

SC706: Longitudinal Data Analysis

Instructor: Natasha Sarkisian

Two Wave Panel Data Analysis

First, let's examine some tools that allow us to examine and describe change in the data. We will use an example from HRS data that is focusing on employment and caregiving.

```
. use http://www.sarkisian.net/sc706/hrs\_hours.dta

. reshape long r@workhours80 r@poorhealth r@married r@totalpar r@siblog h@childlg
r@allparhelptw, i(hhid pn) j(wave)
(note: j = 1 2 3 4 5 6 7 8 9)

Data                                wide  ->  long
-----
Number of obs.                      6591  ->  59319
Number of variables                  75    ->   20
j variable (9 values)                ->   wave
xij variables:
r1workhours80 r2workhours80 ... r9workhours80->rworkhours80
r1poorhealth r2poorhealth ... r9poorhealth-> rpoorhealth
  r1married r2married ... r9married -> rmarried
  r1totalpar r2totalpar ... r9totalpar -> rtotalpar
  r1siblog r2siblog ... r9siblog -> rsiblog
  h1childlg h2childlg ... h9childlg -> hchildlg
r1allparhelptw r2allparhelptw ... r9allparhelptw->rallparhelptw
-----

. tab wave
      wave |          Freq.      Percent      Cum.
-----+-----
          1 |           6,591         11.11      11.11
          2 |           6,591         11.11      22.22
          3 |           6,591         11.11      33.33
          4 |           6,591         11.11      44.44
          5 |           6,591         11.11      55.56
          6 |           6,591         11.11      66.67
          7 |           6,591         11.11      77.78
          8 |           6,591         11.11      88.89
          9 |           6,591         11.11     100.00
-----+-----
      Total |          59,319      100.00

. keep if wave<3
(46137 observations deleted)
```

Stata provides a number of tools for analyzing panel data. The commands all begin with the prefix `xt`. To use these commands, we first need to tell Stata that our dataset is a panel dataset. We need to have a variable that identifies the units (for example, a country or person id) and a time variable. To set a dataset as a panel, we need to use `xtset` command:

```
. xtset hhidpn wave
      panel variable: hhidpn (strongly balanced)
      time variable: wave, 1 to 2
      delta: 1 unit
```

Stata thinks the dataset is strongly balanced, meaning all units are observed at all time points (at the same time and equal number of times). But it is not true – we just have empty rows that were created when we went from wide to long format.

```
. xtides
      hhidpn: 10003020, 10004010, ..., 99564010      n =      6591
      wave: 1, 2, ..., 2                          T =      2
      Delta(wave) = 1 unit
      Span(wave) = 2 periods
      (hhidpn*wave uniquely identifies each observation)
```

```
Distribution of T_i:  min      5%      25%      50%      75%      95%      max
                   2         2         2         2         2         2         2
```

```
-----+-----
      Freq.  Percent  Cum.  |  Pattern
-----+-----
      6591   100.00  100.00 |  11
-----+-----
      6591   100.00          |  XX
```

Xtides also thinks all cases are complete. We will now delete those empty records to have a more accurate picture. Note that those rows are not completely empty – time-invariant variables still have values there, but the time-variant ones are empty. So we will only specify time-varying variables in the egen command:

```
. egen miss=rowmiss( rworkhours80 rpoorhealth rmarried rtotalpar rsiblog hchildlg
rallparhelptw)
```

```
. tab miss
```

```
-----+-----
      miss |      Freq.      Percent      Cum.
-----+-----
      0 |     11,327      85.93      85.93
      1 |      1,017       7.72      93.64
      2 |         115       0.87      94.52
      3 |          90       0.68      95.20
      4 |           7       0.05      95.25
      5 |           3       0.02      95.27
      6 |          364       2.76      98.04
      7 |          259       1.96     100.00
-----+-----
      Total |     13,182     100.00
```

```
. drop if miss==7
(259 observations deleted)
```

```
. xtset hhidpn wave
      panel variable: hhidpn (unbalanced)
      time variable: wave, 1 to 2
      delta: 1 unit
```

This is more accurate now, and xtides also shows that there are missing observations at time 2.

```
. xtides
      hhidpn: 10003020, 10004010, ..., 99564010      n =      6591
      wave: 1, 2, ..., 2                          T =      2
      Delta(wave) = 1 unit
```

```
Span(wave) = 2 periods
(hhidpn*wave uniquely identifies each observation)
```

```
Distribution of T_i:  min      5%      25%      50%      75%      95%      max
                    1         2         2         2         2         2         2
```

