# 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 15, 2019

UPPSALA UNIVERSITET

# Count models I

- Count data: Data that can equal (0, 1, 2, ...)
  - Number of traffic accidents in a location and/or time interval
  - Number of battle-related deaths in a country during a time period
- Poisson regresson model:

$$y_i \sim Poisson(\theta_i)$$

where

$$\theta_i = exp(X_i\beta)$$

UPPSALA UNIVERSITET

# Count models II

- Example from Gelman & Hill:

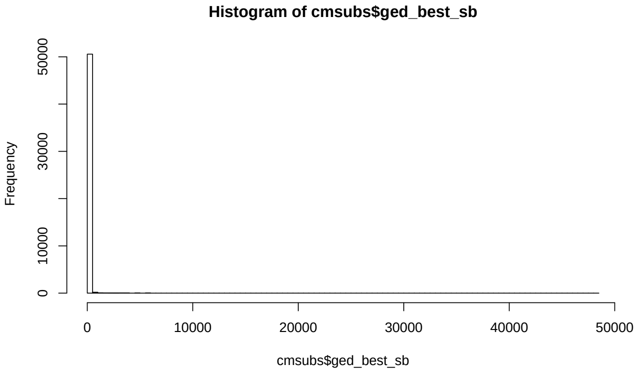
$$y_i \sim Poisson(exp(2.8 + 0.012X_{1i} - 0.20X_{2i})$$

  where $X_1$ is speed limit at intersection and $X_2$ has the value 1 if there is a traffic signal
- Constant term/intercept of model: Prediction if $X_{1i} = X_{2i} = 0$
- The coefficient of $X_{1i}$ is the expected difference in $y$ (on a logarithmic scale) for each additional mph of speed
- Expected multiplicative increase is $e^{0.012} = 1.012$, or a 1.2% positive difference in the rate of traffic accidents per mph

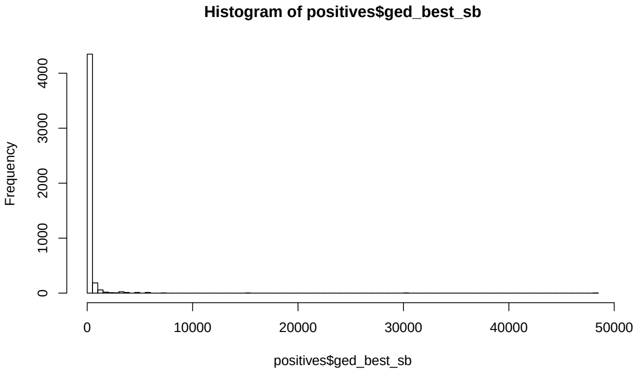
# Count models III

- Coefficient of $X_{2i}$: the predictive difference of having a traffic signal is found by multiplying the accident rate by $exp(-0.20) = 0.82$ – a reduction of 18%
- If the mean of the Poisson process is relatively high, OLS models of log counts perform well

# Distribution of battle-related deaths, country months, zeros included I



# Distribution of battle-related deaths for months with at least 5 deaths I



UPPSALA UNIVERSITET

# Overdispersion I

- Challenge: Many zeroes, high variance
- Fatalities in war may have a power-law distribution
- Under Poisson distribution, variance is equal to the mean
- Mathematically, $E(y_i) = u_i\theta_i$ and $sd(y_i) = \sqrt{u_i\theta_i}$
- Standardizing, residuals are

$$z_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)}$$

- If the Poisson model is true, the standardized residuals should have mean 0 and standard deviation 0
- If there is **overdispersion**, the standard deviation of the standardized residuals is larger than 1

UPPSALA UNIVERSITET

# Overdispersion II

- Script testing for over-dispersion in Gelman & Hill p. 115
- Overdispersed-Poisson or Negative binomial model:

$$y_i \sim overdispersedPoisson(u_i exp(X_i\beta), \omega)$$

- where $\omega$ is the overdispersion parameter

UPPSALA UNIVERSITET

# Zero-inflated models I

- Even more zeroes than in an overdispersed Poisson?
- A DGP with two separate systematic processes:
  1. A process deciding whether an observation produces a zero or a positive count
  2. A process deciding the actual count
- Variants:
  1. zero-inflated Poisson
  2. zero-inflated Negative binomial
  3. zero-inflated OLS (hurdle model)

UPPSALA UNIVERSITET

# Vectors, matrices, arrays

- A vector has one dimension
- A matrix has two dimensions
- An array can have more than two dimensions
- Example: A 2x2x2 array: `box2 <- array(NA, c(2,2,2))`
- `box2 <- array(1:4, c(2,2,2))`
- Useful to store a number of similar matrices:
- `par.est.merror[ , , j] <- par.est` stores parameter estimate matrix number $j$ in the `par.est.merror` array

