SOCY7706: Longitudinal Data Analysis
Instructor: Natasha Sarkisian
Two Wave Panel Data Analysis

In any longitudinal analysis, we can distinguish between analyzing trends vs individual change – that is, model the actual level of DV (Y) vs model the change in DV (ΔY). The predictors also can be either actual levels (X=time-varying, Z=time-invariant) or measures of change (ΔX; because ΔZ=0), as well as time itself (T).

We turn to the main approaches of explaining change in two wave panel dataset. We will review four main approaches.

Lagged dependent variable model: X, Z → Y
Difference score model: X, Z → ΔY
First difference model: ΔX → ΔY
Cross-lagged model: X, Z→ Y and Y, Z → X

Lagged Dependent Variable (LDV) approach
This approach is also known as regressor variable approach. The idea is to predict time 2 outcome using time 1 independent variables while controlling for stability in the outcome variable by including the dependent variable from time 1 into the model.

```
. reg  rworkhours80 l.rworkhours80

      Source |       SS       df       MS              Number of obs =    5897
-------------+------------------------------           F(  1,  5895) = 6301.45
       Model |  1609174.94     1  1609174.94           Prob > F      =  0.0000
    Residual |  1505380.87  5895  255.365712           R-squared     =  0.5167
-------------+------------------------------           Adj R-squared =  0.5166
       Total |  3114555.82  5896  528.248951           Root MSE      =   15.98

------------------------------------------------------------------------------
rworkhours80 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rworkhours80 |
         L1. |   .7368788   .0092827    79.38   0.000     .7186812    .7550763
             |
       _cons |   5.339778   .3551734    15.03   0.000     4.643507    6.036048
------------------------------------------------------------------------------

. reg  rworkhours80 l.rworkhours80 l.rallparhelptw

      Source |       SS       df       MS              Number of obs =    5767
-------------+------------------------------           F(  2,  5764) = 3081.77
       Model |   1573387.1     2  786693.548           Prob > F      =  0.0000
    Residual |  1471395.53  5764   255.27334           R-squared     =  0.5167
-------------+------------------------------           Adj R-squared =  0.5166
       Total |  3044782.63  5766  528.058034           Root MSE      =  15.977

------------------------------------------------------------------------------
rworkhours80 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rworkhours80 |
         L1. |   .7345166   .0094029    78.12   0.000     .7160834    .7529498
```

```
             |
rallparhel~w |
         L1. |  -.1601855   .0719849    -2.23   0.026    -.3013029   -.0190681
             |
       _cons |   5.483749   .3637186    15.08   0.000     4.770724    6.196774
------------------------------------------------------------------------------

. reg  rworkhours80 l.rworkhours80 l.rallparhelptw  l.rpoorhealth l.rmarried
l.rtotalpar l.rsiblog  l.hchildlg raedyrs female age  minority

      Source |       SS       df       MS              Number of obs =    5457
-------------+------------------------------           F( 11,  5445) =  582.89
       Model |  1557155.96    11  141559.633           Prob > F      =  0.0000
    Residual |  1322370.75  5445  242.859642           R-squared     =  0.5408
-------------+------------------------------           Adj R-squared =  0.5398
       Total |  2879526.71  5456  527.772491           Root MSE      =  15.584

------------------------------------------------------------------------------
rworkhours80 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rworkhours80 |
         L1. |    .668734    .010576    63.23   0.000     .6480009    .6894672
             |
rallparhel~w |
         L1. |  -.0942385   .0734988    -1.28   0.200    -.2383254    .0498485
             |
 rpoorhealth |
         L1. |   -4.44369   .5954816    -7.46   0.000    -5.611072   -3.276308
             |
    rmarried |
         L1. |   .4209347    .612163     0.69   0.492    -.7791495    1.621019
             |
   rtotalpar |
         L1. |   .2755657   .2905194     0.95   0.343    -.2939684    .8450998
             |
     rsiblog |
         L1. |    -.42027    .374524    -1.12   0.262    -1.154487    .3139468
             |
    hchildlg |
         L1. |  -.5223844    .400087    -1.31   0.192    -1.306715    .2619461
             |
     raedyrs |   .1235686   .0776308     1.59   0.111    -.0286189    .2757561
      female |  -3.392911     .46171    -7.35   0.000    -4.298048   -2.487775
         age |  -.7810018   .0711669   -10.97   0.000    -.9205174   -.6414862
    minority |  -.7411717   .5320883    -1.39   0.164    -1.784278    .3019342
       _cons |   52.52523   4.385398    11.98   0.000     43.92809    61.12236
```

We can do the same thing in wide format:

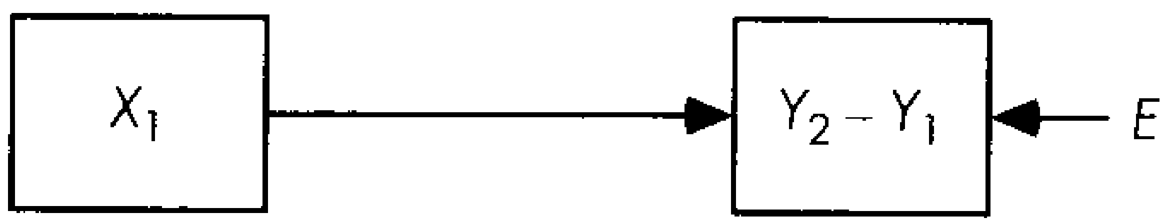
```
. reshape wide
(note: j = 1 2)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                    13182   ->    6591
Number of variables                  20   ->      26
j variable (2 values)              wave   ->   (dropped)
xij variables:
                           rworkhours80   ->   r1workhours80 r2workhours80
                            rpoorhealth   ->   r1poorhealth r2poorhealth
                               rmarried   ->   r1married r2married
```

```
                                      rtotalpar   ->   r1totalpar r2totalpar
                                        rsiblog   ->   r1siblog r2siblog
                                       hchildlg   ->   h1childlg h2childlg
                                  rallparhelptw   ->   r1allparhelptw r2allparhelptw
-----------------------------------------------------------------------------

. reg  r2workhours80  r1workhours80  r1allparhelptw  r1poorhealth r1married r1totalpar
r1siblog h1childlg age minority female raedyrs

      Source |       SS       df       MS              Number of obs =    5457
-------------+------------------------------           F( 11,  5445) =  582.89
       Model |  1557155.96    11  141559.633           Prob > F      =  0.0000
    Residual |  1322370.75  5445  242.859642           R-squared     =  0.5408
-------------+------------------------------           Adj R-squared =  0.5398
       Total |  2879526.71  5456  527.772491           Root MSE      =  15.584
------------------------------------------------------------------------------
r2workhou~80 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r1workhou~80 |    .668734    .010576    63.23   0.000     .6480009    .6894672
r1allparhe~w |  -.0942385   .0734988    -1.28   0.200    -.2383254    .0498485
r1poorhealth |   -4.44369   .5954816    -7.46   0.000    -5.611072   -3.276308
   r1married |   .4209347    .612163     0.69   0.492    -.7791495    1.621019
  r1totalpar |   .2755657   .2905194     0.95   0.343    -.2939684    .8450998
    r1siblog |    -.42027    .374524    -1.12   0.262    -1.154487    .3139468
   h1childlg |  -.5223844    .400087    -1.31   0.192    -1.306715    .2619461
         age |  -.7810018   .0711669   -10.97   0.000    -.9205174   -.6414862
    minority |  -.7411717   .5320883    -1.39   0.164    -1.784278    .3019342
      female |  -3.392911     .46171    -7.35   0.000    -4.298048   -2.487775
     raedyrs |   .1235686   .0776308     1.59   0.111    -.0286189    .2757561
       _cons |   52.52523   4.385398    11.98   0.000     43.92809    61.12236
------------------------------------------------------------------------------
```

This format also allows us to examine interactions of the effects of each of the variables of interest with the lagged DV.

<u>Difference score approach</u>

This approach is also known as the change score approach. There has been a lot of controversy surrounding this approach.




```
. reshape long
. reg d.rworkhours80 l.rallparhelptw

      Source |       SS           df       MS      Number of obs   =    28,330
-------------+----------------------------------   F(1, 28328)     =      2.40
       Model |  630.324788         1  630.324788   Prob > F        =    0.1210
    Residual |  7426545.89    28,328  262.162733   R-squared       =    0.0001
-------------+----------------------------------   Adj R-squared   =    0.0000
       Total |  7427176.22    28,329  262.175729   Root MSE        =    16.191
-------------------------------------------------------------------------------
```

```
D.            |
 rworkhours80 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
rallparhelptw |
          L1. |   .0370821   .0239149     1.55   0.121    -.0097921    .0839564
              |
        _cons |  -3.342122   .1035991   -32.26   0.000    -3.545181   -3.139063
-------------------------------------------------------------------------------

. reg d.rworkhours80 l.(rallparhelptw rpoorhealth rmarried rtotalpar rsiblog hchildlg)
raedyrs female age minority

      Source |       SS           df       MS      Number of obs   =    26,793
-------------+----------------------------------   F(10, 26782)    =      8.87
       Model |  23282.3714        10  2328.23714   Prob > F        =    0.0000
    Residual |  7027170.04    26,782  262.384066   R-squared       =    0.0033
-------------+----------------------------------   Adj R-squared   =    0.0029
       Total |  7050452.42    26,792  263.155136   Root MSE        =    16.198
-------------------------------------------------------------------------------
D.            |
 rworkhours80 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
rallparhelptw |
          L1. |   .0219595   .0250888     0.88   0.381    -.0272159    .0711348
              |
  rpoorhealth |
          L1. |   .7798301   .2583554     3.02   0.003     .2734401     1.28622
              |
     rmarried |
          L1. |   .1212796   .2801567     0.43   0.665    -.4278424    .6704015
              |
    rtotalpar |
          L1. |   .1286168   .1309732     0.98   0.326    -.1280975    .3853312
              |
      rsiblog |
          L1. |  -.2250316   .1797335    -1.25   0.211    -.5773188    .1272556
              |
     hchildlg |
          L1. |   .1074617   .1867357     0.58   0.565    -.2585501    .4734735
              |
      raedyrs |  -.0950416   .0361084    -2.63   0.008     -.165816   -.0242672
       female |   1.299652   .2043352     6.36   0.000     .8991447     1.70016
          age |  -.1017287   .0332564    -3.06   0.002     -.166913   -.0365445
     minority |   .3712403   .2551799     1.45   0.146    -.1289257    .8714063
        _cons |   2.600841   2.005298     1.30   0.195    -1.329648     6.53133
-------------------------------------------------------------------------------
```