Freq.	Percent	Cum.	Pattern
6332	96.07	96.07	11
259	3.93	100.00	1.
6591	100.00		XX

Next, let's examine change in a continuous variable.

```
. xtsum rworkhours80
```

Variable	Mean	Std. Dev.	Min	Max	Observations
rwork~80 overall	29.53971	22.79859	0	80	N = 12477
between		21.33473	0	80	n = 6580
within		8.392351	-10.46029	69.53971	T-bar = 1.8962

Here we see overall standard deviation along with between and within standard deviations – between indicates the amount of variation across individuals (cross-sectional variation, or differences among individuals that are stable over time), and within indicates change over time within individuals (temporal variation). Between variation is essentially variation of average values for individuals over time, and within variation is variation in differences between values at each time point and averages for a given individual (i.e. individual's deviation from their own overall mean). That is why the minimum and maximum differ from those for overall and between, and can be negative. Observation column shows that there are 12477 records, 6580 individuals, and an average of 1.8962 time points per person.

To examine change in categorical variables, we can use both xttab and xttrans.

```
. xttab rmarried
```

rmarried	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
0	2662	21.20	1532	23.24	92.13
1	9895	78.80	5300	80.41	97.73
Total	12557	100.00	6832	103.66	96.47

(n = 6591)

Here we can see that overall, out of all records in the data, 78.8% indicate that the person is currently married, and 21.2% indicate that the person is currently single. Between percent indicates that 80.41% of all individuals in the data were married at some point during the study (or in this case that means that they were married at either wave 1 or wave 2), and 23.24% of individuals were single at some point during the study period. Within percent indicates that among those individuals that were married at some point, they were married 97.73% of all of their data points, and among those who were single at some point, they were single 92.13% of all of their data points.

```
. xttrans rmarried, freq
```

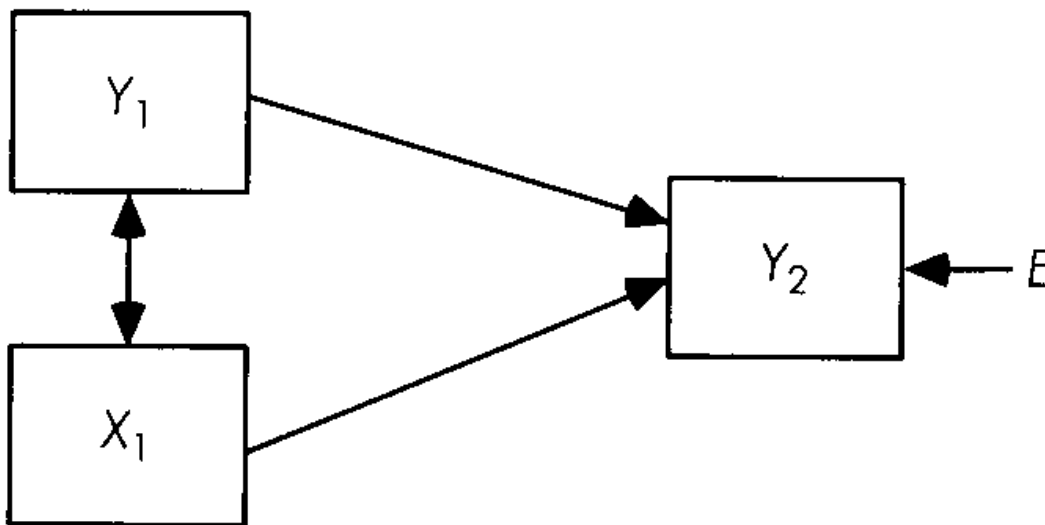
rmarried	rmarried		Total
	0	1	
0	1,130 92.85	87 7.15	1,217 100.00
1	154 3.24	4,595 96.76	4,749 100.00
Total	1,284 21.52	4,682 78.48	5,966 100.00

Xttrans shows transitions among statuses: so here we see that among those who were married at time point 1, 96.76% were still married at time point 2, while 3.24% were no longer married. Of those who were single at time 1, 92.85% were still single at time 2 and 7.15% were no longer single.

