Analyzing collective individual behavior

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2017
Columbia University

Housekeeping

- Today:
  - time-series cross-section analysis
  - Final team progress report
- next week: **Guest speaker:**
- week after that: **Final Presentations**

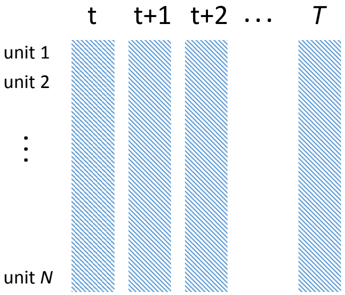
Analyzing individual behaviors collectively

- A number of human behaviors happen repeatedly over time
  - we may want to exploit that the past can predict the future (T)
  - we may also want to leverage correlations in contemporaneous behaviors (N)
- many problems involve drawing inferences from both dimensions simultaneously
- ML is only beginning to explore the time dimension

Models to deal with T and N simultaneously

Panel models

- large N, small T
- a lot of people with few observations over time

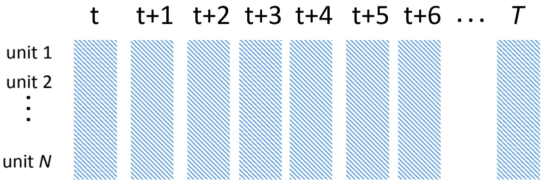


$N > T$

Models to deal with T and N simultaneously

Time-Series Cross-Section models

- small N, large T
- a few people with a lot of observations over time



$N < T$

Modeling Time-Series Cross-Section data

the time (T) dimension

- what you know about time-series applies to TSCS data
  - think of each unit i as having its own time series
- start with a general model - ADL(1,1;1) for illustration purposes - and let the data "tell" the correct specification

$$\mathbf{Y}_{i,t} = \alpha_0 + \mathbf{Y}_{i,t-1}\alpha_1 + \mathbf{X}_{i,t}\beta_0 + \mathbf{X}_{i,t-1}\beta_1 + \mathbf{Z}_i\psi + \epsilon_{i,t}$$

where:

$\mathbf{Y}_{i,t}$: DV at time t
$\mathbf{Y}_{i,t-1}$: DV at time $t-1$
$\mathbf{X}_{i,t}$: exogenous regressor of interest at time t
$\mathbf{X}_{i,t-1}$: exogenous regressor of interest at time $t-1$
$\mathbf{Z}_i$: other exogenous regressors

Modeling Time-Series Cross-Section data

the time (T) dimension

- the "correct" specification is determined by testing:

Table 1 Restrictions of the ADL General Dynamic Model

Type	ADL Model	Restriction
General	$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$	None
Partial Adjustment*	$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_0 X_t + \varepsilon_t$	$\beta_1 = 0$
Static[a]	$Y_t = \alpha_0 + \beta_0 X_t + \varepsilon_t$	$\alpha_1 = \beta_1 = 0$
Finite Distributed Lag[b]	$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$	$\alpha_1 = 0$
Differences[c]	$\Delta Y_t = \alpha_0 + \beta_0 \Delta X_t + \varepsilon_t$	$\alpha_1 = 1, \beta_0 = -\beta_1$
Dead Start	$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_1 X_{t-1} + \varepsilon_t$	$\beta_0 = 0$
Common Factor[d]	$Y_t = \beta_0 X_t + \varepsilon_t, \varepsilon_t = \beta_1 \varepsilon_{t-1} + u_t$	$\beta_1 = -\beta_0 \alpha_1$

*Also known as the Koyck model.
[a] $k_1 = \beta_0$; Dynamic effects at lags beyond zero constrained to be zero.
[b] $k_1 = \sum_{j=1}^{n} \sum_{i=0}^{q-1} \beta_{ji}$.
[c] Infinite mean lag length.
[d] $k_1 = \beta_0$, $\mu = 0$, EC rate 100%.

