Lecture 8: Serial Correlation

Prof. Sharyn O'Halloran

Sustainable Development U9611

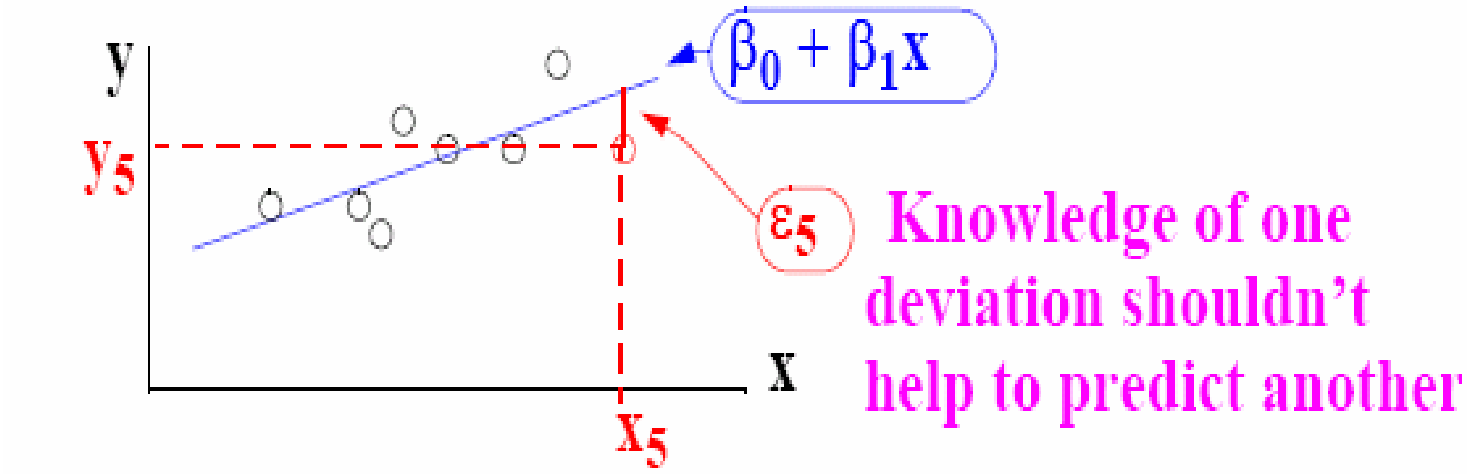
Econometrics II

Midterm Review

- Most people did very well
  - Good use of graphics
  - Good writeups of results
  - A few technical issues gave people trouble
    - F-tests
    - Predictions from linear regression
    - Transforming variables
- A do-file will be available on Courseworks to check your answers

Review of independence assumption

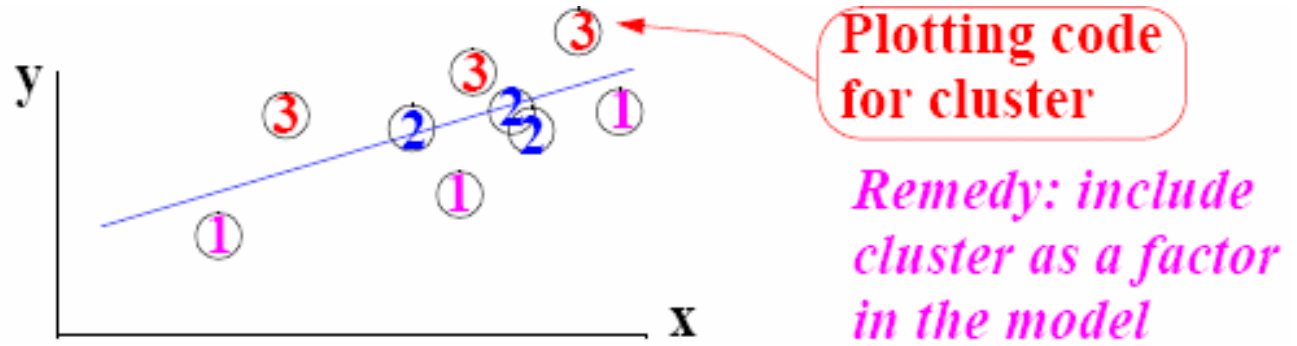
- Model:
  - $y_i = b_0 + b_1 x_i + e_i$ (i = 1, 2, ..., n) e_i is independent of e_j for all distinct indices i, j



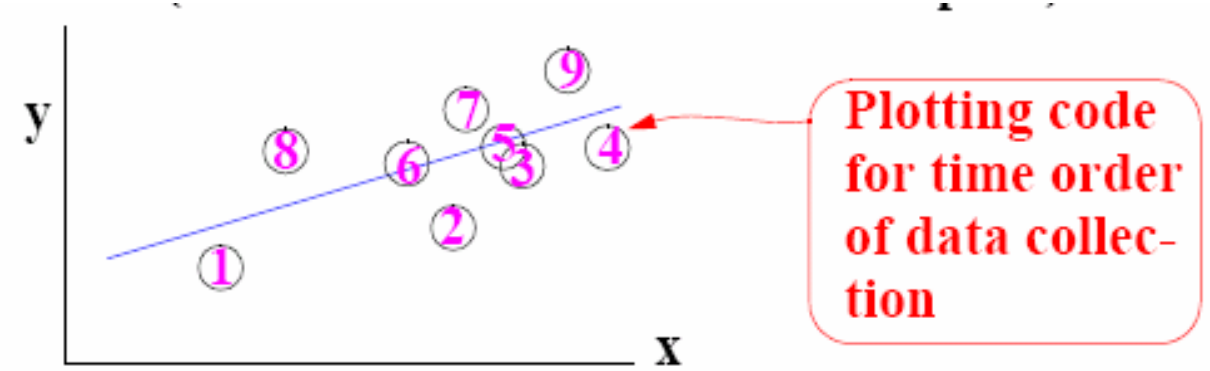
- Consequences of non-independence:
  - SE's, tests, and CIs will be incorrect;
  - LS isn't the best way to estimate β's

Main Violations

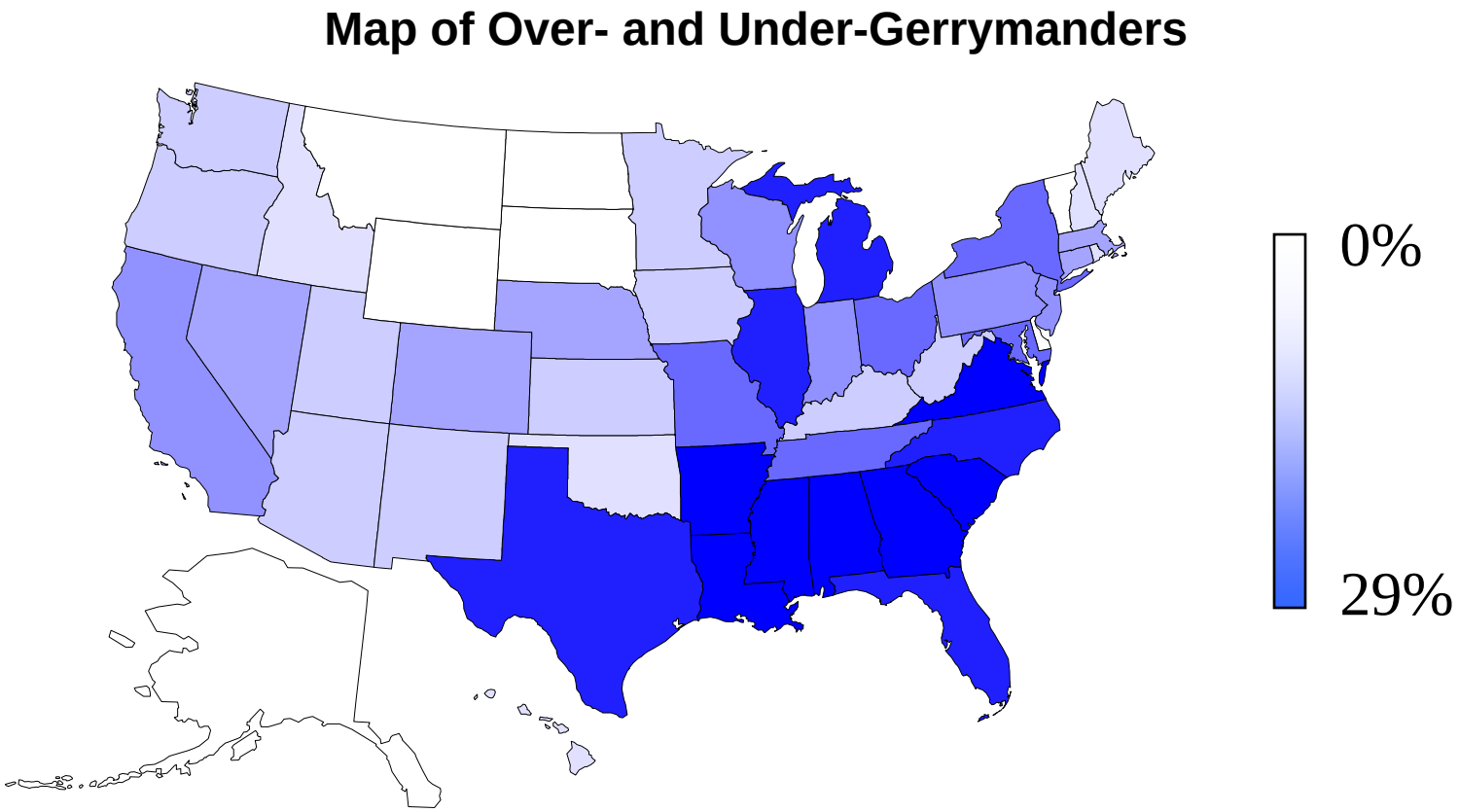
- Cluster effects (ex: mice litter mates)