Now that we know how to describe change, we turn to the main approaches of explaining change in two wave panel dataset. We will review four main approaches.

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	
Model	1609174.94	1	1609174.94	Number of obs = 5897
Residual	1505380.87	5895	255.365712	F(1, 5895) = 6301.45
				Prob > F = 0.0000
				R-squared = 0.5167

```
-----+-----
Total | 3114555.82  5896  528.248951
Adj R-squared = 0.5166
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
-----+-----
Model | 1573387.1      2  786693.548          F( 2, 5764) = 3081.77
Residual | 1471395.53  5764  255.27334          Prob > F      = 0.0000
-----+-----
Total | 3044782.63  5766  528.058034          R-squared     = 0.5167
Adj R-squared = 0.5166
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. |     -0.1601855   .0719849    -2.23   0.026   -0.3013029  -0.0190681
  |
  _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
-----+-----
Model | 1557155.96     11  141559.633          F( 11, 5445) = 582.89
Residual | 1322370.75  5445  242.859642          Prob > F      = 0.0000
-----+-----
Total | 2879526.71  5456  527.772491          R-squared     = 0.5408
Adj R-squared = 0.5398
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. |     -0.0942385   .0734988    -1.28   0.200   -0.2383254   .0498485
rpoorhealth |
  L1. |     -4.44369   .5954816    -7.46   0.000   -5.611072  -3.276308
rmarried |
  L1. |      .4209347   .612163     0.69   0.492   -0.7791495   1.621019
rtotalpar |
  L1. |      .2755657   .2905194     0.95   0.343   -0.2939684   .8450998
-----+-----
```

```

rsiblog |
  L1. |    -.42027    .374524   -1.12   0.262   -1.154487   .3139468
|
hchildlg |
  L1. |   -5.223844   .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
```

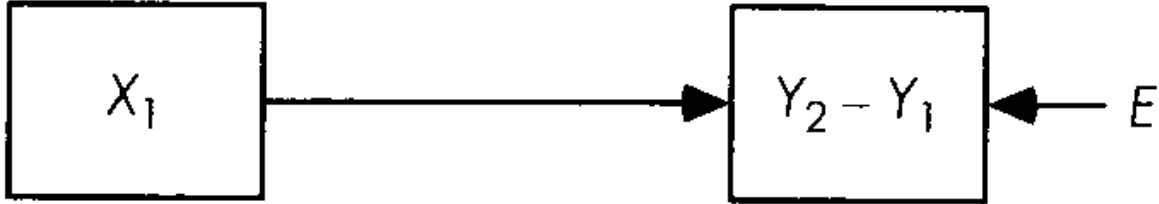
Source	SS	df	MS	Number of obs = 5457		
Model	1557155.96	11	141559.633	F(11, 5445)	=	582.89
Residual	1322370.75	5445	242.859642	Prob > F	=	0.0000
Total	2879526.71	5456	527.772491	R-squared	=	0.5408
				Adj R-squared	=	0.5398
				Root MSE	=	15.584

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	-5.223844	.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.

Difference score approach

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



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

```
. reg diff r1allparhelptw
```

Source	SS	df	MS	Number of obs =	5767
Model	10.7404403	1	10.7404403	F(1, 5765) =	0.04
Residual	1674892.93	5765	290.527828	Prob > F =	0.8475
Total	1674903.67	5766	290.479304	R-squared =	0.0000
				Adj R-squared =	-0.0002
				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
Model	17340.4376	10	1734.04376	F(10, 5446) =	6.05
Residual	1560639.79	5446	286.566249	Prob > F =	0.0000
Total	1577980.23	5456	289.21925	R-squared =	0.0110
				Adj R-squared =	0.0092
				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

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 more recently, 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 error variance, not unreliability.

Furthermore, change score models control for any unobserved factors as long as their effects are stable over time, while the lagged dependent variable models do not, so this is a big advantage of change score models.