Same in wide format:

```
. reshape wide
. gen diff= r2workhours80- r1workhours80
(694 missing values generated)

. reg diff  r1allparhelptw

      Source |       SS       df       MS              Number of obs =    5767
-------------+------------------------------           F(  1,  5765) =    0.04
       Model |  10.7404403     1  10.7404403           Prob > F      =  0.8475
    Residual |  1674892.93  5765  290.527828           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0002
       Total |  1674903.67  5766  290.479304           Root MSE      =  17.045

------------------------------------------------------------------------------
```

```
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r1allparhe~w |  -.0147277    .076598    -0.19   0.848    -.1648885    .1354331
       _cons |  -2.792029   .2297434   -12.15   0.000    -3.242412   -2.341645
------------------------------------------------------------------------------

. reg diff  r1allparhelptw r1poorhealth r1married r1totalpar r1siblog h1childlg
raedyrs female age minority
      Source |       SS       df       MS              Number of obs =    5457
-------------+------------------------------           F( 10,  5446) =    6.05
       Model |  17340.4376    10  1734.04376           Prob > F      =  0.0000
    Residual |  1560639.79  5446  286.566249           R-squared     =  0.0110
-------------+------------------------------           Adj R-squared =  0.0092
       Total |  1577980.23  5456   289.21925           Root MSE      =  16.928
------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r1allparhe~w |  -.0267496   .0798046    -0.34   0.737    -.1831985    .1296994
r1poorhealth |   .2642639   .6259046     0.42   0.673    -.9627592    1.491287
   r1married |   1.383919   .6641307     2.08   0.037     .0819573     2.68588
  r1totalpar |   .0906871   .3155152     0.29   0.774    -.5278488    .7092229
    r1siblog |  -.7903476    .406629    -1.94   0.052    -1.587503    .0068077
   h1childlg |  -.4283254   .4345873    -0.99   0.324     -1.28029    .4236395
     raedyrs |  -.1313198   .0838629    -1.57   0.117    -.2957246     .033085
      female |   1.381293   .4734211     2.92   0.004     .4531982    2.309387
         age |  -.4761804   .0765798    -6.22   0.000    -.6263073   -.3260534
    minority |   -.578333   .5779601    -1.00   0.317    -1.711366    .5546998
       _cons |   25.22486   4.668661     5.40   0.000     16.07242     34.3773
------------------------------------------------------------------------------
```

When interpreting these results, keep in mind that your dependent variable is change – so a positive coefficient would mean a larger positive change OR a smaller negative change. The baseline change (for a case with all zeroes) is represented by a constant. In our case, constant is not a very meaningful value because that would be for someone with age=0; that is why we get 25 years increase in hours of work as our constant, which is not very realistic. You might want to mean center all your continuous variables to ensure a more interpretable constant. But given a positive constant, we could say that women experience even more of an increase in hours of paid work than men (or you can say that being a woman boosts one's hours of paid work), while older individuals experience less of an increase (at some point, that becomes a decrease – after age 53, we have to talk of people experiencing more and more of a decrease in hours of paid work as they age; turning point calculated as 25.225/.476).

For many years, difference scores were criticized. One reason is their presumed unreliability – if the DV for time 1 and time 2 are positively correlated (which is pretty much always the case), then the difference score will have lower reliability than each of the time points individually, and if the correlation across time is high, that decrease in reliability will be substantial.

But Paul Allison (1990) has argued that it is not a problem – "low reliability results from the fact that in calculating the change score we differ out all the stable between-subject variation." He showed that what matters is measurement error, not unreliability – the same amount of error variance that was contained in the individual scores just appears to be more prominent once the stable component is removed, but in fact it has not changed.