- Serial effects (for data collected over time or space)



Spatial Autocorrelation

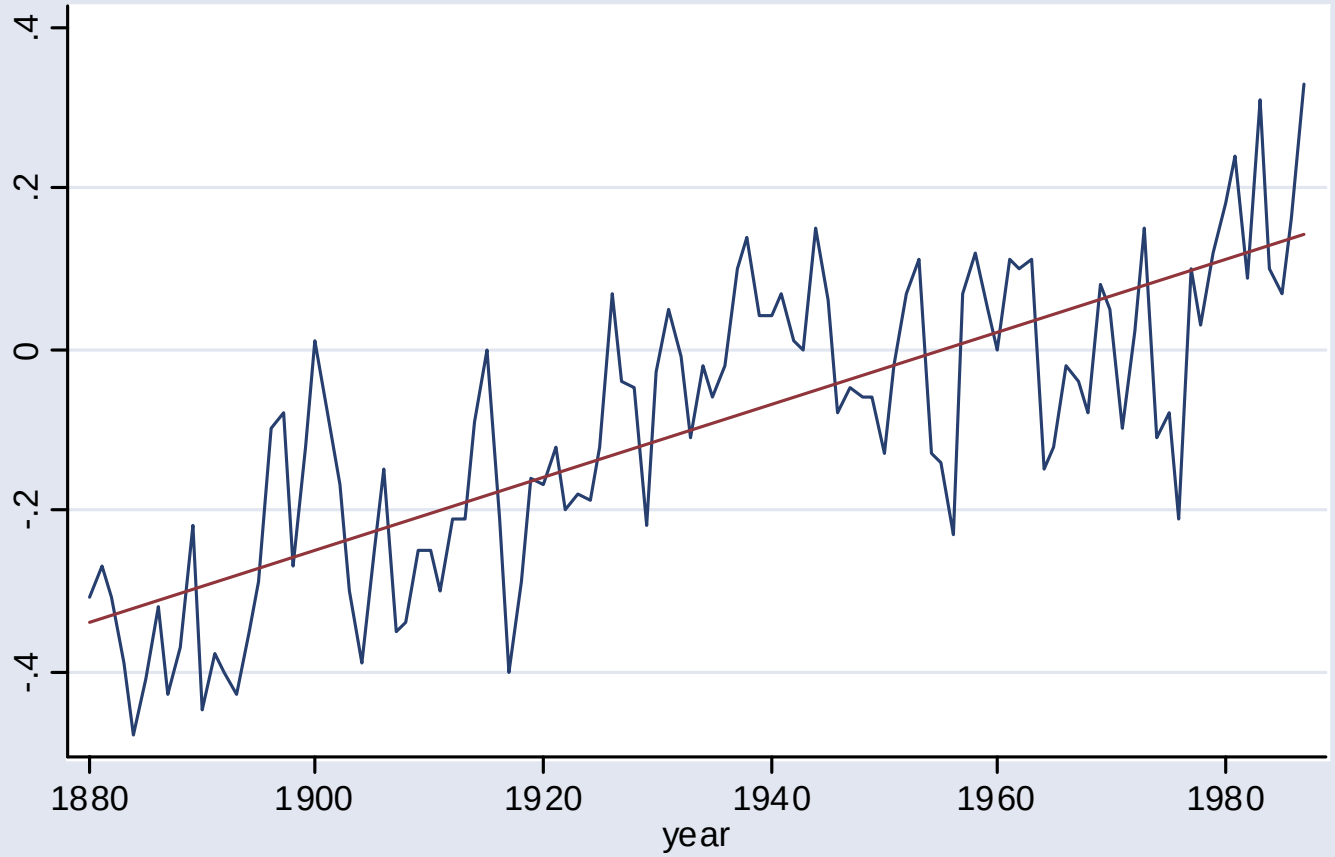


- Clearly, the value for a given state is correlated with neighbors
- This is a hot topic in econometrics these days…

Time Series Analysis

- More usual is correlation over time, or serial correlation: this is *time series* analysis
  - So residuals in one period (ε_t) are correlated with residuals in previous periods (ε_{t-1}, ε_{t-2}, etc.)
  - Examples: tariff rates; debt; partisan control of Congress, votes for incumbent president, etc.
- Stata basics for time series analysis
  - First use **`tsset var`** to tell Stata data are time series, with **`var`** as the time variable
  - Can use **`L.anyvar`** to indicate lags
    - Same with **`L2.anyvar`**, **`L3.anyvar`**, etc.
  - And can use **`F.anyvar`**, **`F2.anyvar`**, etc. for leads

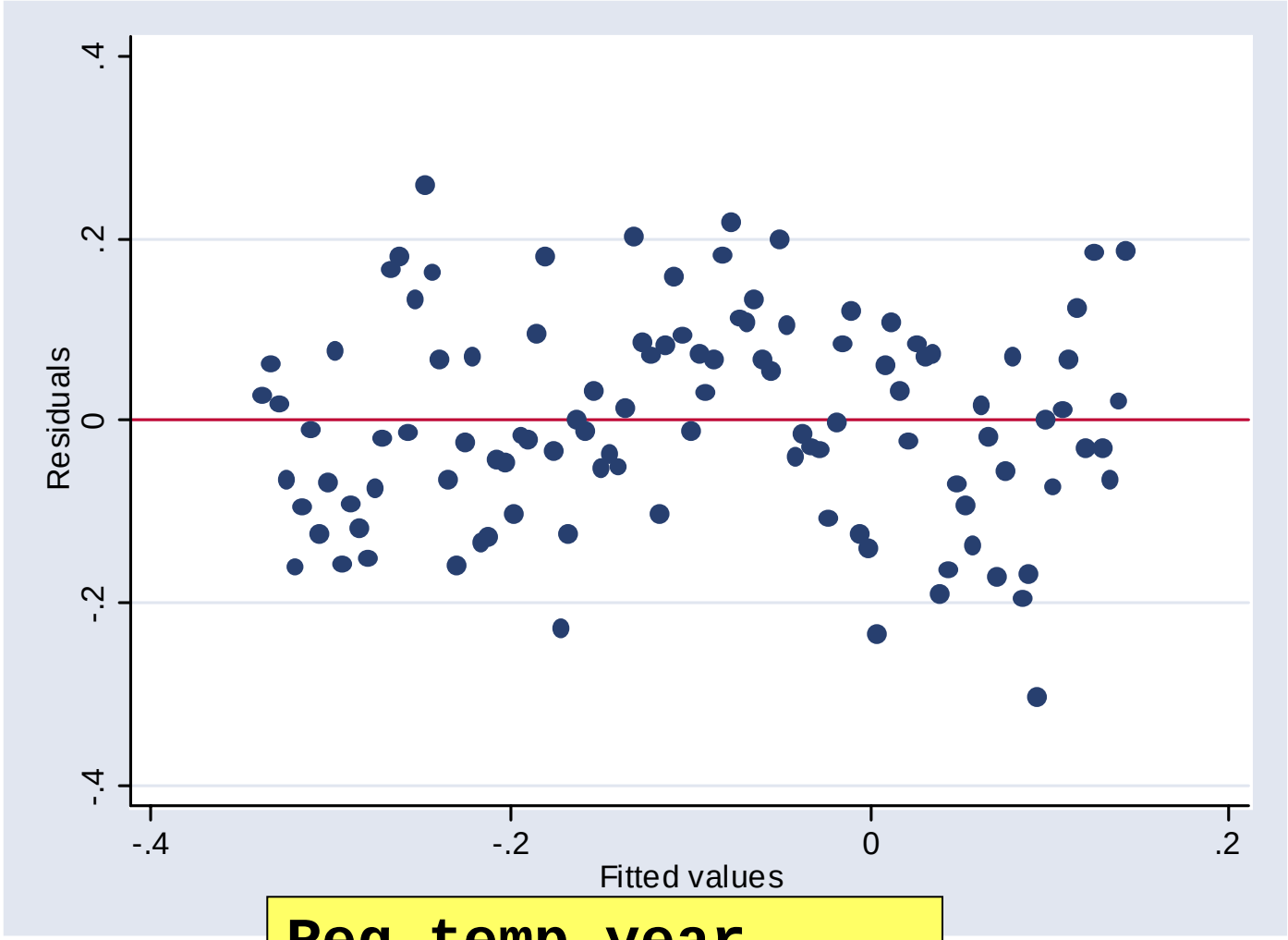
Diagnosing the Problem



Temperature data with linear fit line drawn in

```
tsset year
twoway (tsline temp) lfit temp year
```