Figure: De Boef et al. (2008)

Modeling Time-Series Cross-Section data

the cross-sectional (N) dimension

- how much unit heterogeneity is warranted by the data?
  - a **(pooled) model**

$$\mathbf{Y}_{i,t} = \mathbf{X}_{i,t}\beta + \epsilon_{i,t}$$

  - a **fixed effects model**

$$\mathbf{Y}_{i,t} = \mathbf{X}_{i,t}\beta + f_i + \epsilon_{i,t}$$

  - a **random coefficients model**

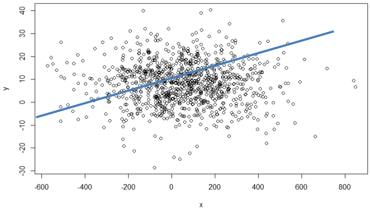
$$\mathbf{Y}_{i,t} = \mathbf{X}_{i,t}\beta_i + \epsilon_{i,t}$$

- TSCS allows modeling the process as equal for all units

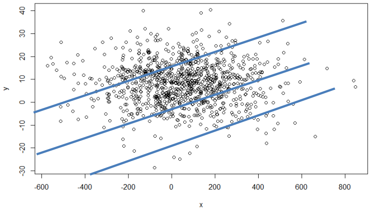
Modeling Time-Series Cross-Section data

the cross-sectional (N) dimension

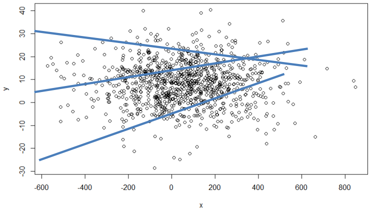
Pooled model



Fixed effects model



Random coefficients model



Modeling Time-Series Cross-Section data

interesting quantities of interest

- **Speed of adjustment** (change of Y_t in subsequent periods)

$$s = 1 - \alpha_1$$

- **Immediate effect** (non-distributed effect of X_t on Y_t)

$$\beta_0$$

- **Long-run multiplier** (total effect of X_t on Y_t over all future periods)

$$k_1 = \frac{\beta_1 + \beta_0}{1 - \alpha_1}$$

- **Mean lag length** (periods it takes Y_t to adjust back to equilibrium)

$$\mu = \frac{\beta_1}{\beta_0 + \beta_1} - \frac{-\alpha_1}{1 - \alpha_1}$$

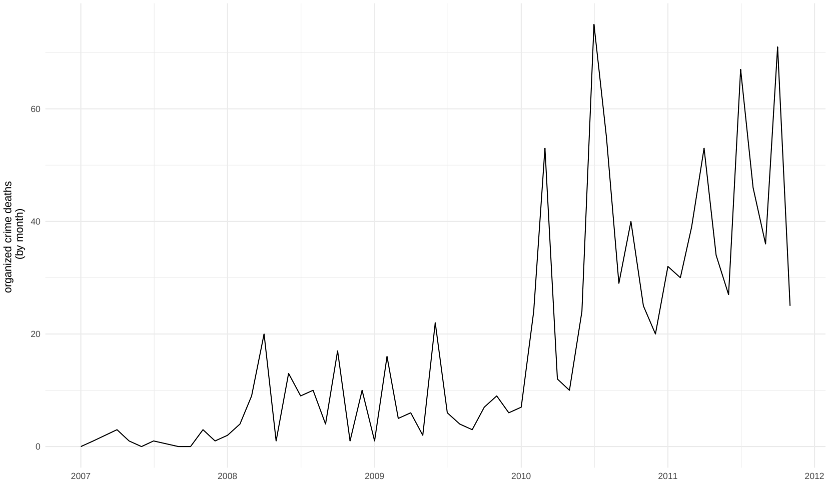
Modeling Time-Series Cross-Section data

back to our working example

- think of each **municipality** as a unit with multiple (daily) observations over time
- we may be able to explore a few interesting questions:
  - do the number of **deaths** in municipality i at time t help us understand **deaths** at time $t+1$?
  - do the number of **wounded** in municipality i at time t help us understand **deaths** at time $t+1$?
- for illustration purposes:
  - we aggregate data at the **monthly** level
  - we only chose the **top 10 most violent** municipalities

