: $ln(OR) = ln(o_2) - ln(o_1) = 0.811$

UPPSALA UNIVERSITET

Application: Urdal (2005)

- Does population increase/dense populations and associated pressure on natural resources lead to a higher risk of internal armed conflict?
- Cross-national time-series study covering 1950–2000
- Logistic regression model

UPPSALA UNIVERSITET

Application: Urdal (2005)

Table I. Risk of Armed Conflict by Neo-Malthusian Population Pressure Variables

	Full sample	*Restricted sample*	*Full sample*			
	Model 1	*Model 2*	*Model 3*	*Model 4 all onsets*	*Model 5*	*Model 6 onset 2*
Explanatory variables	*β st. error*	*β st. error*	*β st. error*	*β st. error*	*β st. error*	*β st. error*
Population pressure variables						
Population growth[a]	–0.009	0.074	–0.020	0.003	–0.013	–0.019
	(0.062)	(0.071)	(0.062)	(0.058)	(0.071)	(0.063)
Population density[a]	–0.088*	0.002	–0.156***	–0.074	–0.068	–0.113**
	(0.053)	(0.061)	(0.052)	(0.049)	(0.060)	(0.055)
Growth * density[a]	0.042	–0.017	0.061	0.041	0.014	0.081**
	(0.039)	(0.050)	(0.041)	(0.036)	(0.045)	(0.037)
Urban growth					–0.025	
					(0.041)	
Control variables						
Total population[a]	0.269***	0.207***	0.266***	0.323***	0.289***	0.285***
	(0.047)	(0.055)	(0.047)	(0.043)	(0.055)	(0.050)
Dependency	–0.890**		–0.663	–1.167***	–0.855	–0.933**
	(0.381)		(0.716)	(0.354)	(0.538)	(0.394)
Infant mortality rate[a]	0.006***	0.006***		0.006***	0.010***	0.006***
	(0.001)	(0.002)		(0.001)	(0.002)	(0.002)
GDP per capita (Ln)			–0.663***			
			(0.102)			
Missing GDP data			0.408			
			(0.729)			
Regime	0.006	0.005	0.009	0.003	0.015	0.011
	(0.014)	(0.014)	(0.014)	(0.012)	(0.015)	(0.014)
Regime, squared	–0.014***	–0.013***	–0.013***	–0.014***	–0.014***	–0.015***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Missing regime data	–0.259	–0.009	–0.313	–0.114	–0.311	–0.235
	(0.314)	(0.331)	(0.317)	(0.274)	(0.346)	(0.332)
Economic growth					–0.054**	
					(0.024)	
Missing economic growth data					0.296	
					(0.245)	
Controls for statistical dependency						
Brevity of peace	1.819***	1.725***	1.763***		1.691***	1.124***
	(0.275)	(0.285)	(0.278)		(0.304)	(0.325)
Brevity of conflict				1.366***		
				(0.318)		
Ongoing conflict in country[a]				–1.218***		
				(0.351)		
Ongoing conflict * total population[a]				0.304**		
				(0.118)		
Constant	–6.078***	–5.385***	–2.273**	–3.845***	–6.302***	–6.157***
	(0.488)	(0.569)	(0.944)	(0.206)	(0.599)	(0.513)
N	7,752	5,490	8,065	8,691	5,851	7,730
Log likelihood	–793.33	–700.47	–795.95	–963.45	–631.85	–733.86
Pseudo R^2	0.107	0.080	0.112	0.096	0.113	0.089

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
[a] Singular terms are centered to avoid multicollinearity when introducing interaction terms (Kleinbaum, Kupper & Muller, 1998: 206–212).