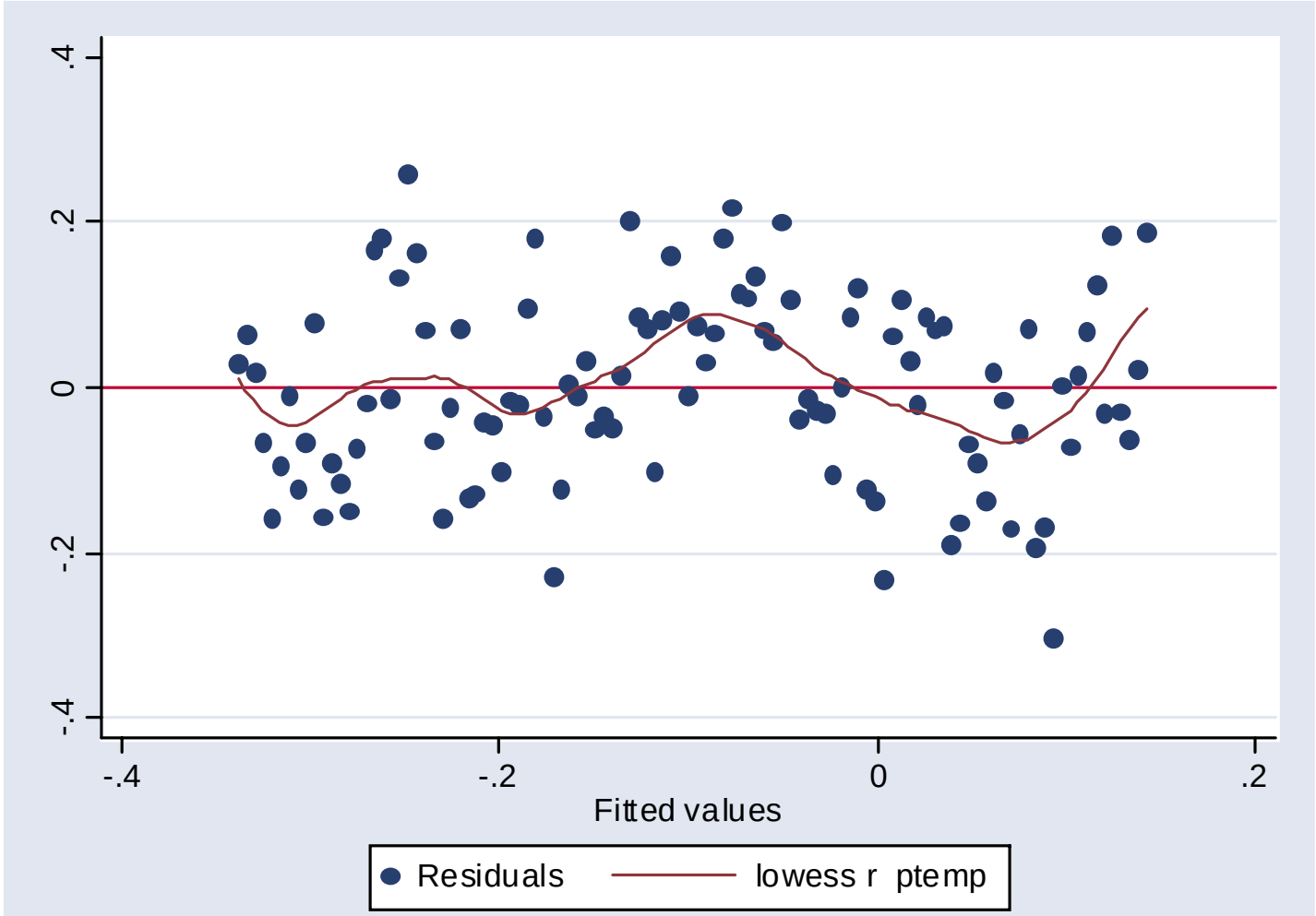
Diagnosing the Problem



Rvfplot doesn't look too bad…

```
Reg temp year
rvfplot, yline(0)
```

Diagnosing the Problem



But adding a lowess line shows that the residuals cycle.

In fact, the amplitude may be <u>increasing</u> over time.

```
predict ptemp; predict r, resid
scatter r ptemp || lowess r ptemp, bw(.3) yline(0)
```

Diagnosing the Problem

- One way to think about the problem is the pattern of residuals: (+,+,+,-,-,+,+,+…)
  - With no serial correlation, the probability of a "+" in this series is independent of history
  - With (positive) serial correlation, the probability of a "+" following a "+" is greater than following a "-"
- In fact, there is a nonparametric test for this:

$$\mu = \frac{2mp}{m+p} + 1$$

$$\sigma = \sqrt{\frac{2mp(2mp - m - p)}{(m+p)^2(m+p-1)}}$$

$$Z = \frac{(\text{number of runs}) - \mu + C}{\sigma}$$

m = # of minuses, p = # of pluses

C = +.5 (-.5) if # of runs < (>) μ

Z distributed standard normal

Calculations

```
. gen plusminus = "+"
. replace plusminus = "-" if r<0
```

	year	plusminus
1.	1880	+
2.	1881	+
3.	1882	+
4.	1883	-
5.	1884	-
6.	1885	-
7.	1886	-
8.	1887	-
9.	1888	-
10.	1889	+
11.	1890	-
12.	1891	-
13.	1892	-
14.	1893	-
15.	1894	-

	year	plusminus
16.	1895	-
17.	1896	+
18.	1897	+
19.	1898	-
20.	1899	+
21.	1900	+
22.	1901	+
23.	1902	+
24.	1903	-
25.	1904	-
26.	1905	-
27.	1906	+
28.	1907	-
29.	1908	-
30.	1909	-

Calculations

```
gen newrun = plusm[_n]~=plusm[_n-1]
gen runs = sum(newrun)
```

	year	plusminus	newrun	runs
1.	1880	+	1	1
2.	1881	+	0	1
3.	1882	+	0	1
4.	1883	-	1	2
5.	1884	-	0	2
6.	1885	-	0	2
7.	1886	-	0	2
8.	1887	-	0	2
9.	1888	-	0	2
10.	1889	+	1	3
11.	1890	-	1	4
12.	1891	-	0	4
13.	1892	-	0	4
14.	1893	-	0	4
15.	1894	-	0	4

Calculations

```
. sum runs
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        runs |       108    18.23148    10.68373          1         39

 . dis (2*39*69)/108 + 1   *This is mu
50.833333

. dis sqrt((2*39*69)*(2*39*69-39-69)/(108^2*107))   *This is sigma
4.7689884

. dis (39-50.83+0.5)/4.73    *This is Z score
-2.3953488
```