The second critique is that difference score models do not account for the regression to the mean effect—the trend wherein extremely low initial scores will be followed by an increase, and extremely high scores – by a decrease. So the initial level might shape change, but if we add the lagged DV to this change score model, we are back to the LDV model, so this strategy is not useful:

```
. reg diff  r1allparhelptw r1poorhealth r1married r1totalpar r1siblog h1childlg
raedyrs female age minority r1workhours80
      Source |       SS       df       MS              Number of obs =    5457
-------------+------------------------------           F( 11,  5445) =   95.68
       Model |  255609.477    11  23237.2252           Prob > F      =  0.0000
    Residual |  1322370.75  5445  242.859642           R-squared     =  0.1620
-------------+------------------------------           Adj R-squared =  0.1603
       Total |  1577980.23  5456   289.21925           Root MSE      =  15.584
------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r1allparhe~w |  -.0942385   .0734988    -1.28   0.200    -.2383254    .0498485
r1poorhealth |   -4.44369   .5954816    -7.46   0.000    -5.611072   -3.276308
   r1married |   .4209347    .612163     0.69   0.492    -.7791495    1.621019
  r1totalpar |   .2755657   .2905194     0.95   0.343    -.2939684    .8450998
    r1siblog |    -.42027    .374524    -1.12   0.262    -1.154487    .3139468
   h1childlg |  -.5223844    .400087    -1.31   0.192    -1.306715    .2619461
     raedyrs |   .1235686   .0776308     1.59   0.111    -.0286189    .2757561
      female |  -3.392911     .46171    -7.35   0.000    -4.298048   -2.487775
         age |  -.7810018   .0711669   -10.97   0.000    -.9205174   -.6414862
    minority |  -.7411717   .5320883    -1.39   0.164    -1.784278    .3019342
r1workhou~80 |   -.331266    .010576   -31.32   0.000    -.3519991   -.3105328
       _cons |   52.52523   4.385398    11.98   0.000     43.92809    61.12236
------------------------------------------------------------------------------
```

But Allison argued that regression to the mean does not always happen (although it is common) – mostly if there are ceiling and/or floor effects (e.g., if the variable was measured in such a way that it cannot go below above a certain value and above a certain value – that is usually the case with scales, by the way); the correlation between the initial score and the increase does not have to be negative – it can be positive and then the variance of scores increases with time. Allison argues that regression to the mean is not a problem when we compare stable groups, and in such cases, difference score approach may produce better results (less bias) than LDV approach.

Evaluating regression to the mean empirically by examining a group with high scores (above 75th percentile) at time 1 and examining their distance from the mean at time 1 and time 2:

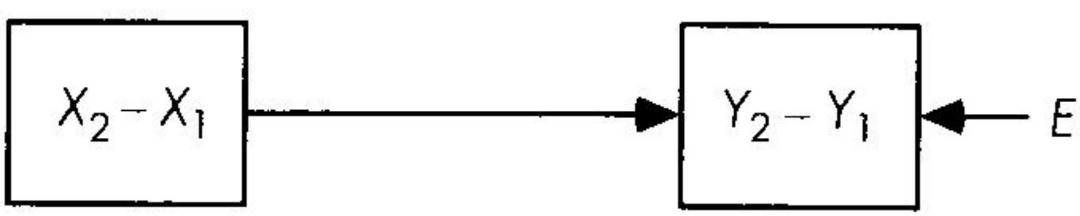
```
. for var  r1workhours80: sum X, det \ scalar Xmean1=r(mean) \ gen sample=1 if
X>r(p75)\ sum X if X>r(p75)\di r(mean)-Xmean1

->  sum r1workhours80, det
                       1 rworkhours80
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                6548
25%            0              0       Sum of Wgt.        6548

50%           40                      Mean           30.73396
                        Largest       Std. Dev.      22.52788
75%           45             80
```

```
90%           57             80       Variance       507.5055
95%           63             80       Skewness       -.175734
99%           80             80       Kurtosis       1.930742

->  scalar r1workhours80mean1=r(mean)

->  gen sample=1 if r1workhours80>r(p75)
(5020 missing values generated)

->  sum r1workhours80 if r1workhours80>r(p75)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
r1workhou~80 |      1528    57.37304     9.33897         46         80

->  di r(mean)-r1workhours80mean1
26.639072

. for var  r2workhours80: sum X \ scalar Xmean1=r(mean) \ sum X if sample==1\di
r(mean)-Xmean1
->  sum r2workhours80
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
r2workhou~80 |      5929    28.22078    23.02388          0         80

->  scalar r2workhours80mean1=r(mean)

->  sum r2workhours80 if sample==1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
r2workhou~80 |      1429    46.80476    19.63145          0         80

->  di r(mean)-r2workhours80mean1
18.583979
```

These individuals moved closer to the mean. So we conclude that regression to the mean is a problem for our data, so LDV will be better, especially if we want to document interactions between the starting level of DV and the IVs.

Moreover, recent research increasingly suggest that we should examine both LDV and change score types of models and compare findings because if assumptions are violated, they may be biased in opposite directions; e.g., see:
Ding, Peng and Fan Li. 2019. "A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. *Political Analysis* 27:605–615.