UPPSALA UNIVERSITET

Application: Urdal (2005)

	Full sample	*Restricted sample*	*Full sample*			
	Model 1	*Model 2*	*Model 3*	*Model 4 all onsets*	*Model 5*	*Model 6 onset 2*
Explanatory variables	*β* *st. error*	*β* *st. error*	*β* *st. error*	*β* *st. error*	*β* *st. error*	*β* *st. error*
Population pressure variables						
Population growth[a]	–0.009	0.074	–0.020	0.003	–0.013	–0.019
	(0.062)	(0.071)	(0.062)	(0.058)	(0.071)	(0.063)
Population density[a]	–0.088*	0.002	–0.156***	–0.074	–0.068	–0.113**
	(0.053)	(0.061)	(0.052)	(0.049)	(0.060)	(0.055)
Growth * density[a]	0.042	–0.017	0.061	0.041	0.014	0.081**
	(0.039)	(0.050)	(0.041)	(0.036)	(0.045)	(0.037)
Urban growth					–0.025	
					(0.041)	
Control variables						
Total population[a]	0.269***	0.207***	0.266***	0.323***	0.289***	0.285***
	(0.047)	(0.055)	(0.047)	(0.043)	(0.055)	(0.050)
Dependency	–0.890**		–0.663	–1.167***	–0.855	–0.933**
	(0.381)		(0.716)	(0.354)	(0.538)	(0.394)
Infant mortality rate[a]	0.006***	0.006***		0.006***	0.010***	0.006***
	(0.001)	(0.002)		(0.001)	(0.002)	(0.002)
GDP per capita (Ln)			–0.663***			
			(0.102)			

Application: Urdal (2005)

- Interpretation of estimate for Infant mortality rate (IMR) in model 1; $\hat{\beta} = 0.006$:
  - Log odds of armed conflict increases by 0.006 when IMR increases by one unit
  - Log odds of armed conflict increases by 0.6 when IMR increases from 20 to 120
  - Odds of armed conflict increases by a factor of $exp(0.6) = 1.82$, or by 82%
- Interpretation of estimate for Population density (PD) in model 1; $\hat{\beta} = -0.088$:
  - Log odds of armed conflict increases by 0.006 when IMR increases by one unit
  - Log odds of armed conflict increases by 0.6 when IMR increases from 20 to 120
  - Odds of armed conflict increases by a factor of $exp(0.6) = 1.82$, or by 82%

The logistic regression model I

- An R function to compute the inverse logit:

```
inv.logit <- function(p){
 return(exp(p)/(1+exp(p)))
}
```

- Simulating the logit model:

```
Y <- rbinom(n, 1, inv.logit(b0 + b1*X))
```

- Estimating the logit model:

```
model <- glm(Y ~ X, family = binomial (link =
logit))
```