$$\mu = \frac{2mp}{m+p} + 1$$

$$\sigma = \sqrt{\frac{2mp(2mp-m-p)}{(m+p)^2(m+p-1)}}$$

$$Z = \frac{(\text{number of runs}) - \mu + C}{\sigma}$$

m = # of minuses, p = # of pluses

C = +.5 (-.5) if # of runs < (>) μ

Z distributed standard normal

Calculations

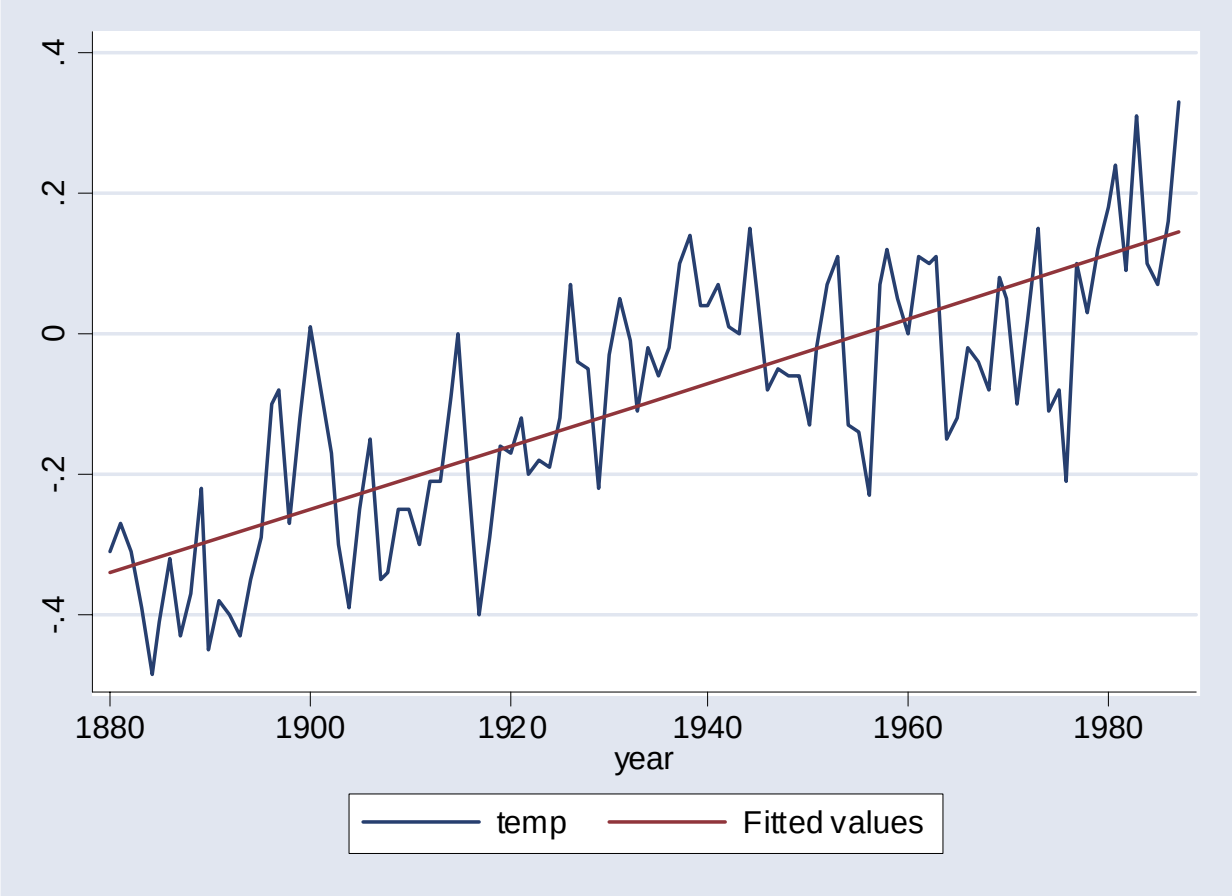
```
. sum runs

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        runs |       108    18.23148    10.68373          1         39


. dis (2*39*69)/108 + 1   *This is mu
50.833333

. dis sqrt((2*39*69)*(2*39*69-39-69)/(108^2*107))   *This is sigma
4.7689884

. dis (39-50.83+0.5)/4.73    *This is Z score
-2.3953488
```

The Z-score is significant, so we can reject the null that the number of runs was generated randomly.

Autocorrelation and partial autocorrelation coefficients

- *(a) Estimated autocorrelation coefficients of lag k* are (essentially)
  - The correlation coefficients between the residuals and the lag *k* residuals
- *(b) Estimated partial autocorrelation coefficients of lag k* are (essentially)
  - The correlation coefficients between the residuals and the lag *k* residuals, after accounting for the lag 1,...,lag (*k*-1) residuals
  - I.e., from multiple regression of residuals on the lag 1, lag 2,...,lag *k* residuals
- Important: in checking to see what order of autoregressive (AR) model is necessary, it is (b), not (a) that must be used.

Example: Temperature Data



Save residuals from ordinary regression fit

Test lag structure of residuals for autocorrelation

```
tsset year
twoway (tsline temp) lfit temp year
```

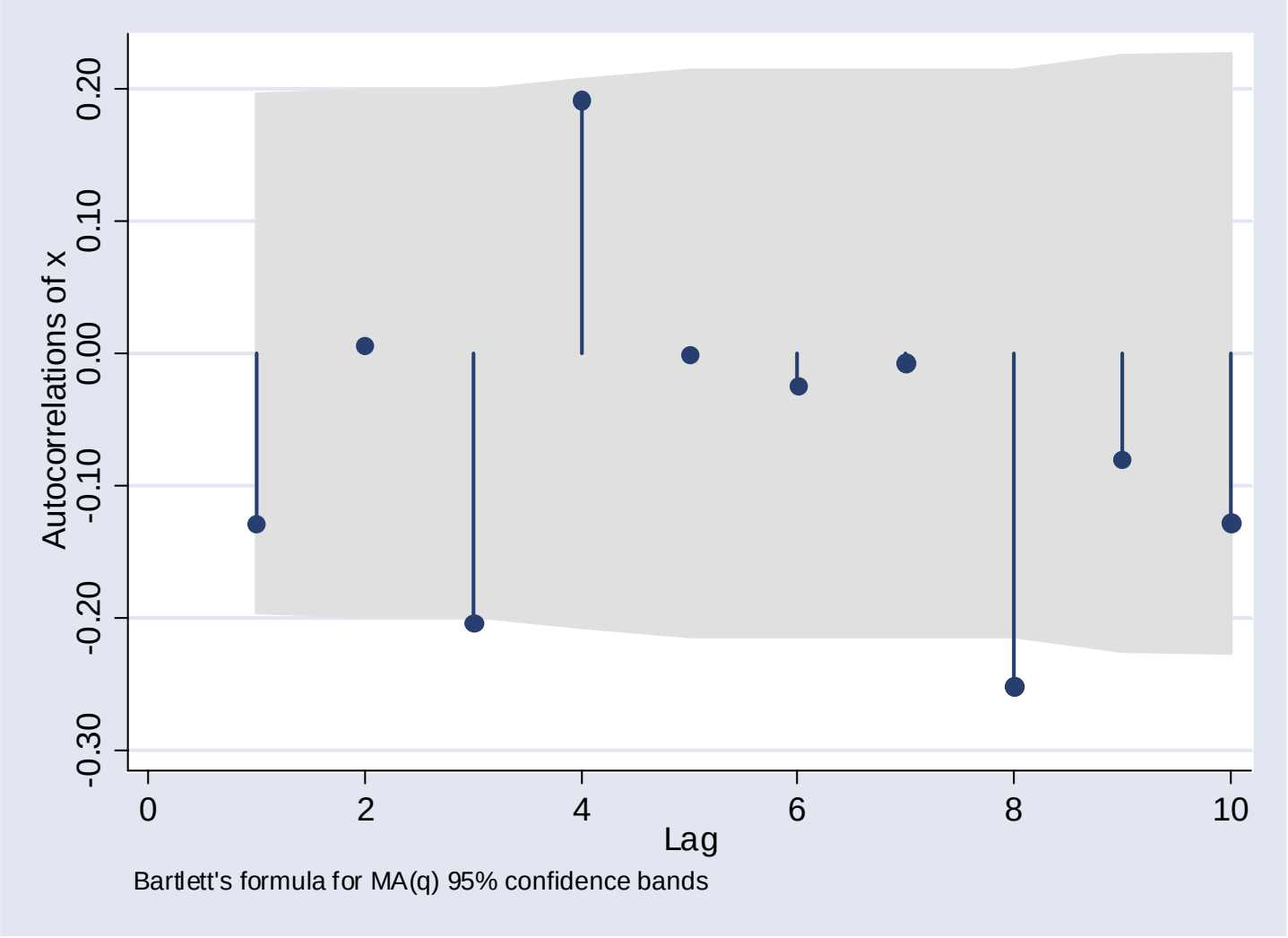
Examining Autocorrelation

- One useful tool for examining the degree of autocorrelation is a correlogram
  - This examines the correlations between residuals at times t and t-1, t-2, …
- If no autocorrelation exists, then these should be 0, or at least have no pattern
  - **`corrgram var, lags(t)`** creates a text correlogram of variable **`var`** for t periods
  - **`ac var, lags(t)`**: autocorrelation graph
  - **`pac var`**: partial autocorrelation graph

Example: Random Data

```
. gen x = invnorm(uniform())
. gen t = _n
. tsset t
        time variable:  t, 1 to 100
. corrgram x, lags(20)
                                                -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1       -0.1283  -0.1344    1.696  0.1928          -|                  -|
2        0.0062  -0.0149   1.6999  0.4274           |                   |
3       -0.2037  -0.2221   6.0617  0.1086          -|                  -|
4        0.1918   0.1530   9.9683  0.0410           |-                  |-
5       -0.0011   0.0457   9.9684  0.0761           |                   |
6       -0.0241  -0.0654   10.032  0.1233           |                   |
7       -0.0075   0.0611   10.038  0.1864           |                   |
8       -0.2520  -0.3541   17.078  0.0293         --|                --|
9       -0.0811  -0.2097   17.816  0.0374           |                  -|
10      -0.1278  -0.2059   19.668  0.0326          -|                  -|
11       0.1561  -0.0530   22.462  0.0210           |-                  |
12      -0.1149  -0.0402   23.992  0.0204           |                   |
13       0.1168   0.1419   25.591  0.0193           |                   |-
14      -0.1012  -0.0374   26.806  0.0204           |                   |
15       0.0400  -0.0971   26.998  0.0288           |                   |
16       0.0611   0.0639   27.451  0.0367           |                   |
17       0.0947  -0.1022   28.552  0.0389           |                   |
18      -0.0296  -0.1728   28.661  0.0527           |                  -|
19      -0.0997  -0.0916   29.914  0.0529           |                   |
20       0.0311  -0.0789   30.037  0.0693           |                   |
```