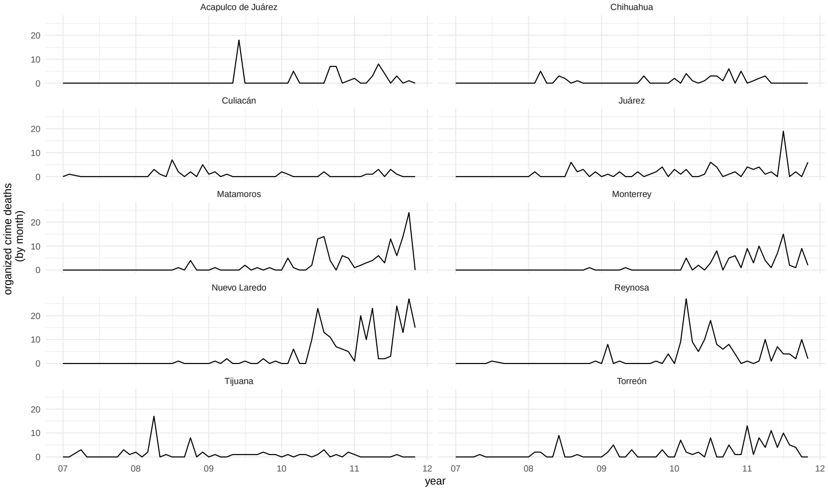
Modeling Time-Series Cross-Section data

back to our working example



Modeling Time-Series Cross-Section data

back to our working example



Modeling Time-Series Cross-Section data

back to our working example

- (simple) unit-root testing (and stationarity)

$$y_t = \alpha_1 y_{t-1} + \epsilon_t; \;\; H_0 : \alpha_1 = 1$$

```
Call:
lm(formula = organized.crime.dead ~ organized.crime.dead.L1,
    data = panel)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2405  -1.0806  -1.0806  -0.0806  22.1095

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.08059    0.16753    6.45 2.42e-10 ***
organized.crime.dead.L1  0.42333    0.03883   10.90  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 3.604 on 558 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.1756,Adjusted R-squared:  0.1741
F-statistic: 118.8 on 1 and 558 DF,  p-value: < 2.2e-16
```

Modeling Time-Series Cross-Section data

back to our working example

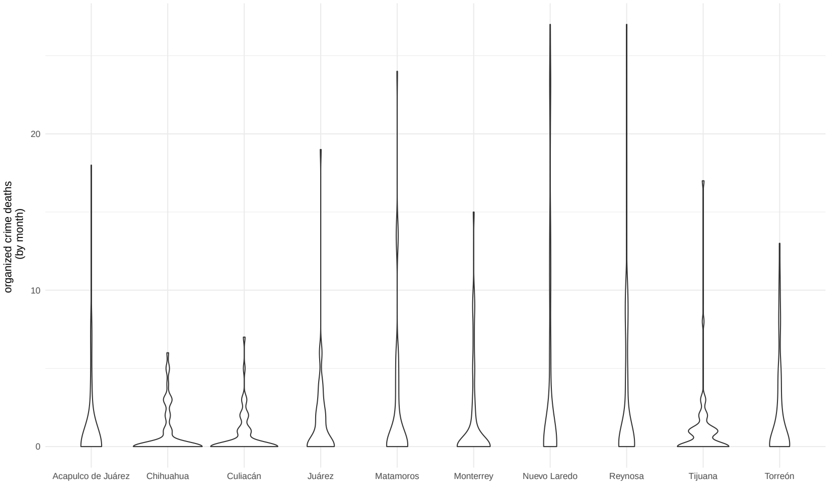
- (simple) unit-root testing (and stationarity)

$$\epsilon_t = \gamma_1 \epsilon_{t-1} + \nu_t; \quad H_0 : \gamma_1 = 1$$

```
Call:
lm(formula = res[-1] ~ res[-n])

Residuals:
    Min      1Q  Median      3Q     Max
-9.5721 -1.2508 -1.2087 -0.1875 22.8759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02070    0.15423   0.134   0.8933
res[-n]     -0.09940    0.04303  -2.310   0.0213 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 3.617 on 548 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.009644,Adjusted R-squared:  0.007837
F-statistic: 5.336 on 1 and 548 DF,  p-value: 0.02125
```