- Note that there is no error term in the model

UPPSALA UNIVERSITET

Maximum likelihood estimation I

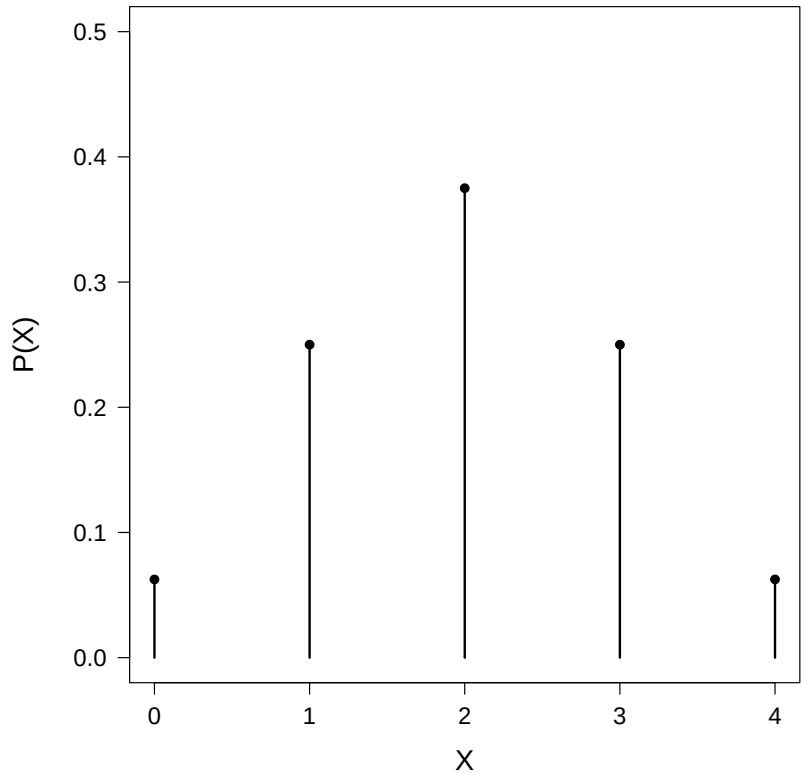

Figure 1: PDF for number of heads = (0,1,2,3,4) out of four, given $p = 0.5$

- Assume we toss a coin four times and obtain one head out of four
- We then have a dataset $y = (0, 1, 0, 0)$
- What is the probability of obtaining this if $\theta = p = 0.5$?

UPPSALA UNIVERSITET

Maximum likelihood estimation I

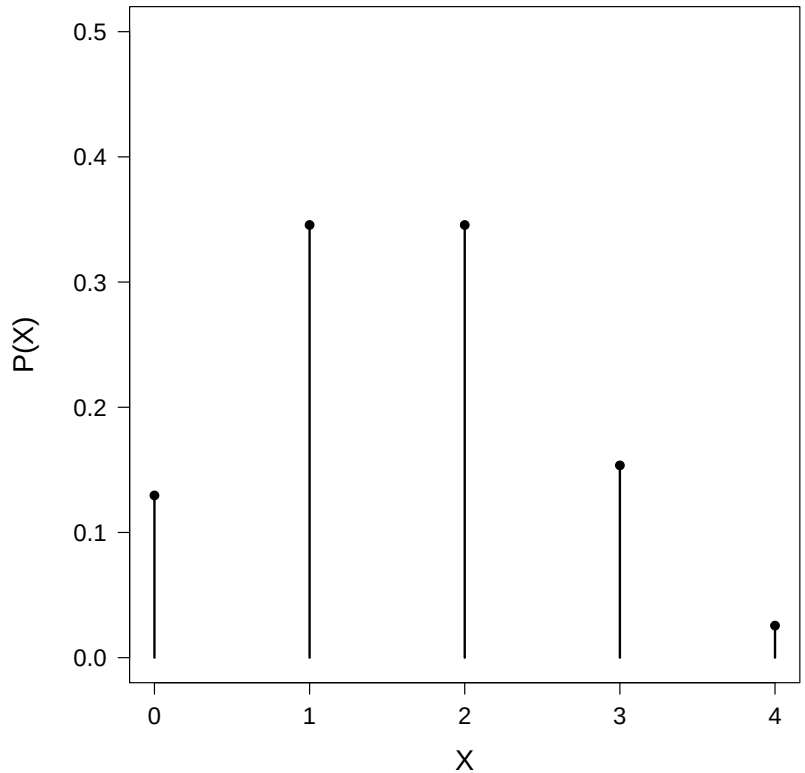

Figure 2: PDF for number of heads = (0,1,2,3,4) out of four, given $p = 0.4$

- The dataset would be more likely if $\theta = 0.4$ ($Pr(y) = 0.346$ rather than $Pr(y) = 0.25$)
- We can estimate a parameter by finding the parameter value that maximizes the probability of observing our data
- To do so, we must assume a model (binomial in this case)