The second critique is that difference score models do not account for the regression to the mean effect—the phenomenon when 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 rlallparhelptw rlpoorhealth rlmarrried rltotalpar rlsiblog hlchildlg
raedyrs female age minority rlworkhours80
```

Source	SS	df	MS	Number of obs =	5457
Model	255609.477	11	23237.2252	F(11, 5445) =	95.68
Residual	1322370.75	5445	242.859642	Prob > F =	0.0000
				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]
rlallparhe~w	-.0942385	.0734988	-1.28	0.200	-.2383254 .0498485
rlpoorhealth	-4.44369	.5954816	-7.46	0.000	-5.611072 -3.276308
rlmarrried	.4209347	.612163	0.69	0.492	-.7791495 1.621019
rltotalpar	.2755657	.2905194	0.95	0.343	-.2939684 .8450998
rlsiblog	-.42027	.374524	-1.12	0.262	-1.154487 .3139468
hlchildlg	-.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
rlworkhou~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; 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 at time 1 and examining their distance from the mean at time 1 and time 2:

```
. for var v661972: 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 v661972, det
```

1972 v66

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	113
25%	720	0	Sum of Wgt.	113
50%	8500		Mean	267416.8
		Largest	Std. Dev.	1068867
75%	95600	1806401		
90%	482000	2419401	Variance	1.14e+12
95%	1227901	4540001	Skewness	7.259941
99%	4540001	9900000	Kurtosis	61.94597

```
-> scalar v661972mean1=r(mean)
```

```
-> gen sample=1 if v661972>r(p75)
(85 missing values generated)
```

```
-> sum v661972 if v661972>r(p75)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
v661972	28	1038998	1979140	98000	9900000

```
-> di r(mean)-v661972mean1
771581.47
```

```
. for var v661973: sum X \ scalar Xmean1=r(mean) \ sum X if sample==1\di r(mean)-Xmean1
```

```
-> sum v661973
```

Variable	Obs	Mean	Std. Dev.	Min	Max
v661973	113	199197	593450.7	0	4920001

```
-> scalar v661973mean1=r(mean)
```

```
-> sum v661973 if sample==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
v661973	28	759401.6	1013181	99500	4920001

```
-> di r(mean)-v661973mean1
560204.62
```

The difference between overall mean and the mean for the selected “high” group was much larger in 1972 than in 1973, which means that these people experienced regression toward the

mean. We could also select instead cases that are one SD above the mean (rather than above 75%) but here that would be too few countries. We can do the same test for lower 25th percentile:

```
. 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
                                Largest      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
```

```

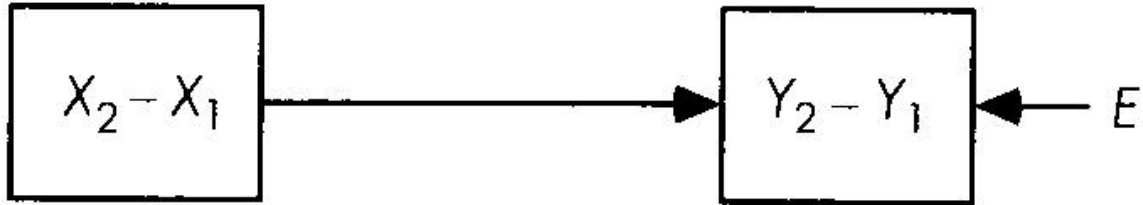
Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
r2workhou~80 |    1429    46.80476    19.63145         0        80

```

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

These individuals also 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.

First difference model



```
. 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
Model	4995.34416	6	832.55736	F(6, 5222) =	2.88
Residual	1510362.04	5222	289.230571	Prob > F =	0.0084
Total	1515357.38	5228	289.854129	R-squared =	0.0033
				Adj R-squared =	0.0022
				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.

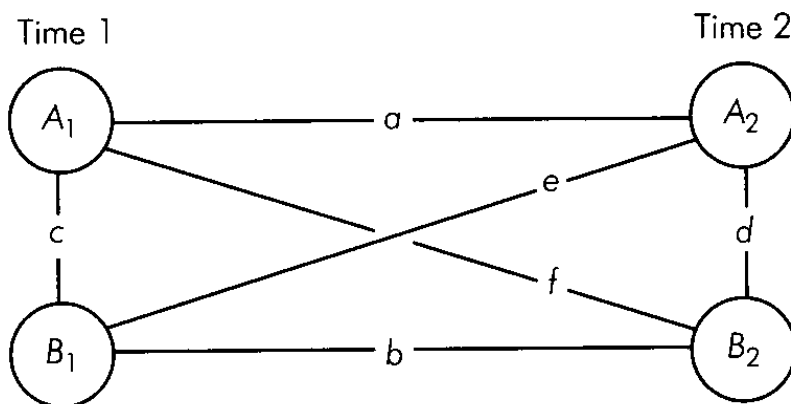
```
. reg diff allparhelptwdiff poorhealthdiff marrieddiff totalpardiff siblogdiff
childlgdiff raedyrs female age minority
```