Modeling Time-Series Cross-Section data

back to our working example



Modeling Time-Series Cross-Section data

back to our working example

- a nice feature of TSCS data: can be estimated by OLS
  - estimated parameters will be consistent...
  - but inefficient if Gauss-Markov assumptions not met
- for illustration purposes, we estimated the model:

$$\mathbf{Y}_{i,t} = \alpha_0 + \mathbf{Y}_{i,t-1}\alpha_1 + \mathbf{X}_{i,t}\beta_0 + \mathbf{X}_{i,t-1}\beta_1 + \epsilon_{i,t}$$

- note that:
  - the model is (artificially) restricted to one lag
  - no municipality information is included (unavailable)
  - the data suggests an ADL(1,1;1) is appropriate

Modeling Time-Series Cross-Section data

back to our working example

```
Call:
lm(formula = organized.crime.dead ~ organized.crime.dead.L1 +
    organized.crime.wounded + organized.crime.wounded.L1, data = panel)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2499  -0.8942  -0.6366   0.2378  20.6463

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.63663    0.16271   3.913 0.000103 ***
organized.crime.dead.L1     0.43293    0.03881  11.154  < 2e-16 ***
organized.crime.wounded     0.82792    0.07754  10.678  < 2e-16 ***
organized.crime.wounded.L1 -0.23762    0.08465  -2.807 0.005174 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.282 on 556 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.319,Adjusted R-squared:  0.3153
F-statistic: 86.82 on 3 and 556 DF,  p-value: < 2.2e-16
```

- note the negative coefficient on the lag of `organized.crime.wounded`
- what does that mean?

Modeling Time-Series Cross-Section data

back to our working example

i) do the number of **deaths** in municipality i at time t help us understand **deaths** at time $t+1$?

- the answer is yes, but...
- how fast do the number of deaths change on every period? (**speed of adjustment**)

$$s = 1 - \alpha_1 = 1 - .43 = \mathbf{0.57}$$

- how many periods does it take for the number of deaths to adjust back to equilibrium? (**mean lag length**)

$$\mu = \frac{\beta_1}{\beta_0 + \beta_1} - \frac{-\alpha_1}{1 - \alpha_1} = \frac{-.23}{.82 - .23} - \frac{-.64}{1 - .43} = \mathbf{0.73}$$

Modeling Time-Series Cross-Section data

back to our working example

ii) do the number of **wounded** in municipality i at time t help us understand **deaths** at time $t + 1$?

- the answer is yes, but...
- what is the total effect of wounded on deaths on *this period*? (**immediate effect**)

$$\beta_0 = \mathbf{0.82}$$

- what is the total effect of wounded on deaths *over all periods*? (**long run multiplier**)

$$k_1 = \frac{\beta_1 + \beta_0}{1 - \alpha_1} = \frac{-.23 + .82}{1 - .43} = \mathbf{1.20}$$