UPPSALA UNIVERSITET

Maximum likelihood estimation I

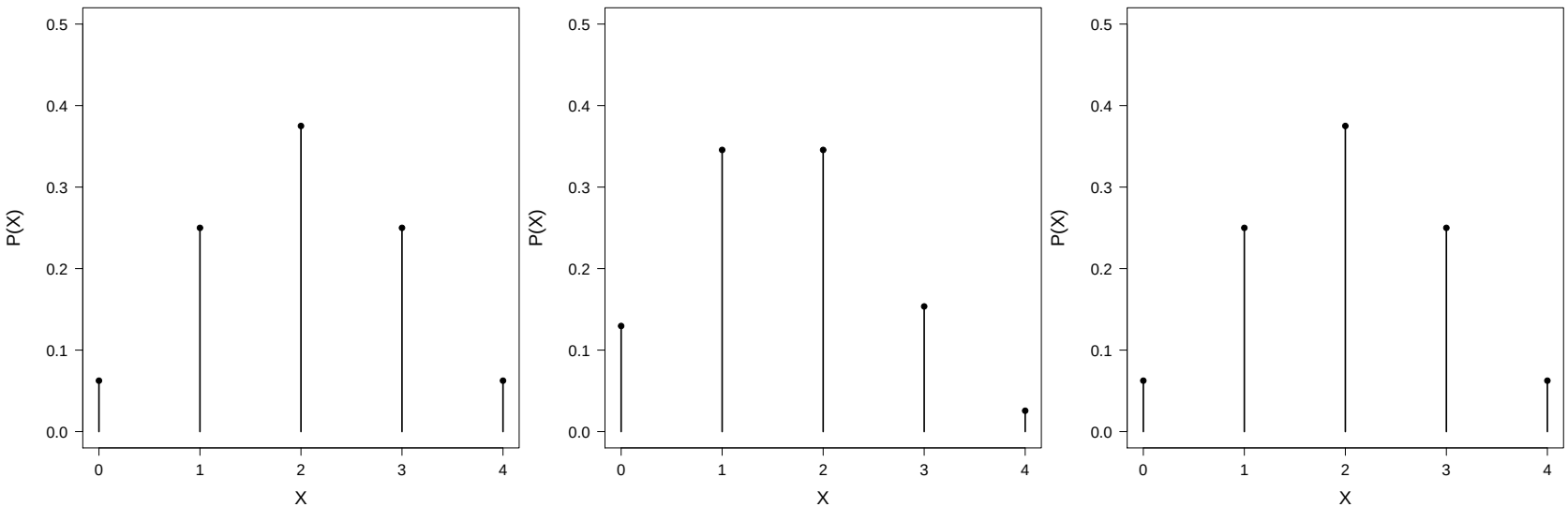

Figure 3: The probability density function (PDF) for number of heads = (0,1,2,3,4) out of four, given $p = 0.5$ (left), $p = 0.4$ (middle), $p = 0.25$ (right)

- The Pr(y) is highest at $\theta = 0.25$

UPPSALA UNIVERSITET

Maximum likelihood estimation II

- Define a likelihood function:

$$L(\tilde{\theta}|y, M^*) = k(y)Pr(y|\tilde{\theta})$$

- The likelihood of a given parameter value $\tilde{\theta}$ depends on the data y and on our model M^* that implies a probability for observing a given dataset if the parameter is true
- For instance, we have calculated from the binomial distribution for $N = 4$ that the probability of obtaining (y: 0 heads out of four tosses) is 0.0625 if $p = 0.5$
- Maximum likelihood estimation implies searching for the parameter value that maximizes the likelihood of generating these data

UPPSALA UNIVERSITET

Predictions

- Predicted probabilities are calculated using the inverse logit function:

```
inv.logit <- function(p){
 return(exp(p)/(1+exp(p)))
}
```

- In order to predict the dichotomous Y, however, we need to define a threshold
  - A value over which we say we expect the event to happen
  - $p(Y = 1) = 0.5$ is a natural threshold
  - But others can also be useful

UPPSALA UNIVERSITET

Example confusion matrix

- From Hegre et al. (2019) (ViEWS project; *cm* ensemble model)
- Model trained on data 1990–2014, evaluated against test partition: 2015–2017

	Observed		
Predicted	Pos	Neg	Sum
Pos	437	282	719
Neg	17	1208	1225
Sum	454	1490	1944

Note. State-based conflict at cm-level, January 2015 to December 2017. Accuracy = 0.846, F1 = 0.745, precision = 0.608, recall = 0.963, threshold = 0.126.