<u>First difference model</u>




```
. for any poorhealth married totalpar siblog allparhelptw: gen Xdiff=r2X-r1X
```

```
->  gen poorhealthdiff=r2poorhealth-r1poorhealth
(627 missing values generated)

->  gen marrieddiff=r2married-r1married
(625 missing values generated)

->  gen totalpardiff=r2totalpar-r1totalpar
(691 missing values generated)

->  gen siblogdiff=r2siblog-r1siblog
(325 missing values generated)

->  gen allparhelptwdiff=r2allparhelptw-r1allparhelptw
(864 missing values generated)

. for any childlg: gen Xdiff=h2X-h1X

->  gen childlgdiff=h2childlg-h1childlg
(1132 missing values generated)

. reg  diff allparhelptwdiff poorhealthdiff marrieddiff totalpardiff siblogdiff
childlgdiff
      Source |       SS       df       MS              Number of obs =    5229
-------------+------------------------------           F(  6,  5222) =    2.88
       Model |  4995.34416     6   832.55736           Prob > F      =  0.0084
    Residual |  1510362.04  5222  289.230571           R-squared     =  0.0033
-------------+------------------------------           Adj R-squared =  0.0022
       Total |  1515357.38  5228  289.854129           Root MSE      =  17.007
------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
allparhelp~f |  -.0796341   .0596778    -1.33   0.182    -.1966276    .0373593
poorhealth~f |  -2.450367   .6794682    -3.61   0.000    -3.782409   -1.118325
 marrieddiff |  -.8902544   1.360583    -0.65   0.513    -3.557567    1.777058
totalpardiff |   .5724302    .494059     1.16   0.247    -.3961321    1.540993
  siblogdiff |  -1.649011   2.908561    -0.57   0.571    -7.351007    4.052985
 childlgdiff |   1.415648   1.658858     0.85   0.393    -1.836407    4.667703
       _cons |  -2.515716    .260116    -9.67   0.000    -3.025652    -2.00578
------------------------------------------------------------------------------
```

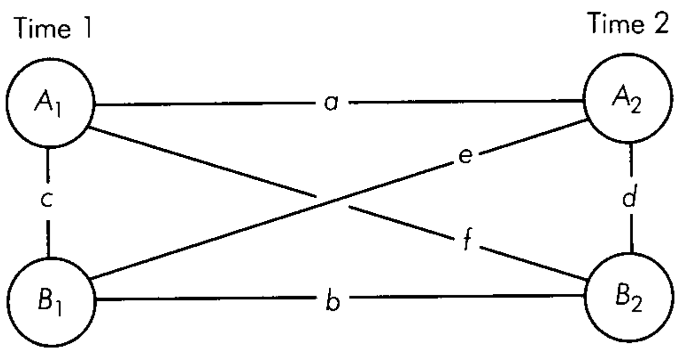
Once we created a first difference model, can we introduce time-invariant variables as well? We can; by doing that, we are assuming that the effect of this time-invariant variable is not stable over time, and interpret the resulting coefficient as an interaction term for time and that variable. That would allow us to assess how the effect of that time-invariant variable changes over time, but we would not have an estimate of that estimate at baseline.

```
 . reg  diff allparhelptwdiff poorhealthdiff marrieddiff totalpardiff siblogdiff
childlgdiff raedyrs female age minority
      Source |       SS       df       MS              Number of obs =    5227
-------------+------------------------------           F( 10,  5216) =    6.43
       Model |  18452.1386    10  1845.21386           Prob > F      =  0.0000
    Residual |  1496890.64  5216  286.980567           R-squared     =  0.0122
-------------+------------------------------           Adj R-squared =  0.0103
       Total |  1515342.77  5226  289.962261           Root MSE      =  16.941
------------------------------------------------------------------------------
        diff |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
allparhelp~f |  -.0779137   .0595081    -1.31   0.190    -.1945744     .038747
poorhealth~f |  -2.417475   .6781968    -3.56   0.000    -3.747025   -1.087926
 marrieddiff |  -.7896093   1.355911    -0.58   0.560    -3.447763    1.868545
totalpardiff |   .4298372   .4928697     0.87   0.383    -.5363938    1.396068
```

```
   siblogdiff |  -1.740446   2.905442    -0.60   0.549    -7.436328    3.955437
  childlgdiff |    1.10057   1.654093     0.67   0.506    -2.142146    4.343286
      raedyrs |  -.0944175   .0800286    -1.18   0.238    -.2513072    .0624722
       female |   1.262989   .4708219     2.68   0.007     .3399806    2.185997
          age |  -.4535023   .0760188    -5.97   0.000     -.602531   -.3044735
     minority |  -.9362349   .5703079    -1.64   0.101    -2.054277    .1818075
        _cons |   23.36358   4.426971     5.28   0.000     14.68486     32.0423
------------------------------------------------------------------------------
```