UPPSALA UNIVERSITET

# Serial correlation I

- Serial correlation: where one or more of a model's errors influence one or more other errors
- If we have time-series data, with one observation for each point in time $t$, and $Y_t = \beta_0 + \beta_1 X_1 + \epsilon_t$
- Autoregressive (AR) and moving average (MA) processes:
  - AR(1) process: $\epsilon_t = \rho\epsilon_{t-1} + u_t$
  - MA(1) process: $\epsilon_t = u_t - \lambda u_{t-1}$
  - AR(2) process: $\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + u_t$
- Simulating in `R`: Replace the `rnorm` function with `arima.sim(list(order=c(1, 0, 0), ar = ac), n= n)`
- `list(order=c(1, 0, 0), ar = ac)` specifies an AR(1) process with $\rho$=ac MC experiment:

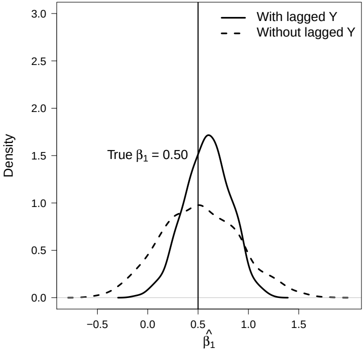
UPPSALA UNIVERSITET

# Serial correlation II

- Compare two models given an autocorrelated DGP
- One without and one which includes a lagged dependent variable
  - $\hat{Y} = \beta_0 + \beta_1 X_{1,t}$ which ignores error structure
  - $\hat{Y} = \beta_0 + \beta_1 X_{1,t} + \beta_2 Y_{t-1}$ which partially corrects for the serial correlation but introduces some bias due to the stochastic component of $Y_t$
- Autocorrelation may also cause over-confidence

# Serial correlation III

Figure 1: The distribution of $\beta_1$ estimates with and without lagged $Y$, $\rho = .75$



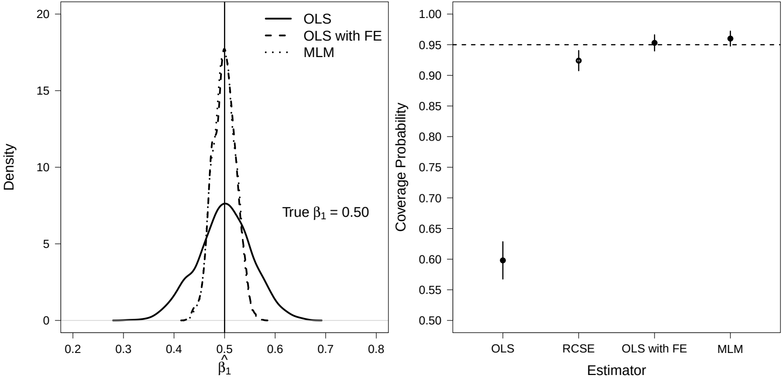
UPPSALA UNIVERSITET

# Clustered data I

- Clustering occurs if the errors within groups of observations are correlated
  - regions within countries
  - students within schools
  - repeated observations for the same actors (TSCS data)
- Leads to:
  - Overconfidence
  - Omitted variable bias

# Clustered data II

Figure 2: Comparison of estimators for clustered data



UPPSALA UNIVERSITET

# Handling clustered data I

- Strategies for handling it:
  - Naïve OLS, ignoring the problem
  - OLS with 'fixed effects'
  - Adjusting OLS standard errors
    - Clustered standard errors
    - Bootstrapping
  - Multilevel modeling (random effects)

UPPSALA UNIVERSITET

# Clustered standard errors

- If observations are clustered, they are not truly independent
- Even in estimates of $\beta$ may be unbiased, estimates of standard errors will be unless the source of clustering is accounted for
- Clustered standard errors are calculated according to a formula assuming no residual correlation between groups, but accounting for correlation within groups
- See Stock and Watson (2007, p. 379-382 for a detailed, technical explanation)
- Also see `http://www.ne.su.se/polopoly_fs/1.216115.1426234213!/menu/standard/file/clustering1.pdf`

UPPSALA UNIVERSITET

# Omitted variable bias (repetition) I

- OVB: When omitting a relevant independent variable that is part of the true DGP from the statistical model that is estimated
- A problem only if $X_o$ is correlated with both $X$ and $Y$
  - Causes bias to regression coefficients because it produces correlation between $X$ and the error term $\epsilon$
- A special case: unobserved unit heterogeneity (Green, Kim and Yoon, 2001)