UPPSALA UNIVERSITET

Example confusion matrix

	Observed		
Predicted	Pos	Neg	Sum
Pos	437	282	719
Neg	17	1208	1225
Sum	454	1490	1944

- In the evaluation data, 454 of 1944 country months with at least one fatality
- Predicting $Y = 1$ if $p(Y = 1) > 0.126$
- This yields 719 positive predictions

UPPSALA UNIVERSITET

PR curves

By FeanDoe - Modified version from Walber's Precision and Recall https://commons.wikimedia.org/wiki/File:Precisionrecall.svg, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=65826093

UPPSALA UNIVERSITET

Example confusion matrix

	Observed		
Predicted	Pos	Neg	Sum
Pos	437	282	719
Neg	17	1208	1225
Sum	454	1490	1944

True positives (TP):

- Observed positives that were correctly predicted

True negative (TN):

- Observed negatives that were correctly predicted

False positives (FP):

- Observed negatives that were predicted as positives

False negatives (FN):

- Observed positives that were predicted as negatives

UPPSALA UNIVERSITET

Sensitivity, specificity, recall

Sensitivity: $\frac{TP}{TP+FN}$:

- Number of true positive predictions divided by number of actual positives
- Also called Recall or True Positive Rate
- Here, $437/454 = 0.963$

Specificity: $\frac{TN}{FP+TN}$:

- Number of true negatives divide by number of actual negatives
- Also called True Negative Rate
- Here, $1208/1490 = 0.811$

Precision: $\frac{TP}{TP+FP}$:

- Number of true positive predictions divided by total number of predicted positives
- Also called Positive Predictive Value
- Here, $437/719 = 0.608$

UPPSALA UNIVERSITET

Sensitivity, specificity, recall

Accuracy:

- Proportion of true predictions of all predictions:
  $A = \frac{TP+TN}{TP+TN+FP+FN} = \frac{437+1208}{437+17+282+1208} = 0.846$
- Accuracy will be high if FN is high

F_1 measure – the harmonic mean of precision and recall:

- $F_1 == 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

$$F_1 = 2 \cdot \frac{0.608 \cdot 0.963}{0.608 + 0.963} = 0.799$$

UPPSALA UNIVERSITET

Loss functions

- These metrics assume losses are dichotomous: correct or incorrect
- Classification loss is either 0 or 1
- But what is best when evaluating an actual positive: a prediction of 0.49, or one of 0.01?
- A quadratic loss function (for outcome variable with a_k categories):

$$\sigma_j(p_j - a_j)^2$$

- Brier score:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

UPPSALA UNIVERSITET

ROC curves

- ROC curves depict the performance of a classifier without regard to class distribution or error costs
- A confusion matrix depends on your choice of cut-off for distinguishing countries with conflict from those without
- The higher the cut-off, the lower sensitivity
- The higher the cut-off, the higher specificity
- But the number of actual positives is the same
- Plots the True Positive Rate = 1 - True Negative Rate on the vertical axis against True Negative Rate on the horizontal, for all possible cut-off probabilities

UPPSALA UNIVERSITET

PR curves

- Plots the True Positive Rate on the vertical axis against True Negative Rate on the horizontal

UPPSALA UNIVERSITET

Bibliography I

Carsey, Thomas M. and Jeffrey J. Harden. 2014. *Monte Carlo Simulation and Resampling. Methods for Social Science*. Los Angeles, London, and New Dehli: Sage.

Hegre, Håvard, Marie Allansson, Matthias Basedau, Mike Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayeed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull and Jonas Vestby. 2019. "ViEWS: A political Violence Early Warning System." *Journal of Peace Research* 56(2):155–174.
URL: *https://doi.org/10.1177/0022343319823860*

Urdal, Henrik. 2005. "People vs. Malthus: Population Pressure, Environmental Degradation and Armed Conflict Revisited." *Journal of Peace Research* 42(4):417–434.