

**SC706: Longitudinal Data Analysis**  
**Instructor: Natasha Sarkisian**

**Event history analysis: Discrete-time Models**

Event history analysis is the method for analyzing change over time in a discrete outcome. It focuses on explaining the occurrence and timing of events.

**Synonyms:**

Hazard rate modeling, survival analysis, life history (or lifetime) analysis, failure time (or reliability) analysis

**Examples of possible outcomes:**

Mortality (dead/alive)

Labor force (durations employed, unemployed, out of the labor force, job mobility)

Life course (first sex, cohabitation, marriage, divorce, remarriage, fertility)

Organizations (births, deaths, mergers)

Income (spells of poverty, spells on welfare)

Education (school dropout, college enrollment, attentiveness of school children)

Criminology (recidivism, arrests)

Collective action (strikes, lynching)

**Types of events:**

\* Repeatable vs. non-repeatable events:

Non-repeatable – the transition from 0 to 1 happens only once (e.g. death, first childbirth)

Repeatable – the transitions may happen multiple times (e.g. job mobility, spells on welfare)

\*Single vs. multiple event types: There may be only one event type that we consider (e.g. death) or we may consider as alternative types of events: death from cancer, death from heart attack, etc.

For now we'll deal with single, non-repeatable events.

**Two main types of methods:**

Discrete-time (grouped data) models and continuous-time models. Time is always measured in discrete units, but when those are large and relatively few, we use discrete time methodology, and otherwise – continuous-time methods.

**Discrete-time models**

Example: childcare.dta

A hypothetical sample of 300 mothers. They were observed for a maximum of 5 years, beginning with the time when they child was 1 y.o.

First, let's examine what we have:

```
. tab1 time- marriage5
```

```
-> tabulation of time
```

time	Freq.	Percent	Cum.
1	17	5.67	5.67
2	33	11.00	16.67
3	55	18.33	35.00
4	68	22.67	57.67
5	127	42.33	100.00
Total	300	100.00	

```
-> tabulation of childcare
```

child placed in child care	Freq.	Percent	Cum.
0	91	30.33	30.33
1	209	69.67	100.00
Total	300	100.00	

```
-> tabulation of girl
```

child's gender	Freq.	Percent	Cum.
0	176	58.67	58.67
1	124	41.33	100.00
Total	300	100.00	

```
-> tabulation of age
```

mother's age at childbirth	Freq.	Percent	Cum.
18	2	0.67	0.67
19	3	1.00	1.67
21	3	1.00	2.67
22	9	3.00	5.67
23	11	3.67	9.33
24	34	11.33	20.67
25	24	8.00	28.67
26	6	2.00	30.67
27	30	10.00	40.67
29	6	2.00	42.67
30	24	8.00	50.67
31	24	8.00	58.67
32	13	4.33	63.00
33	31	10.33	73.33
34	6	2.00	75.33
35	10	3.33	78.67
36	11	3.67	82.33
37	16	5.33	87.67
38	6	2.00	89.67
39	5	1.67	91.33
40	13	4.33	95.67
42	12	4.00	99.67
52	1	0.33	100.00
Total	300	100.00	

-> tabulation of marriage1

mother's marital status at time 1	Freq.	Percent	Cum.
0	49	16.33	16.33
1	251	83.67	100.00
Total	300	100.00	

-> tabulation of marriage2

mother's marital status at time 2	Freq.	Percent	Cum.
0	38	13.43	13.43
1	245	86.57	100.00
Total	283	100.00	

-> tabulation of marriage3

mother's marital status at time 3	Freq.	Percent	Cum.
0	26	10.40	10.40
1	224	89.60	100.00
Total	250	100.00	

-> tabulation of marriage4

mother's marital status at time 4	Freq.	Percent	Cum.
0	12	6.15	6.15
1	183	93.85	100.00
Total	195	100.00	

-> tabulation of marriage5

mother's marital status at time 5	Freq.	Percent	Cum.
0	5	3.94	3.94
1	122	96.06	100.00
Total	127	100.00	

Note that observations for any specific respondent may stop before time 5 for two reasons: either because the respondent experienced the event (oftentimes called **failure**), or because we lost track of the respondent (the respondent is **censored**). All the respondents who did not experience an event by the time the study ended are also censored. For censored cases, we don't know when (and whether) the failure occurred.

Because of censoring as well as because the dependent variable is not continuous, we cannot use regular regression. But there is still a fairly easy solution -- to transform the data and use logistic regression.

Since marriage is a special case here (marital status changes over time), we'll come back to it later -- for now, we'll just use marriage at the first time point and drop the rest.

```
. drop marriage2-marriage5
```

We want to have one record for every year that the respondent is in the dataset:

```
*Generating subject identifier
. gen id = _n
. lab var id "subject ID"

*Episode-splitting: transforming the data into the person-year format
. expand time
. sort id

*Generating spell yeah identifier, by subject
. quietly by id: gen t = _n
. lab var t "spell year"

*Generating the binary dependent variable for discrete-time hazard model
. quietly by id: gen care = (childcare==1 & _n==_N)
. lab var care "placed in child care"
```

Now let's examine the results:

```
. tabl t care
```

```
-> tabulation of t
spell year |          Freq.      Percent      Cum.
-----+-----
          1 |             300       25.97       25.97
          2 |             283       24.50       50.48
          3 |             250       21.65       72.12
          4 |             195       16.88       89.00
          5 |             127       11.00      100.00
-----+-----
        Total |           1,155      100.00
```

```
-> tabulation of care
placed in |
child care |          Freq.      Percent      Cum.
-----+-----
          0 |             946       81.90       81.90
          1 |             209       18.10      100.00
-----+-----
        Total |           1,155      100.00
```

```
. tab care t, col
```

```

+-----+
| Key |
+-----+
|     |
|     |
|     |
|     |
|     |
+-----+