Source	SS	df	MS	Number of obs =	5227
Model	18452.1386	10	1845.21386	F(10, 5216) =	6.43
Residual	1496890.64	5216	286.980567	Prob > F =	0.0000
				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

Cross-lagged panel model

Another type of change model, in many ways similar to LDV, is useful if you are interested in mutual effects of two variables on one another:



```
. reg r2workhours80 r1workhours80 rlallparhelptw
```

Source	SS	df	MS	Number of obs =	5767
Model	1573387.1	2	786693.548	F(2, 5764) =	3081.77
Residual	1471395.53	5764	255.27334	Prob > F =	0.0000
				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]
--------------	-------	-----------	---	------	----------------------

```
-----+-----
```

rlworkhou~80		.7345166	.0094029	78.12	0.000	.7160834	.7529498
rlallparhe~w		-.1601855	.0719849	-2.23	0.026	-.3013029	-.0190681
_cons		5.483749	.3637186	15.08	0.000	4.770724	6.196774

```
-----+-----
```

```
. reg r2allparhelptw rlallparhelptw rlworkhours80
```

```
-----+-----
```

Source		SS	df	MS		Number of obs =	5697
Model		3376.80486	2	1688.40243		F(2, 5694) =	151.12
Residual		63615.6175	5694	11.1723951		Prob > F =	0.0000
						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]
rlallparhe~w		.261863	.0151334	17.30	0.000	.2321957 .2915302
rlworkhou~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 rlallparhelptw rlworkhours80, beta
```

```
-----+-----
```

Source		SS	df	MS		Number of obs =	5697
Model		3376.80486	2	1688.40243		F(2, 5694) =	151.12
Residual		63615.6175	5694	11.1723951		Prob > F =	0.0000
						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
rlallparhe~w		.261863	.0151334	17.30	0.000	.2240261
rlworkhou~80		-.0008848	.0019782	-0.45	0.655	-.0057909
_cons		1.129847	.0764159	14.79	0.000	.

```
-----+-----
```

```
. reg r2workhours80 rlworkhours80 rlallparhelptw, beta
```

```
-----+-----
```

Source		SS	df	MS		Number of obs =	5767
Model		1573387.1	2	786693.548		F(2, 5764) =	3081.77
Residual		1471395.53	5764	255.27334		Prob > F =	0.0000
						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
rlworkhou~80		.7345166	.0094029	78.12	0.000	.7171015
rlallparhe~w		-.1601855	.0719849	-2.23	0.026	-.0204278
_cons		5.483749	.3637186	15.08	0.000	.

```
-----+-----
```

Simultaneous estimation with correlated residuals:

```
. reg3 ( r2workhours80 rlworkhours80 rlallparhelptw) (r2allparhelptw rlallparhelptw
rlworkhours80), corr(unstr)
```

Three-stage least-squares regression

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
r2workhou~80	5651	2	15.9836	0.5185	6086.13	0.0000
r2allparhe~w	5651	2	3.349652	0.0505	300.40	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
r2workhou~80						
rlworkhou~80	.7377979	.0095035	77.63	0.000	.7191713	.7564244
rlallparhe~w	-.1516555	.0723852	-2.10	0.036	-.2935278	-.0097831
_cons	5.366347	.3672892	14.61	0.000	4.646473	6.086221
r2allparhe~w						
rlallparhe~w	.2616838	.0151696	17.25	0.000	.2319519	.2914157
rlworkhou~80	-.0008713	.0019916	-0.44	0.662	-.0047748	.0030323
_cons	1.133529	.0769721	14.73	0.000	.9826666	1.284392

Endogenous variables: r2workhours80 r2allparhelptw

Exogenous variables: rlworkhours80 rlallparhelptw

Unfortunately, no beta option, but we can standardize the variables of interest:

```
. for var rlworkhours80 rlallparhelptw r2workhours80 r2allparhelptw: sum X \ gen
X_st=(X-r(mean))/r(sd)
```