<u>Cross-lagged panel model</u>

This type of model, in many ways similar to LDV (in that it models level rather than change), is useful if you are interested in mutual effects of two variables on one another:




```
. reg r2workhours80 r1workhours80 r1allparhelptw
      Source |       SS       df       MS              Number of obs =    5767
-------------+------------------------------           F(  2,  5764) = 3081.77
       Model |   1573387.1     2  786693.548           Prob > F      =  0.0000
    Residual |  1471395.53  5764   255.27334           R-squared     =  0.5167
-------------+------------------------------           Adj R-squared =  0.5166
       Total |  3044782.63  5766  528.058034           Root MSE      =  15.977
------------------------------------------------------------------------------
r2workhou~80 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r1workhou~80 |   .7345166   .0094029    78.12   0.000     .7160834    .7529498
r1allparhe~w |  -.1601855   .0719849    -2.23   0.026    -.3013029   -.0190681
       _cons |   5.483749   .3637186    15.08   0.000     4.770724    6.196774
------------------------------------------------------------------------------

. reg r2allparhelptw r1allparhelptw r1workhours80
      Source |       SS       df       MS              Number of obs =    5697
-------------+------------------------------           F(  2,  5694) =  151.12
       Model |  3376.80486     2  1688.40243           Prob > F      =  0.0000
    Residual |  63615.6175  5694  11.1723951           R-squared     =  0.0504
-------------+------------------------------           Adj R-squared =  0.0501
       Total |  66992.4223  5696  11.7613101           Root MSE      =  3.3425
------------------------------------------------------------------------------
r2allparhe~w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r1allparhe~w |    .261863   .0151334    17.30   0.000     .2321957    .2915302
r1workhou~80 |  -.0008848   .0019782    -0.45   0.655    -.0047629    .0029932
       _cons |   1.129847   .0764159    14.79   0.000     .9800425    1.279651
------------------------------------------------------------------------------
```

To establish causal predominance, we can compare standardized effects:

```
. reg r2allparhelptw r1allparhelptw r1workhours80, beta
      Source |       SS       df       MS              Number of obs =    5697
-------------+------------------------------           F(  2,  5694) =  151.12
```

```
       Model |  3376.80486     2  1688.40243           Prob > F      =  0.0000
    Residual |  63615.6175  5694  11.1723951           R-squared     =  0.0504
-------------+------------------------------           Adj R-squared =  0.0501
       Total |  66992.4223  5696  11.7613101           Root MSE      =  3.3425
------------------------------------------------------------------------------
r2allparhe~w |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
r1allparhe~w |    .261863   .0151334    17.30   0.000                 .2240261
r1workhou~80 |  -.0008848   .0019782    -0.45   0.655                -.0057909
       _cons |   1.129847   .0764159    14.79   0.000                        .
------------------------------------------------------------------------------

. reg   r2workhours80 r1workhours80 r1allparhelptw, beta
      Source |       SS       df       MS              Number of obs =    5767
-------------+------------------------------           F(  2,  5764) = 3081.77
       Model |   1573387.1     2  786693.548           Prob > F      =  0.0000
    Residual |  1471395.53  5764   255.27334           R-squared     =  0.5167
-------------+------------------------------           Adj R-squared =  0.5166
       Total |  3044782.63  5766  528.058034           Root MSE      =  15.977

------------------------------------------------------------------------------
r2workhou~80 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
r1workhou~80 |   .7345166   .0094029    78.12   0.000                 .7171015
r1allparhe~w |  -.1601855   .0719849    -2.23   0.026                -.0204278
       _cons |   5.483749   .3637186    15.08   0.000                        .
```

A better way of modeling these same relationships is to perform simultaneous estimation with correlated residuals. We can do this with structural equation modeling (SEM).

```
. sem (r1workhours80 -> r2workhours80, ) (r1workhours80 -> r2allparhelptw, )
(r1allparhelptw -> r2workhours80, ) (r1allparhelptw -> r2allparhelptw, ), cov(
r1allparhelptw*r1workhours80 e.r2workhours80*e.r2allparhelptw) nocapslatent

Structural equation model                       Number of obs      =      5651
Estimation method  = ml
Log likelihood     =  -78231.18
--------------------------------------------------------------------------------------
                     |                 OIM
                     |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+----------------------------------------------------------------
Structural           |
  r2workhours80 <-   |
       r1workhours80 |   .7377979   .0095035    77.63   0.000     .7191713    .7564244
      r1allparhelptw |  -.1516555   .0723852    -2.10   0.036    -.2935278   -.0097831
               _cons |   5.366347   .3672892    14.61   0.000     4.646473    6.086221
  -------------------+----------------------------------------------------------------
  r2allparhelptw <-  |
       r1workhours80 |  -.0008713   .0019916    -0.44   0.662    -.0047748    .0030323
      r1allparhelptw |   .2616838   .0151696    17.25   0.000     .2319519    .2914157
               _cons |   1.133529   .0769721    14.73   0.000     .9826666    1.284392
---------------------+----------------------------------------------------------------
Mean                 |
       r1workhours80 |   30.77579   .2983912   103.14   0.000     30.19096    31.36063
      r1allparhelptw |   .6459817    .039176    16.49   0.000      .569198    .7227653
---------------------+----------------------------------------------------------------
Variance             |
     e.r2workhours80 |   255.4756     4.8062                       246.2272    265.0714
    e.r2allparhelptw |   11.22017   .2110823                       10.81399     11.6416
       r1workhours80 |   503.1497   9.465631                       484.9353    522.0483
      r1allparhelptw |   8.672941   .1631619                       8.358973    8.998701
```

```
---------------------+----------------------------------------------------------------
Covariance           |
  e.r2workhours80    |
    e.r2allparhelptw |  -1.446145   .7124758    -2.03   0.042    -2.842572   -.0497185
  -------------------+----------------------------------------------------------------
  r1workhours80      |
      r1allparhelptw |  -4.739734   .8810168    -5.38   0.000    -6.466495   -3.012972
```