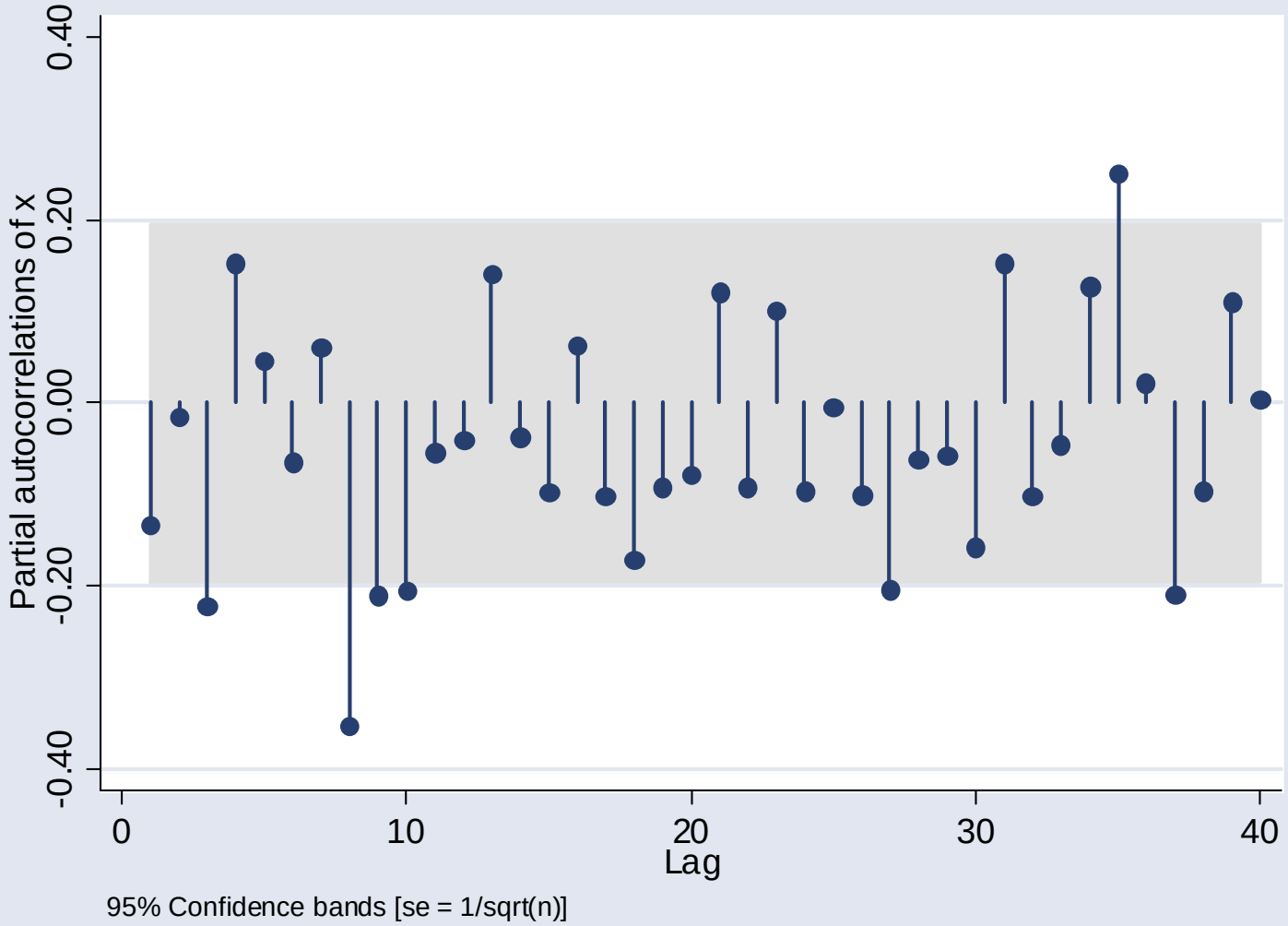
Example: Random Data



No pattern is apparent in the lag structure.

```
ac x, lags(10)
```

Example: Random Data



Still no pattern…

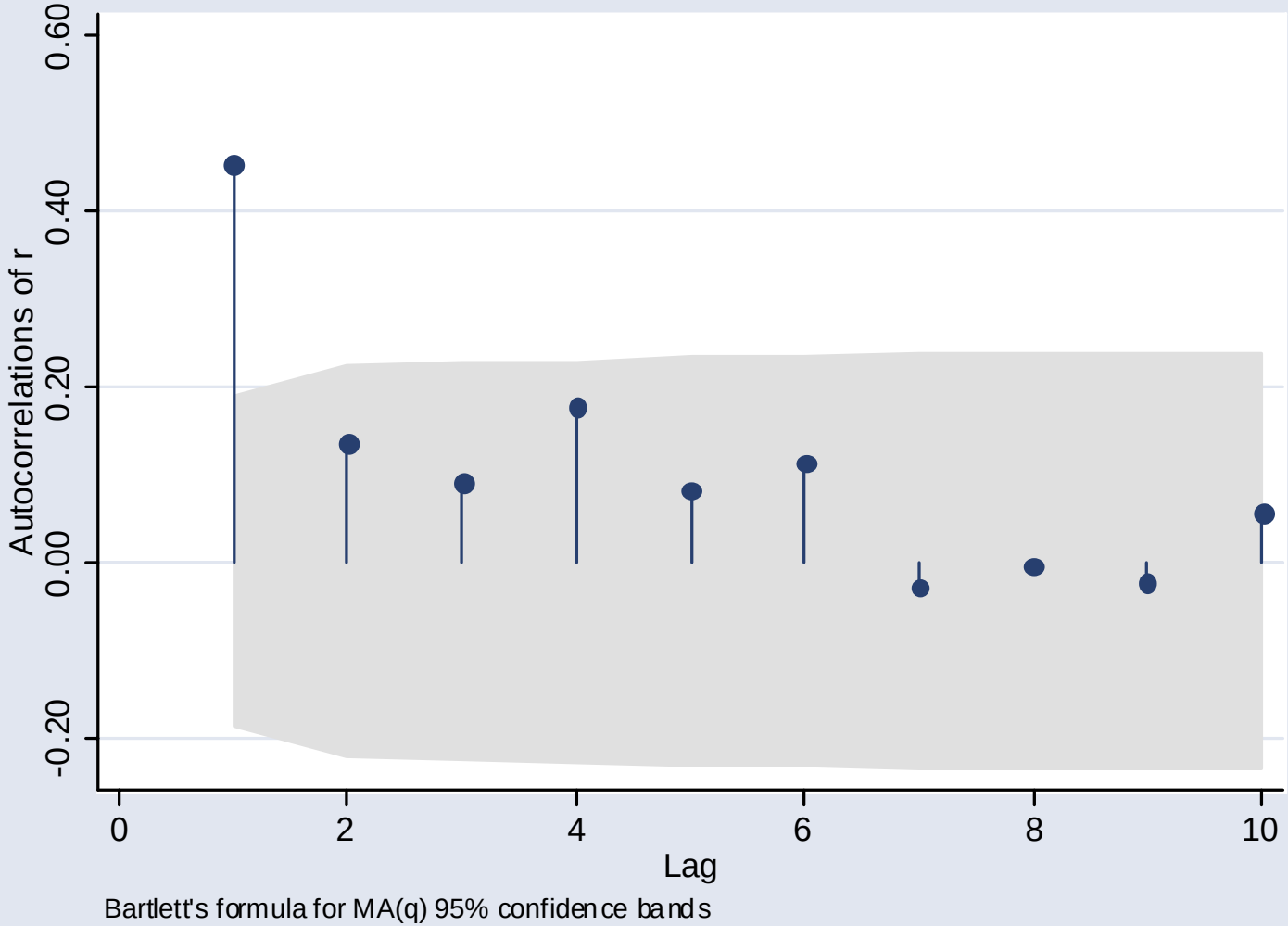
```
pac x
```

Example: Temperature Data

```
. tsset year
        time variable:  year, 1880 to 1987

. corrgram r, lags(20)
```

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.4525	0.4645	22.732	0.0000	|---	|---
2	0.1334	-0.0976	24.727	0.0000	|-	|
3	0.0911	0.0830	25.667	0.0000	|	|
4	0.1759	0.1451	29.203	0.0000	|-	|-
5	0.0815	-0.0703	29.969	0.0000	|	|
6	0.1122	0.1292	31.435	0.0000	|	|-
7	-0.0288	-0.1874	31.533	0.0000	|	-|
8	-0.0057	0.0958	31.537	0.0001	|	|
9	-0.0247	-0.0802	31.61	0.0002	|	|
10	0.0564	0.1007	31.996	0.0004	|	|
11	-0.0253	-0.0973	32.075	0.0007	|	|
12	-0.0678	-0.0662	32.643	0.0011	|	|
13	-0.0635	0.0358	33.147	0.0016	|	|
14	0.0243	0.0037	33.222	0.0027	|	|
15	-0.0583	-0.1159	33.656	0.0038	|	|
16	-0.0759	0.0009	34.399	0.0048	|	|
17	-0.0561	-0.0180	34.81	0.0066	|	|
18	-0.0114	0.0252	34.827	0.0099	|	|
19	-0.0202	-0.0007	34.882	0.0144	|	|
20	0.0437	0.0910	35.139	0.0194	|	|