```
-> sum rlworkhours80
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rlworkhou~80	6548	30.73396	22.52788	0	80

```
-> gen rlworkhours80_st=(rlworkhours80-r(mean))/r(sd)
(43 missing values generated)
```

```
-> sum rlallparhelptw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rlallparhe~w	6438	.6381431	2.93342	0	19.23077

```
-> gen rlallparhelptw_st=(rlallparhelptw-r(mean))/r(sd)
(153 missing values generated)
```

```
-> sum r2workhours80
```

Variable	Obs	Mean	Std. Dev.	Min	Max
r2workhou~80	5929	28.22078	23.02388	0	80

```
-> gen r2workhours80_st=(r2workhours80-r(mean))/r(sd)
(662 missing values generated)
```

```
-> sum r2allparhelptw
```

Variable	Obs	Mean	Std. Dev.	Min	Max
r2allparhe~w	5764	1.266799	3.431998	0	19.23077

```
-> gen r2allparhelptw_st=(r2allparhelptw-r(mean))/r(sd)
(827 missing values generated)
```

```
. reg3 ( r2workhours80_st rlworkhours80_st rlallparhelptw_st) (r2allparhelptw_st
rlallparhelptw_st rlworkhours80_st), corr(unstr)
```

Three-stage least-squares regression

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
r2workhour~t	5651	2	.6942185	0.5185	6086.13	0.0000
r2allparhe~t	5651	2	.9760063	0.0505	300.40	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
r2workhour~t						
rlworkhour~t	.7219036	.0092988	77.63	0.000	.7036784	.7401289
rlallparhe~t	-.0193221	.0092224	-2.10	0.036	-.0373977	-.0012464
_cons	-.011977	.009235	-1.30	0.195	-.0300772	.0061233
r2allparhe~t						
rlallparhe~t	.2236681	.0129659	17.25	0.000	.1982554	.2490807
rlworkhour~t	-.0057191	.0130732	-0.44	0.662	-.0313421	.0199039
_cons	.0020235	.0129835	0.16	0.876	-.0234238	.0274707

Endogenous variables: r2workhours80_st r2allparhelptw_st

Exogenous variables: rlworkhours80_st rlallparhelptw_st

```
. test [r2allparhelptw_st]rlworkhours80_st = [r2workhours80_st]rlallparhelptw_st
( 1) - [r2workhours80_st]rlallparhelptw_st + [r2allparhelptw_st]rlworkhours80_st = 0
      chi2( 1) =    0.72
      Prob > chi2 =    0.3956
```

We could do the same thing with the long dataset as well:

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

Data	wide	->	long
Number of obs.	6591	->	13182
Number of variables	38	->	32
j variable (2 values)		->	wave
xij variables:			
rlworkhours80 r2workhours80		->	rworkhours80
rlpoorhealth r2poorhealth		->	rpoorhealth
rlmarried r2married		->	rmarried
r1totalpar r2totalpar		->	rtotalpar
r1siblog r2siblog		->	rsiblog
h1childlg h2childlg		->	hchildlg
rlallparhelptw r2allparhelptw		->	rallparhelptw

```

. xtset hhidpn wave
      panel variable:  hhidpn (strongly balanced)
      time variable:  wave, 1 to 2
              delta:  1 unit

. reg3 ( rworkhours80 l.rworkhours80 l.rallparhelptw) ( rallparhelptw l.rallparhelptw
l.rworkhours80), corr(unstr)

```

Three-stage least-squares regression

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
rworkhours80	5651	2	15.9836	0.5185	6086.13	0.0000
rallparhel~w	5651	2	3.349652	0.0505	300.40	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
rworkhours80						
rworkhours80						
L1.	.7377979	.0095035	77.63	0.000	.7191713	.7564244
rallparhel~w						
L1.	-.1516555	.0723852	-2.10	0.036	-.2935278	-.0097831
_cons	5.366347	.3672892	14.61	0.000	4.646473	6.086221
rallparhel~w						
rallparhel~w						
L1.	.2616838	.0151696	17.25	0.000	.2319519	.2914157
rworkhours80						
L1.	-.0008713	.0019916	-0.44	0.662	-.0047748	.0030323
_cons	1.133529	.0769721	14.73	0.000	.9826666	1.284392

Endogenous variables: rworkhours80 rallparhelptw
Exogenous variables: L.rworkhours80 L.rallparhelptw

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 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 $A \rightarrow B$ and $B \rightarrow A$ causal lag is of the same length.

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.