We can also request standardized coefficients in SEM by using the "<u>stand</u>ardized" option.

```
. sem (r1workhours80 r1allparhelptw -> r2workhours80) (r1workhours80 r1allparhelptw ->
r2allparhelptw), cov(r1allparhelptw*r1workhours80 e.r2workhours80*e.r2allparhelptw)
nocapslatent stand

Structural equation model                       Number of obs      =      5651
Estimation method  = ml
Log likelihood     =  -78231.18
--------------------------------------------------------------------------------------
                     |                 OIM
        Standardized |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+----------------------------------------------------------------
Structural           |
  r2workhours80 <-   |
       r1workhours80 |    .718444   .0064574   111.26   0.000     .7057877    .7311003
      r1allparhelptw |  -.0193887   .0092543    -2.10   0.036    -.0375268   -.0012505
               _cons |   .2329623   .0172625    13.50   0.000     .1991284    .2667962
  -------------------+----------------------------------------------------------------
  r2allparhelptw <-  |
       r1workhours80 |  -.0056853   .0129958    -0.44   0.662    -.0311567    .0197861
      r1allparhelptw |   .2241885    .012667    17.70   0.000     .1993617    .2490153
               _cons |   .3297509   .0227067    14.52   0.000     .2852467    .3742552
---------------------+----------------------------------------------------------------
Mean                 |
       r1workhours80 |   1.372021   .0185342    74.03   0.000     1.335694    1.408347
      r1allparhelptw |   .2193497   .0134617    16.29   0.000     .1929653    .2457341
---------------------+----------------------------------------------------------------
Variance             |
     e.r2workhours80 |   .4814634    .009224                      .4637198    .4998858
    e.r2allparhelptw |   .9495243   .0056756                      .9384651    .9607137
       r1workhours80 |          1          .                             .           .
      r1allparhelptw |          1          .                             .           .
---------------------+----------------------------------------------------------------
Covariance           |
  e.r2workhours80    |
    e.r2allparhelptw |  -.0270108   .0132929    -2.03   0.042    -.0530644   -.0009572
  -------------------+----------------------------------------------------------------
  r1workhours80      |
      r1allparhelptw |    -.07175   .0132341    -5.42   0.000    -.0976885   -.0458116
--------------------------------------------------------------------------------------
```

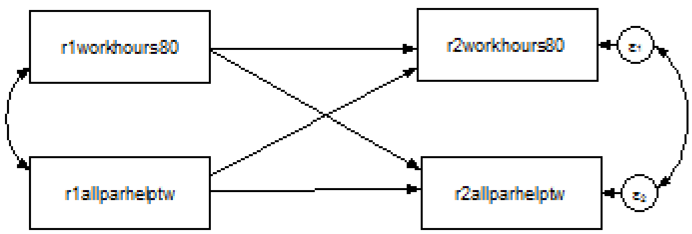
We can also test for the equivalence of coefficients to determine causal predominance. It is important to compare the standardized coefficients for this test, since the units for the b coefficients are not identical.

```
. estat stdize: test
(_b[r2workhours80:r1allparhelptw]==_b[r2allparhelptw:r1workhours80])
 ( 1)  [r2workhours80]r1allparhelptw - [r2allparhelptw]r1workhours80 = 0
           chi2(  1) =    0.74
         Prob > chi2 =    0.3909
```

Here, neither standardized coefficient is significantly larger than the other - we cannot reject the null hypothesis that they are equal ($p = 0.39$).

There are a number of advantages to using SEM for two-wave analysis (e.g., construction of latent variables, direct modeling of mediation, management of missing data via MLMV, etc.), but one of the most practical for us here is the diagramming of paths. Stata allows you to specify SEM models not only with syntax, but also by using path diagrams via its SEM Builder. (Many SEM software packages will produce path diagrams as output along with results tables; Stata only produces path diagram outputs if you specify the model using the SEM Builder.)