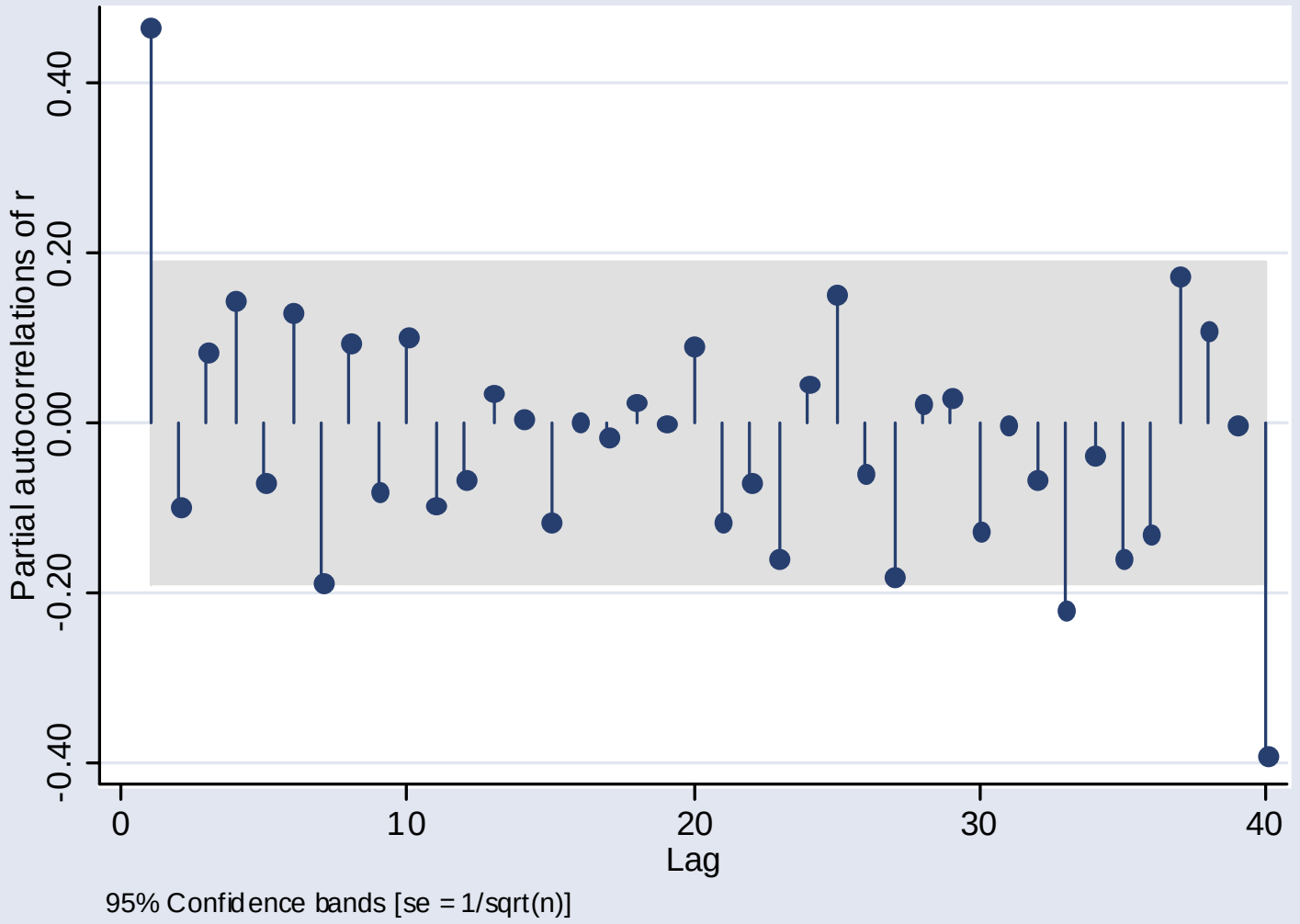
Example: Temperature Data



Now clear pattern emerges

```
ac r, lags(10)
```

Example: Temperature Data



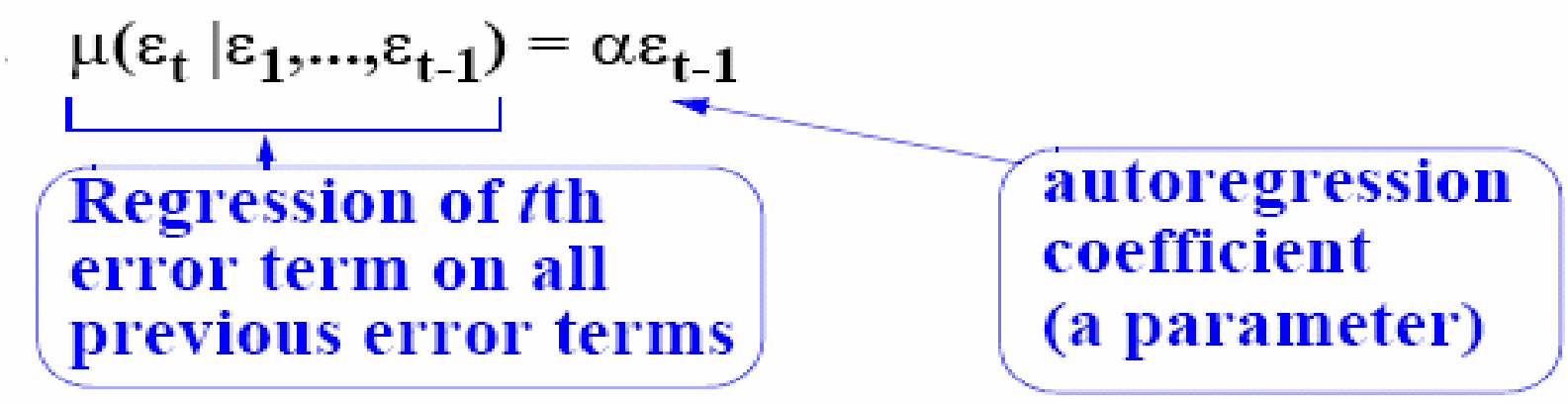
Clear correlation with first lag

So we have an AR(1) process

```
pac r
```

Autoregressive model of lag 1: AR(1)

- Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are random error terms from measurements at *equally-spaced* time points, and $\mu(\varepsilon_t) = 0$



- Notice that the ε's are not independent, but...
  - The dependence is only through the *previous* error term (time t-1), not any other error terms

Estimating the first serial correlation coefficient from residuals of a single series

- Let e_1, e_2, ..., e_n be the *residuals* from the series (Note $\bar{e} = 0$).
- $$\text{Let } c_1 = \sum_{t=2}^{n} e_t e_{t-1} \text{ and } c_0 = \sum_{t=2}^{n} e_t^2$$
- The estimate of the first serial correlation coefficient (α) is $r_1 = c_1/c_0$
- Note: this is (almost) the *sample correlation* of residuals e_2, e_3, ...,e_n with the "*lag 1*" residuals e_1, e_2, ..., e_{n-1}

Example: Global Warming Data

```
. tsset year
        time variable:  year, 1880 to 1987

. reg temp year

      Source |       SS       df       MS              Number of obs =     108
-------------+------------------------------           F(  1,   106) =  163.50
       Model |  2.11954255     1  2.11954255           Prob > F      =  0.0000
    Residual |  1.37415745   106   .01296375           R-squared     =  0.6067
-------------+------------------------------           Adj R-squared =  0.6030
       Total |      3.4937   107  .032651402           Root MSE      =  .11386

------------------------------------------------------------------------------
        temp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0044936   .0003514    12.79   0.000     .0037969    .0051903
       _cons |  -8.786714   .6795784   -12.93   0.000    -10.13404   -7.439384
------------------------------------------------------------------------------

. predict r, resid

. corr r L.r
(obs=107)

             |       r       L.r
-------------+------------------
           r |   1.0000
         L.r |   0.4586   1.0000
```

Example: Global Warming Data

```
. tsset year
        time variable:  year, 1880 to 1987

. reg temp year

      Source |       SS       df       MS              Number of obs =     108
-------------+------------------------------           F(  1,   106) =  163.50
       Model |  2.11954255     1  2.11954255           Prob > F      =  0.0000
    Residual |  1.37415745   106   .01296375           R-squared     =  0.6067
-------------+------------------------------           Adj R-squared =  0.6030
       Total |      3.4937   107  .032651402           Root MSE      =  .11386

------------------------------------------------------------------------------
        temp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0044936   .0003514    12.79   0.000     .0037969    .0051903
       _cons |  -8.786714   .6795784   -12.93   0.000    -10.13404   -7.439384
------------------------------------------------------------------------------

. predict r, resid

. corr r L.r
(obs=107)

             |        r      L.r
-------------+------------------
           r |   1.0000
         L.r |   0.4586   1.0000
```

Estimate of r_1

Regression in the AR(1) Model

- Introduction: Steps in global warming analysis
  1. Fit the usual regression of TEMP (Y_t) on YEAR (X_t).
  2. Estimate the 1st ser. corr. coeff, r_1, from the *residuals*
  3. Is there serial correlation present? (Sect. 15.4)
  4. Is the serial correlation of the AR(1) type? (Sect. 15.5)
  5. If yes, use the filtering transformation (Sect. 15.3.2):
     - $V_t = Y_t - r_1 Y_{t-1}$
     - $U_t = X_t - r_1 X_{t-1}$
  6. Regress V_t on U_t to get AR(1)-adjusted regression estimates

Filtering, and why it works:

- Simple regression model with AR(1) error structure:
  - $Y_t=\beta_0+\beta_1 X_t+\varepsilon_t$; $\mu(\varepsilon_t|\varepsilon_1,\ldots\varepsilon_{t-1}) = \alpha\varepsilon_{t-1}$
- Ideal "filtering transformations:"
  - $V_t=Y_t-\alpha Y_{t-1}$ and $U_t=X_t-\alpha X_{t-1}$
- Algebra showing the induced regression of V_t on U_t

$$V_t = Y_t - \alpha Y_{t-1} = (\beta_0 + \beta_1 X_t + \varepsilon_t) - \alpha(\beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1})$$
$$= (\beta_0 - \alpha\beta_0) + \beta_1(X_t - \alpha X_{t-1}) + (\varepsilon_t - \alpha\varepsilon_{t-1})$$
$$= \gamma_0 + \beta_1 U_t + \varepsilon_t^*$$

Reg of V on U has same slope as Y on X

There is no serial corr. present in the ε^*'s

Filtering, and why it works: (cont.)