Modeling Time-Series Cross-Section data

back to our working example: heterogeneity (fixed-effects model)

```
Call:
lm(formula = organized.crime.dead ~ organized.crime.dead.L1 +
    organized.crime.wounded + organized.crime.wounded.L1 + factor(municipality),
    data = panel)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6842  -1.3342  -0.3462   0.1856  18.9967

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.178746   0.438571   0.408 0.683753
organized.crime.dead.L1           0.375723   0.039992   9.395  < 2e-16 ***
organized.crime.wounded           0.854192   0.077100  11.079  < 2e-16 ***
organized.crime.wounded.L1       -0.166434   0.085109  -1.956 0.051030 .
factor(municipality)Chihuahua     0.088425   0.612670   0.144 0.885295
factor(municipality)Culiacan     -0.235237   0.611851  -0.384 0.700781
factor(municipality)Juarez       -0.009005   0.614229  -0.015 0.988308
factor(municipality)Matamoros     0.760130   0.613805   1.238 0.216103
factor(municipality)Monterrey     0.411768   0.612173   0.673 0.501465
factor(municipality)Nuevo Laredo  2.155425   0.623830   3.455 0.000593 ***
factor(municipality)Reynosa       1.309019   0.617391   2.120 0.034435 *
factor(municipality)Tijuana      -0.100518   0.611966  -0.164 0.869591
factor(municipality)Torreon       0.526026   0.612788   0.858 0.391040
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 3.237 on 547 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.3483,Adjusted R-squared:  0.334
F-statistic: 24.36 on 12 and 547 DF,  p-value: < 2.2e-16
```

Modeling Time-Series Cross-Section data

back to our working example: heterogeneity (random coefficients model)

```
Call:
lm(formula = organized.crime.dead ~ factor(municipality) * organized.crime.dead.L1 +
    factor(municipality) * organized.crime.wounded + factor(municipality) *
    organized.crime.wounded.L1, data = panel)

Residuals:
     Min       1Q   Median       3Q      Max
-14.5040  -0.8529  -0.5059   0.3693  18.4216

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           0.815235   0.436144   1.869  0.06216 .
factor(municipality)Chihuahua                        -0.298263   0.667429  -0.447  0.65514
factor(municipality)Culiacan                         -0.309319   0.634886  -0.487  0.62632
factor(municipality)Juarez                            0.015858   0.642814   0.025  0.98033
factor(municipality)Matamoros                         0.037692   0.632712   0.060  0.95252
factor(municipality)Monterrey                        -0.653681   0.657703  -0.994  0.32074
factor(municipality)Nuevo Laredo                     -0.404602   0.640288  -0.632  0.52773
factor(municipality)Reynosa                          -0.184562   0.639011  -0.289  0.77283
factor(municipality)Tijuana                          -0.490054   0.631085  -0.777  0.43779
factor(municipality)Torreon                           0.620371   0.674076   0.920  0.35783
organized.crime.dead.L1                               0.044951   0.137319   0.327  0.74354
organized.crime.wounded                               0.246583   0.242266   1.018  0.30924
organized.crime.wounded.L1                            0.027641   0.235494   0.117  0.90661
factor(municipality)Chihuahua:organized.crime.dead.L1     0.052567   0.313627   0.168  0.86695
factor(municipality)Culiacan:organized.crime.dead.L1     -0.089649   0.343555  -0.261  0.79424
factor(municipality)Juarez:organized.crime.dead.L1       -0.415334   0.324187  -1.281  0.20071
factor(municipality)Matamoros:organized.crime.dead.L1     0.407657   0.173959   2.343  0.01948 *
factor(municipality)Monterrey:organized.crime.dead.L1     0.001697   0.194372   0.009  0.99304
factor(municipality)Nuevo Laredo:organized.crime.dead.L1  0.371823   0.152801   2.433  0.01529 *
factor(municipality)Reynosa:organized.crime.dead.L1       0.535074   0.165625   3.231  0.00131 **
factor(municipality)Tijuana:organized.crime.dead.L1      -0.227522   0.271611  -0.838  0.40260
factor(municipality)Torreon:organized.crime.dead.L1       0.081801   0.185365   0.441  0.65918
factor(municipality)Chihuahua:organized.crime.wounded     0.484621   0.546482   0.887  0.37560
factor(municipality)Culiacan:organized.crime.wounded      0.031603   0.336098   0.094  0.92512
factor(municipality)Juarez:organized.crime.wounded        0.526326   0.275800   1.908  0.05689 .
factor(municipality)Matamoros:organized.crime.wounded     0.778769   0.319314   2.439  0.01507 *
factor(municipality)Monterrey:organized.crime.wounded     0.713751   0.428343   1.666  0.09625 .
factor(municipality)Nuevo Laredo:organized.crime.wounded  3.007188   0.433448   6.938 1.19e-11 ***
factor(municipality)Reynosa:organized.crime.wounded       2.663122   0.475259   5.604 3.41e-08 ***
factor(municipality)Tijuana:organized.crime.wounded       0.645871   0.303601   2.127  0.03386 *
factor(municipality)Torreon:organized.crime.wounded       0.188931   0.320248   0.590  0.55548
factor(municipality)Chihuahua:organized.crime.wounded.L1 -0.130658   0.574804  -0.227  0.82027
factor(municipality)Culiacan:organized.crime.wounded.L1   0.004471   0.336762   0.013  0.98941
factor(municipality)Juarez:organized.crime.wounded.L1     0.249126   0.361941   0.688  0.49157
factor(municipality)Matamoros:organized.crime.wounded.L1 -0.485304   0.336701  -1.441  0.15009
factor(municipality)Monterrey:organized.crime.wounded.L1  1.134507   0.456030   2.488  0.01317 *
factor(municipality)Nuevo Laredo:organized.crime.wounded.L1  1.211474   0.493106   2.457  0.01434 *
factor(municipality)Reynosa:organized.crime.wounded.L1   -1.682596   0.537712  -3.129  0.00185 **
factor(municipality)Tijuana:organized.crime.wounded.L1    0.180451   0.361944   0.499  0.61830
factor(municipality)Torreon:organized.crime.wounded.L1   -0.036993   0.319468  -0.116  0.90786
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 2.902 on 520 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.5019,Adjusted R-squared:  0.4645
F-statistic: 13.43 on 39 and 520 DF,  p-value: < 2.2e-16
```