# Omitted variable bias (repetition) II

UPPSALA UNIVERSITET

# Green, Kim and Yoon (2001): 'Dirty pool' I

- Background: time-series cross-section data, e.g. collecting data on a number of countries every year over 40 years
  - or all pairs of countries every year
  - or any other form of clustered observations, e.g. administrative units within countries, pupils within classes
- 'Fixed unobserved differences': unmeasured predictors of the dependent variable that would cause each dyad to have its own base rate
  - For instance the cultural and geographical relationship between Norway and Sweden (or Argentina and Uruguay, or Senegal and Gambia)

UPPSALA UNIVERSITET

# Green, Kim and Yoon (2001): ‘Dirty pool’ II

- Ignoring ‘fixed effects’ leads to a special case of omitted variables
  - May lead to bias (if the omitted variables are correlated with both $X$ and $Y$)
  - But here also to over-confidence since residuals are not independent within clusters

UPPSALA UNIVERSITET

# Green, Kim and Yoon (2001): 'Dirty pool' III

- The gravity model of trade:

$$T_{ij} = \beta_0 \times \frac{GDP_i^{\beta_1} \times GDP_j^{\beta_2}}{Distance_{ij}^{\beta_3}}$$

- In log form:

$$ln(T_{ij}) = \beta_0 + \beta_1 ln(GDP_i) + \beta_2(GDP_j) - \beta_3(Distance_{ij})$$

- Add repeated observations:

$$ln(T_{ijt}) = \beta_0 + \beta_1 ln(GDP_{it}) + \beta_2(GDP_{jt}) - \beta_3(Distance_{ijt})$$

- Since distance between states is constant, that variable is not distinguishable from the dyad-specific intercepts