- The AR(1) serial correlation has been filtered out; that is
  - $\mu(\varepsilon_t^* \mid \varepsilon_1^*, \ldots, \varepsilon_{t-1}^*) = \mu(\varepsilon_t - \alpha\varepsilon_{t-1} \mid \varepsilon_1, \ldots, \varepsilon_{t-1}) = 0$
  - Since ε_t-$\alpha\varepsilon_{t-1}$ is independent of all previous residuals.
- So, least squares inferences about β_1 in the regression of V on U is correct.
- Since α is unknown, use its estimate, r_1, instead.

Example: Global Warming Data

- Estimate of the warming trend
- Fit simple regression of TEMP on YEAR
  - Estimate Slope: .004499 (SE=.00035)
- Get the estimate of 1st serial correlation coefficient from the residuals: $r_1 = .452$
- Create new columns of data:
  - $V_t = TEMP_t - r_1 TEMP_{t-1}$ and $U_t = YEAR_t - r_1 YEAR_{t-1}$
- Fit the simple reg. of U on V
  - Estimated slope: .00460 (SE = .00058)
- Use above in the summary of statistical findings

```
. gen yearF = year - 0.4525*L.year
. gen tempF = temp - 0.4525*L.temp
. reg tempF yearF

      Source |       SS       df       MS              Number of obs =     107
-------------+------------------------------           F(  1,   105) =   62.79
       Model |  .648460599     1  .648460599           Prob > F      =  0.0000
    Residual |  1.08435557   105  .010327196           R-squared     =  0.3742
-------------+------------------------------           Adj R-squared =  0.3683
       Total |  1.73281617   106  .016347322           Root MSE      =  .10162

------------------------------------------------------------------------------
       tempF |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       yearF |   .0046035    .000581     7.92   0.000     .0034516    .0057555
       _cons |  -4.926655   .6154922    -8.00   0.000    -6.147063   -3.706248
------------------------------------------------------------------------------

. reg temp year

      Source |       SS       df       MS              Number of obs =     108
-------------+------------------------------           F(  1,   106) =  163.50
       Model |  2.11954255     1  2.11954255           Prob > F      =  0.0000
    Residual |  1.37415745   106   .01296375           R-squared     =  0.6067
-------------+------------------------------           Adj R-squared =  0.6030
       Total |      3.4937   107  .032651402           Root MSE      =  .11386

------------------------------------------------------------------------------
        temp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .0044936   .0003514    12.79   0.000     .0037969    .0051903
       _cons |  -8.786714   .6795784   -12.93   0.000    -10.13404   -7.439384
------------------------------------------------------------------------------
```

Filtering in multiple regression

$$V_t = Y_t - \alpha Y_{t-1}$$

$$U_{1t} = X_{1t} - \alpha X_{1(t-1)}$$

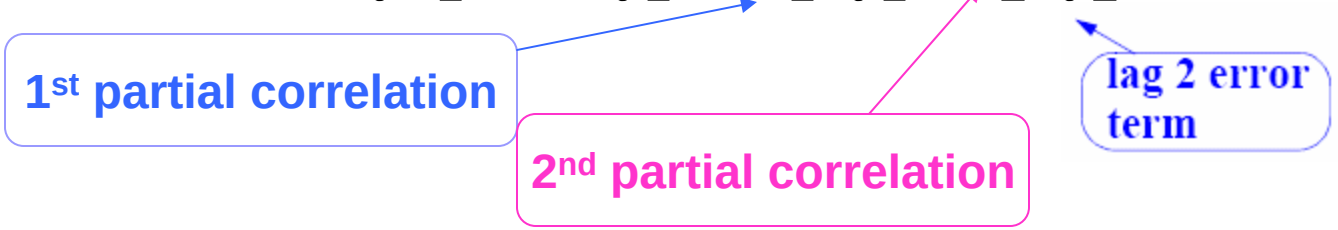
$$U_{2t} = X_{2t} - \alpha X_{2(t-1)}$$

$$etc...$$

- Fit the least squares regression of V on U_1, U_2, etc… using r_1 as an estimate of α.

AR(2) Models, etc…

- AR(2): $\mu(\varepsilon_t|\varepsilon_1,\ldots,\varepsilon_{t-1})=\alpha_1\varepsilon_{t-1}+\alpha_2\varepsilon_{t-2}$



- In this model the deviations from the regression depend on the previous *two* deviations
- Estimate of $\alpha_1=r_1$ as before (estimate of 1st serial correlation coefficient is al so the first partial autocorrelation
- Estimate of $\alpha_2 = r_2$ comes from the multiple regression of residuals on the lag 1 and lag 2 residuals

Example: Democracy Scores

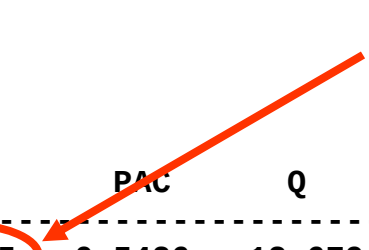
- Data for Turkey, 1955-2000
  - Dep. Var. = Polity Score (-10 to 10)
  - Independent Variables
    - GDP per capita
    - Trade openness
- Theory: Higher GDP per capita and trade openness lead to more democracy
- But democracy level one year may affect the level in following years.

Example: Democracy Scores

```
. reg  polxnew  gdp  trade

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  2,    37) =    1.01
       Model |  34.2273459     2  17.1136729            Prob > F      =  0.3729
    Residual |  624.872654    37  16.8884501            R-squared     =  0.0519
-------------+------------------------------           Adj R-squared =  0.0007
       Total |       659.1    39        16.9           Root MSE      =  4.1096

------------------------------------------------------------------------------
     polxnew |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |  -6.171461   5.103651    -1.21   0.234    -16.51244    4.169518
       trade |   3.134679   2.206518     1.42   0.164    -1.336152     7.60551
       _cons |   49.16307   37.21536     1.32   0.195    -26.24241    124.5686
------------------------------------------------------------------------------

. predict r, resid

. corr r L.r
(obs=39)

             |
             |        r      L.r
-------------+------------------
r            |
          -- |   1.0000
          L1 |   0.5457   1.0000
```