Team Progress Review

Analyzing collective individual behavior

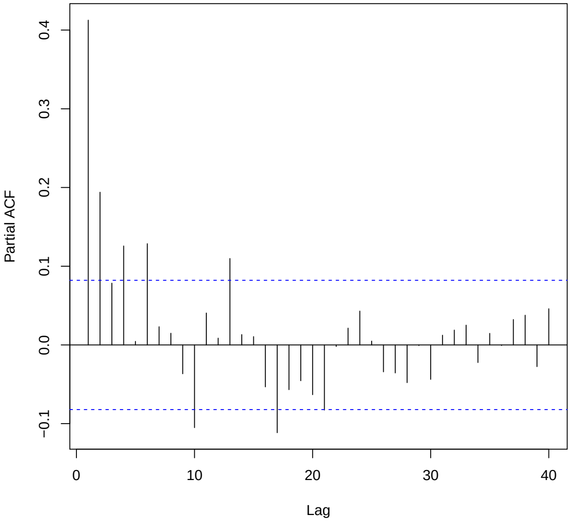
Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2017
Columbia University

Modeling Time-Series Cross-Section data

back to our working example: the "right" lag specification?



Modeling Time-Series Cross-Section data

back to our working example: the "right" lag specification?

```
Call:
lm(formula = organized.crime.dead ~ organized.crime.dead.L1 +
    organized.crime.dead.L2 + organized.crime.dead.L3 + organized.crime.wounded +
    organized.crime.wounded.L1, data = panel)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0489  -1.1404  -0.3522  -0.1404  20.7192

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.35224    0.17261   2.041   0.0418 *
organized.crime.dead.L1     0.30905    0.04346   7.112 3.70e-12 ***
organized.crime.dead.L2     0.21336    0.04341   4.914 1.19e-06 ***
organized.crime.dead.L3     0.09394    0.04335   2.167   0.0307 *
organized.crime.wounded     0.78812    0.07689  10.250  < 2e-16 ***
organized.crime.wounded.L1 -0.18941    0.08392  -2.257   0.0244 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 3.234 on 534 degrees of freedom
  (30 observations deleted due to missingness)
Multiple R-squared:  0.3601,Adjusted R-squared:  0.3541
F-statistic:  60.1 on 5 and 534 DF,  p-value: < 2.2e-16
```