# From ‘Dirty Pool’: Pooled DGP

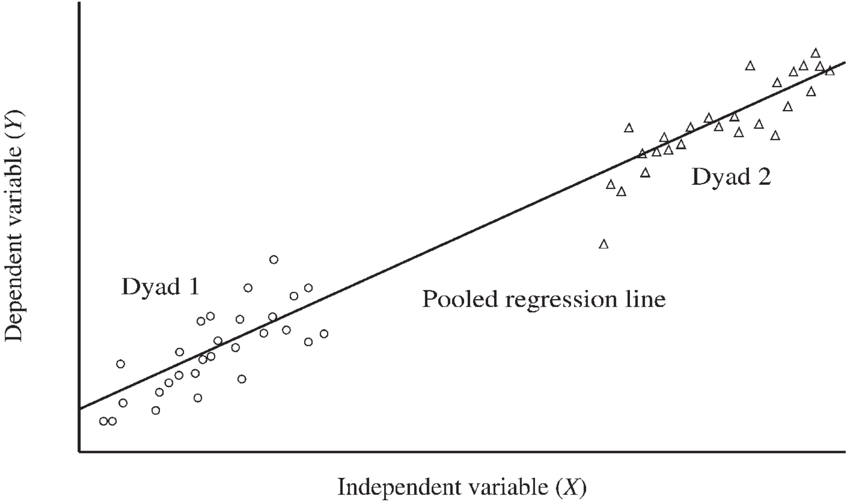



**FIGURE 1.** *Pooling homogenous observations*

# From 'Dirty Pool': Non-pooled DGP

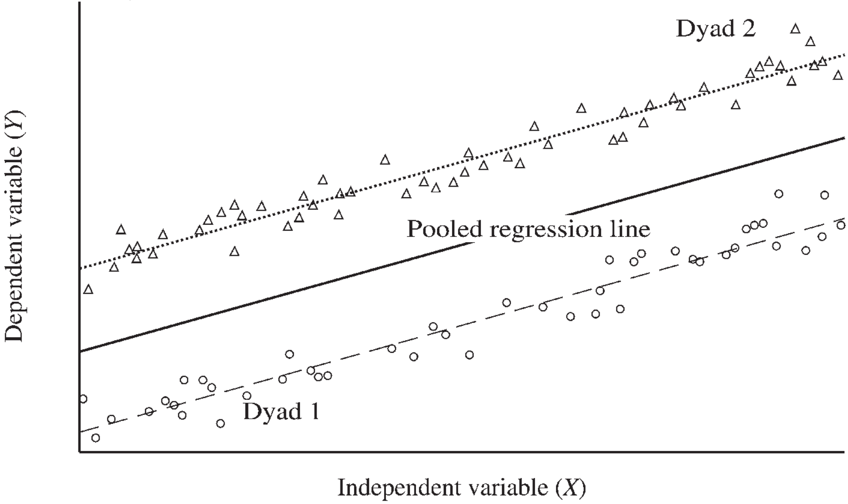



**FIGURE 2.** *Pooling dyads with randomly varying intercepts*

UPPSALA UNIVERSITET

# From ‘Dirty Pool’: Correlation with X

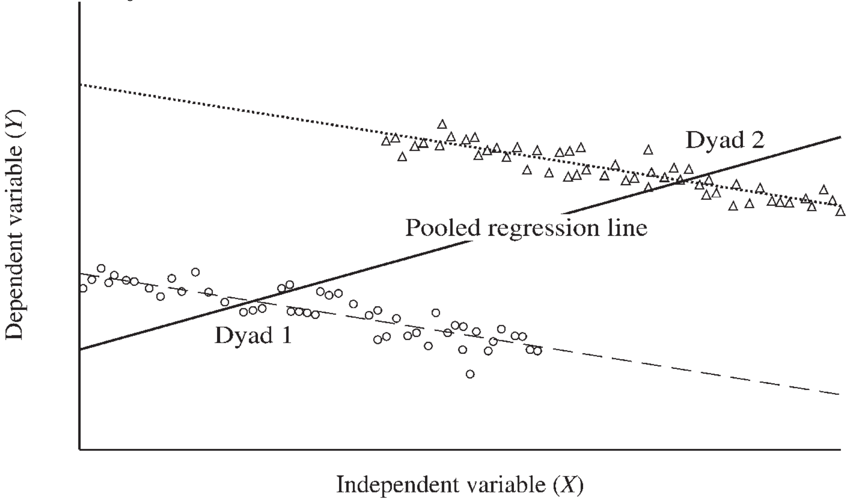



**FIGURE 3.** *Pooling observations ignoring fixed effects*

UPPSALA UNIVERSITET

# 'Dirty pool': Table 2, bilateral trade

| *Variable*[a] | *Pooled* | *Fixed effects* | *Pooled with dynamics* | *Fixed effects with dynamics* |
|---|---|---|---|---|
| GDP | 1.182** | 0.810** | 0.250** | 0.342** |
| | (0.008) | (0.015) | (0.006) | (0.013) |
| Population | −0.386** | 0.752** | −0.059** | 0.143* |
| | (0.010) | (0.082) | (0.006) | (0.068) |
| Distance | −1.342** | Dropped: no within-group variation | −0.328** | Dropped: no within-group variation |
| | (0.018) | | (0.012) | |
| Alliance | −0.745** | 0.777** | −0.247** | 0.419** |
| | (0.042) | (0.136) | (0.027) | (0.121) |
| Democracy[b] | 0.075** | −0.039** | 0.022** | −0.009** |
| | (0.002) | (0.003) | (0.001) | (0.002) |
| Lagged bilateral trade | | | 0.736** | 0.533** |
| | | | (0.002) | (0.003) |
| Constant | −17.331** | −47.994** | −3.046** | −13.745** |
| | (0.265) | (1.999) | (0.177) | (1.676) |
| | $N =$ 93,924 | $NT =$ 93,924<br>$N =$ 3,079<br>$T \geqslant 20$ | $N =$ 88,946 | $NT =$ 88,946<br>$N =$ 3,079<br>$T \geqslant 20$ |
| Adjusted $R^2$ | 0.36 | 0.63 | 0.73 | 0.76 |

*Note:* Estimates obtained using *areg* and *xtreg* procedures in STATA, version 6.0.
[a]GDP, population, distance, and bilateral trade are natural-log transformed. Method of analysis is OLS and fixed-effects regression.

# Some notation

- Index for unit: $i, i = 1, ..., N$
- Index for point in time: $t, t = 1, ..., T$
- $Y_{it}$ is observed $Y$ for unit $i$ at time $t$
- $X_{it}$ is observed $X$ for unit $i$ at time $t$
- $X_{1,it}$ is observed $X_1$ for unit $i$ at time $t$
- In a **balanced panel** we have obervations for all $i$ and all $t$
- In a **unbalanced panel** observations are missing, e.g. the three first time points for unit no. 14

UPPSALA UNIVERSITET

# Degree of pooling

Clustered data:

$$Y_{jt} = \beta_j Z_j + \beta_1 X_j t + \epsilon_{jt}$$

- Complete pooling:

$$Y_{jt} = \beta_0 + \beta_1 X_j t + \epsilon_{jt}$$

- No pooling:

$$Y_{jt} = \beta_j Z_j + \beta_{1j} X_j t + \beta_{xj} X_j t * Z_j + \epsilon_{jt}$$