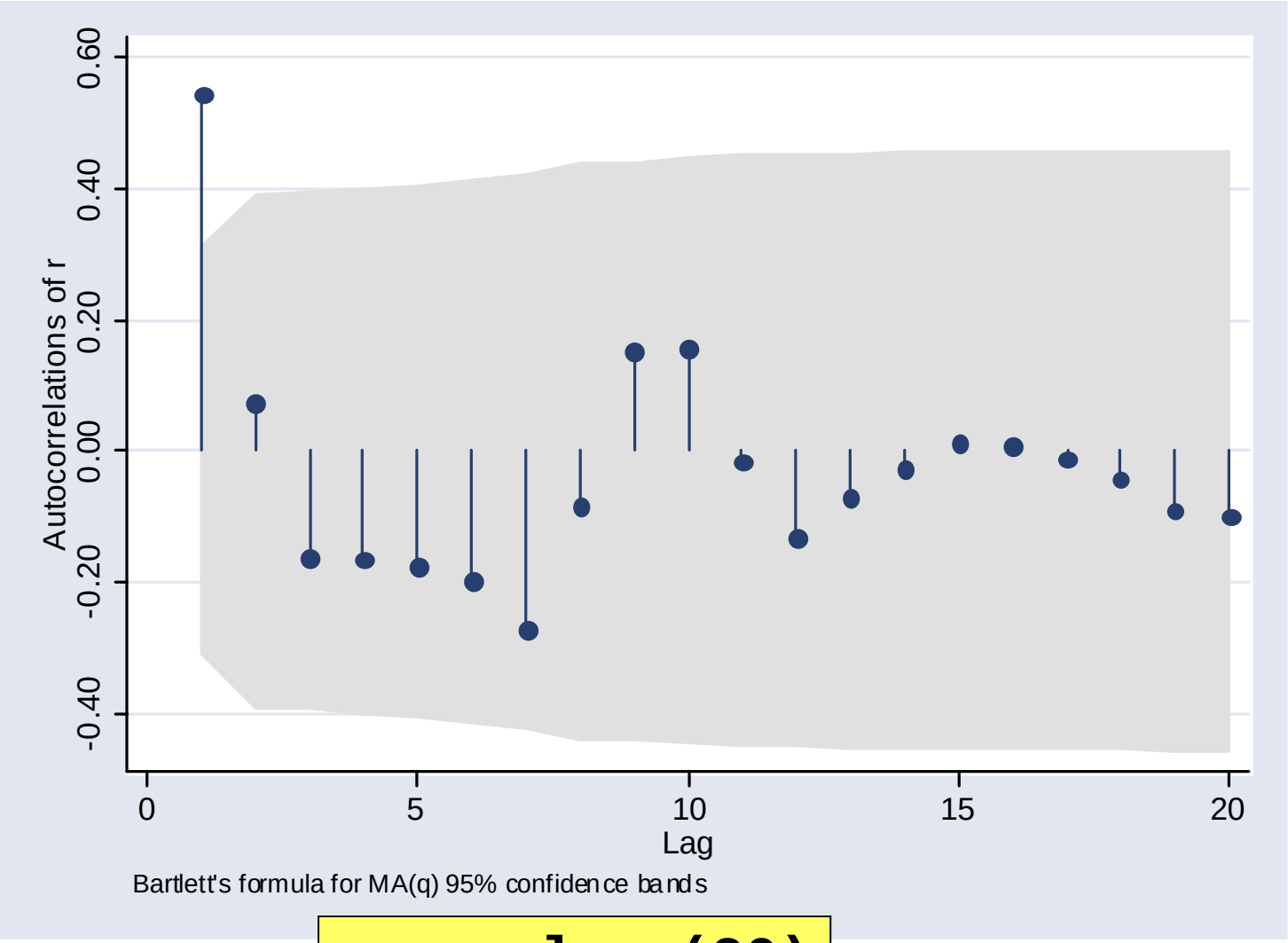
High correlation suggests that residuals are not independent

Example: Democracy Scores




```
. corrgram r

                                                      -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.5425   0.5426   12.679  0.0004           |----                |----
2        0.0719  -0.3189   12.908  0.0016           |                 --|
3       -0.1621  -0.0597   14.102  0.0028         -|                   |
4       -0.1643   0.0100   15.362  0.0040         -|                   |
5       -0.1770  -0.1686   16.865  0.0048         -|                  -|
6       -0.1978  -0.0910   18.799  0.0045         -|                   |
7       -0.2724  -0.2192   22.575  0.0020        --|                  -|
8       -0.0849   0.2005   22.953  0.0034           |                   |-
9        0.1509   0.0509   24.187  0.0040           |-                  |
10       0.1549  -0.1693   25.531  0.0044           |-                 -|
11      -0.0164  -0.1411   25.547  0.0076           |                  -|
12      -0.1315  -0.0853   26.584  0.0089         -|                   |
13      -0.0708   0.1423   26.896  0.0129           |                   |-
14      -0.0276  -0.2305   26.945  0.0196           |                  -|
15       0.0109   0.1018   26.953  0.0291           |                   |
16       0.0081   0.0870   26.958  0.0420           |                   |
17      -0.0126  -0.2767   26.969  0.0585           |                 --|
18      -0.0420  -0.2606   27.104  0.0771           |                 --|
```

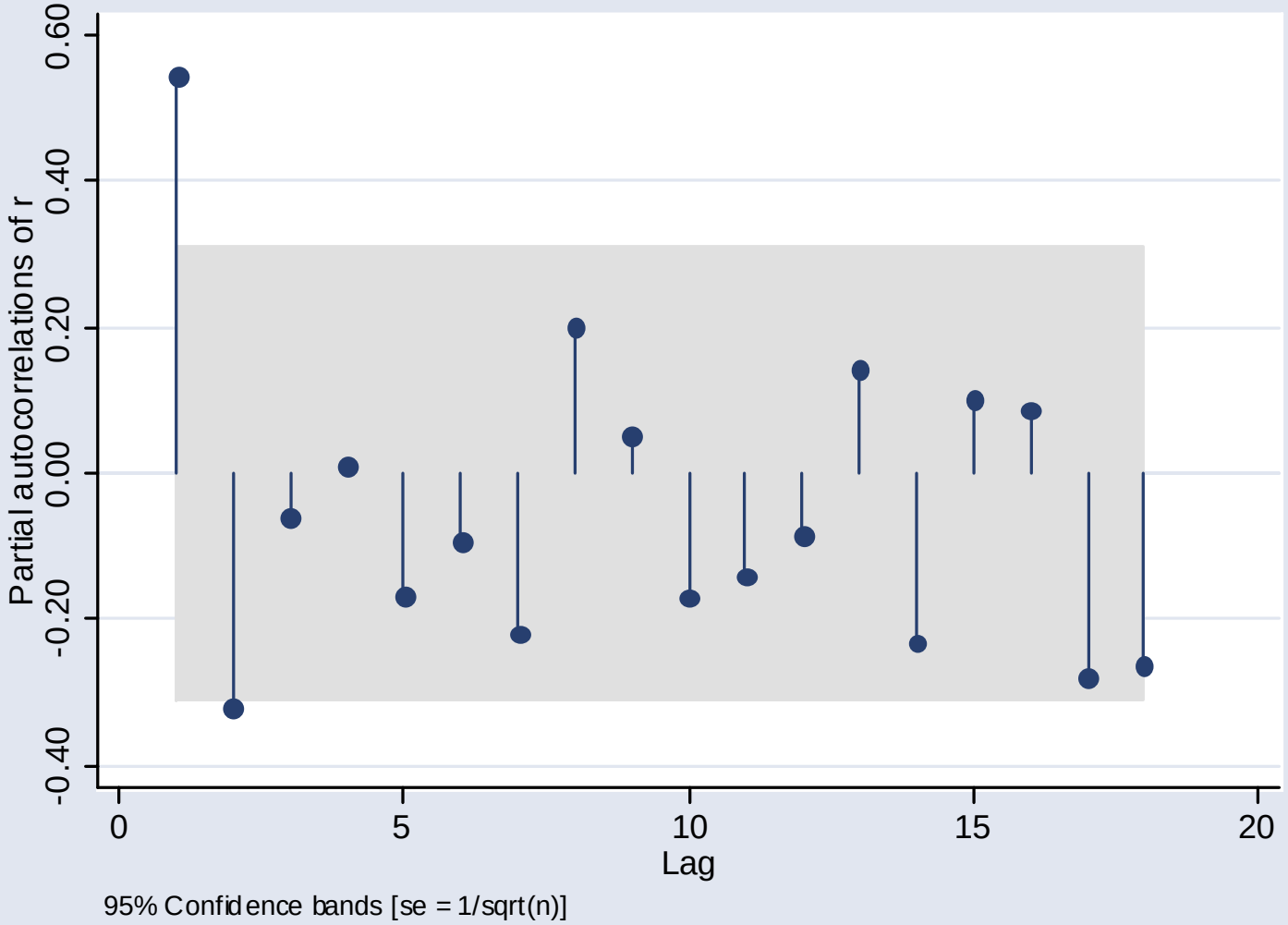
Example: Democracy Scores



First lag seems to be significant

```
ac r, lags(20)
```

Example: Democracy Scores



PAC graph confirms that process is AR(1)

```
pac r
```

```
. gen polxnewF = polxnew - .542*L.polxnew
. gen  gdpF =  gdp- .542*L.gdp
. gen  tradeF =  trade - .542*L.trade
. reg polxnewF gdpF tradeF

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  2,    36) =    0.53
       Model |  12.7652828     2   6.3826414           Prob > F      =  0.5911
    Residual |  430.727616    36   11.964656           R-squared     =  0.0288
-------------+------------------------------           Adj R-squared = -0.0252
       Total |  443.492899    38  11.6708658           Root MSE      =   3.459

------------------------------------------------------------------------------
    polxnewF |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gdpF |  -2.915114    6.85027    -0.43   0.673     -16.8081    10.97788
      tradeF |   2.538289   2.671951     0.95   0.348    -2.880678    7.957256
       _cons |   10.69697    23.8567     0.45   0.657    -37.68667     59.0806

. reg  polxnew Lgdp Ltrade
      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  2,    37) =    1.01
       Model |  34.2273459     2  17.1136729           Prob > F      =  0.3729
    Residual |  624.872654    37  16.8884501           R-squared     =  0.0519
-------------+------------------------------           Adj R-squared =  0.0007
       Total |       659.1    39        16.9           Root MSE      =  4.1096

------------------------------------------------------------------------------
     polxnew |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Lgdp |  -6.171461   5.103651    -1.21   0.234    -16.51244    4.169518
      Ltrade |   3.134679   2.206518     1.42   0.164    -1.336152     7.60551
       _cons |   49.16307   37.21536     1.32   0.195    -26.24241    124.5686
```