Using the dropdown menus, select: Statistics → SEM (structural equation modeling) → Model building and estimation. As a note, in SEM measured variables are represented using rectangles, while latent variables are represented by ellipses. Ordinary regression *only uses measured variables*, so for our purposes here all you need to know is that our variables will be represented using rectangles. In the SEM Builder, we can specify a model that matches the path diagram in the notes above:

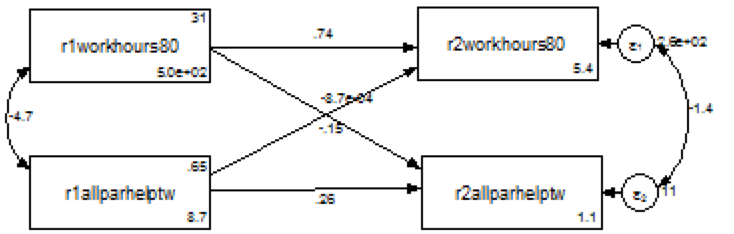


Click the "Estimate" button in the upper right-hand corner and hit "OK" for Maximum likelihood estimation, and Stata will perform this simple cross-lagged SEM model.

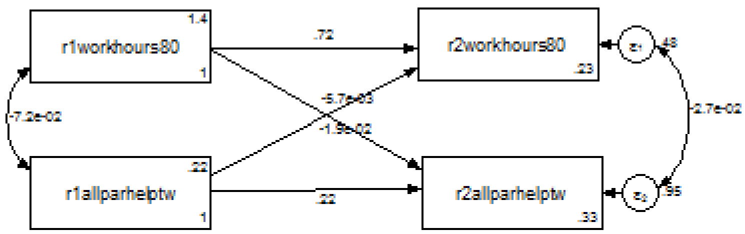
```
Structural equation model                       Number of obs      =      5651
Estimation method  = ml
Log likelihood     =  -78231.18
------------------------------------------------------------------------------
                     |                 OIM
                     |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------
Structural           |
  r2workhours80 <-   |
       r1workhours80 |   .7377979   .0095035    77.63   0.000     .7191713    .7564244
      r1allparhelptw |  -.1516555   .0723852    -2.10   0.036    -.2935278   -.0097831
               _cons |   5.366347   .3672892    14.61   0.000     4.646473    6.086221
  -------------------+--------------------------------------------------------
  r2allparhelptw <-  |
       r1workhours80 |  -.0008713   .0019916    -0.44   0.662    -.0047748    .0030323
      r1allparhelptw |   .2616838   .0151696    17.25   0.000     .2319519    .2914157
               _cons |   1.133529   .0769721    14.73   0.000     .9826666    1.284392
---------------------+--------------------------------------------------------
Mean                 |
       r1workhours80 |   30.77579   .2983912   103.14   0.000     30.19096    31.36063
      r1allparhelptw |   .6459817    .039176    16.49   0.000      .569198    .7227653
---------------------+--------------------------------------------------------
Variance             |
     e.r2workhours80 |   255.4756     4.8062                       246.2272    265.0714
    e.r2allparhelptw |   11.22017   .2110823                       10.81399     11.6416
       r1workhours80 |   503.1497   9.465631                       484.9353    522.0483
      r1allparhelptw |   8.672941   .1631619                       8.358973    8.998701
---------------------+--------------------------------------------------------
Covariance           |
  e.r2workhours80    |
    e.r2allparhelptw |  -1.446145   .7124758    -2.03   0.042    -2.842572   -.0497185
```

```
-------------------+----------------------------------------------------------------
r1workhours80      |
    r1allparhelptw |  -4.739734   .8810168    -5.38   0.000    -6.466495   -3.012972
```

Note that the results are exactly the same whether the model is estimated using syntax or the SEM Builder.



SEM also allows for standardized coefficients to be reported in the SEM Builder diagram, even when the "standardized" option wasn't requested at estimation. You can report the standardized coefficients on the paths in the diagram by selecting View → Standardized estimates.



Special assumptions of this type of analysis:

- Finite causal lag corresponding to our measurement: In such models, we are assuming that causal process happens with a specific lag, and the distance between time points in our dataset reflects, or closely approximates that lag.
- Continuity of causal process: This model assumes that the causal processes are continuous and ongoing so we can observe that at any time.
- Equality of causal lags: We assume that A→B and B→A causal lag is of the same length.

Cross-lagged models can be used for more than two waves, but some recent work has suggested a useful modification for such analyses (using SEM) – if interested, see:
Hamaker, Ellen L., Rebecca M. Kuiper, and Raoul P. P. P. Grasman. 2015. "A critique of the cross-lagged panel model." *Psychological Methods 20(1)*: 102-116.
Diagnostics for longitudinal data with two time points:

Since the vast majority of the models we discussed can be estimated using OLS regression, diagnostics should be conducted the same way as they are for OLS.