- Partial pooling:

$$Y_{jt} = \beta_j Z_j + \beta_1 X_j t + \epsilon_{jt}$$

# Fixed-effects model

$$Y_{jt} = \beta_j Z_j + \beta_1 X_j t + \epsilon_{jt}$$

- Each group $j$ has its own intercept term $Z_j$
- Here, each group is estimated repeatedly over time $t$, but can also be clustered differently
- 'Time-invariant' group intercepts
- The interpretation of $\beta$ estimates change subtly when you include fixed effects – they are now within-estimates only
  - Estimated change in $Y$ when $X$ increases by one *holding observation unit* $Z_j$ *constant*

UPPSALA UNIVERSITET

# How conflict affects infant mortality rates

Table 1: FE model, Infant Mortality Rates , 1970–2008

| | OLS, 2005 | OLS | OLS, td | FE | FE, td |
|---|---|---|---|---|---|
| Log fatalities | 0.092*** | 0.0552*** | 0.0506*** | 0.00431 | 0.0129*** |
| | (0.0269) | (0.00830) | (0.00765) | (0.00381) | (0.00280) |
| Log population | 0.0433 | 0.0242 | 0.0405* | -1.016*** | 0.331*** |
| | (0.0526) | (0.0178) | (0.0164) | (0.0351) | (0.0545) |
| 1975-79 | | | -0.133 | | -0.196*** |
| | | | (0.114) | | (0.0304) |
| 1980-84 | | | -0.325** | | -0.423*** |
| | | | (0.111) | | (0.0319) |
| 1985-89 | | | -0.452*** | | -0.632*** |
| | | | (0.111) | | (0.0352) |
| 1990-94 | | | -0.620*** | | -0.813*** |
| | | | (0.110) | | (0.0387) |
| 1995-99 | | | -0.809*** | | -0.965*** |
| | | | (0.106) | | (0.0424) |
| 2000-04 | | | -0.969*** | | -1.148*** |
| | | | (0.107) | | (0.0460) |
| 2005-08 | | | -1.105*** | | -1.327*** |
| | | | (0.106) | | (0.0493) |
| Constant | 2.770*** | 3.510*** | 3.977*** | 12.78*** | 1.644*** |
| | (0.464) | (0.153) | (0.161) | (0.310) | (0.458) |
| Observations | 151 | 1043 | 1043 | 1043 | 1043 |
| Log likelihood | -212.2 | -1378.4 | -1286.6 | -145.5 | 190.4 |
| $\chi^2$ | | | | | |

Standard errors in parentheses

UPPSALA UNIVERSITET

# Random-effects model I

Some notational controversy: Gelman and Hill (2007):

- ‘fixed-effects’ models (usually) defined as models where the unit-level coefficients are not themselves modeled
- ‘random-effects’ models: unit-level coefficients are modeled
- Gelman and Hill (2007) refer to the latter as ‘multi-level’ models

$$Y_{jt} = Z_j + \beta_1 X_{jt} + \epsilon_{jt}$$

where $Z_j \sim N(0, \sigma_Z^2)$ and $\epsilon_{jt} \sim N(0, \sigma_\epsilon^2)$

# Random-effects model II

- In the notation of Gelman and Hill (2007):

$$y_i = \alpha_{j[i]} + \beta X_i + \epsilon_i$$

or

$$y_i \sim N(\alpha_{j[i]} + \beta X_i, \sigma_i^2), \alpha_j \sim N(\mu_j, \sigma_\alpha^2)$$

# Simulating clustered data

Modeling of this in the MC analysis in the book:

- $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}$
- Four uncorrelated 'effects':
  1. $E1 \sim N(0, 1), n = N$
  2. $E2 \sim N(0, 1 - p), n = N$
  3. $E3 \sim N(0, 1), n = nc = J$
  4. $E4 \sim N(0, p), n = nc = J$
- $\epsilon_{ij} = E2_{ij} + E4_j$
- $X_{ij} = E1_{ij} + E3_j$
- Both $\epsilon$ and $X$ are clustered on $j$ – values within each $j$ are more similar than values across $j$s
- $p$ specifies correlation between $X_{ij}$ and unit effects

# Bibliography I

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press.

Green, Donald P., Soo Yeon Kim and David Yoon. 2001. "Dirty Pool." *International Organization* 55(2):441–468.

Stock, James H. and Mark W. Watson. 2007. *Introduction to econometrics, 2nd ed.* Boston etc: Pearson/Addison Wesley.