```

placed in child care	spell year					Total
	1	2	3	4	5	
0	283 94.33	254 89.75	204 81.60	137 70.26	68 53.54	946 81.90
1	17 5.67	29 10.25	46 18.40	58 29.74	59 46.46	209 18.10
Total	300 100.00	283 100.00	250 100.00	195 100.00	127 100.00	1,155 100.00

At each time, the risk set includes those women who are still “at risk” of placing the child in child care. 300 women are at risk during the first year, 283 at risk during the second year, etc.

Hazard rate  $H(t)$  is the probability that someone who is at risk at time  $T$  will experience the event at that time. Using the risk set and the number of women who experienced the event, we can calculate the hazard rate at each time point. In fact, the table above displays the hazard rate. Here, the hazard rate increases over time. Note that it can increase even if the same number of women place children in the childcare in a given year, because the risk set became smaller.

We can also examine the hazard rate graphically. To do that, however, we need to let Stata know about our time and event variables. For that, we use `stset` command – that’s a command that is a part of the survival analysis module in Stata.

Basic syntax of `stset`:

`Stset time_of_failure_or_censoring_var, id(id_var) failure(one_if_failure_var)`

```
. stset t, id(id) failure(care==1)
```

```

          id:  id
failure event:  care == 1
obs. time interval:  (t[_n-1], t]
exit on or before:  failure

```

```

-----
1155 total obs.
   0 exclusions
-----
1155 obs. remaining, representing
   300 subjects
   209 failures in single failure-per-subject data
1155 total analysis time at risk, at risk from t =           0
          earliest observed entry t =           0
          last observed exit t =           5

```

stset defines four new variables:

\_t0 indicates when the observation for that individual starts

\_t indicates when it ends

\_d indicates the outcome at the end of the span

\_st indicates whether this observation is to be used in this analysis; we'll see later why we would need that

Note: As soon as that is set, those are the variables that will be used in calculations instead of the original time and event variables. So if you change those, you need to do stset again – it is not automatically updated.

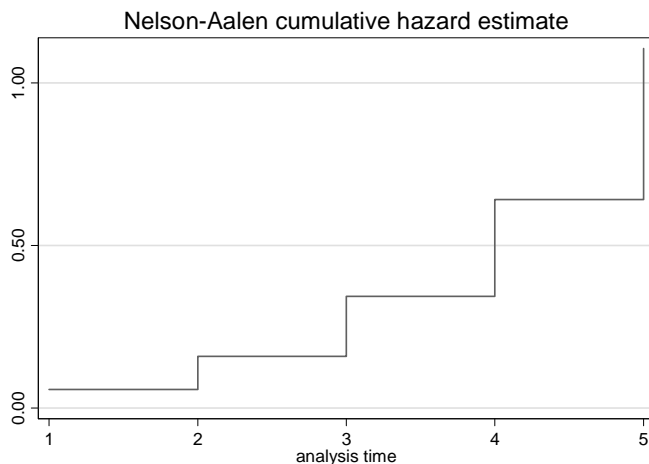
To check the dataset, we need to examine the output of stset, look at the actual data (use data browser or list command), and use stdes:

```
. stdes
      failure _d:  care == 1
      analysis time _t:  t
                   id:  id
```

Category	total	per subject			
		mean	min	median	max
no. of subjects	300				
no. of records	1155	3.85	1	4	5
(first) entry time		0	0	0	0
(final) exit time		3.85	1	4	5
subjects with gap	0				
time on gap if gap	0	.	.	.	.
time at risk	1155	3.85	1	4	5
failures	209	.6966667	0	1	1

Now we can examine hazard function graphically, using sts command with na option:

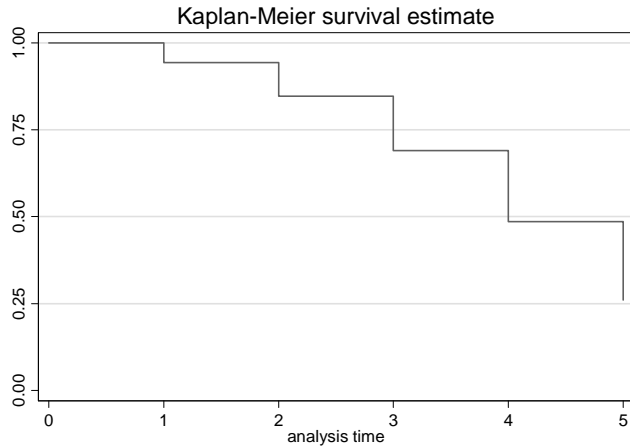
```
. sts, na
      failure _d:  care == 1
      analysis time _t:  t
                   id:  id
```



Another graphical examination tool is survivor function – it shows how the population at risk changes over time:

```
. sts
```

```
      failure _d: care == 1
analysis time _t: t
           id: id
```



Next, we want to find out how the hazard rate depends on explanatory variables. For now, we only deal with the independent variables that don't change over time.

Since  $H(t)$  is a probability, it is bound between 0 and 1. Therefore, need to transform it in order to be able to use a linear model. It is the same transformation that is used for logistic regression:  $\text{Log}(H(t)/(1-H(t)))$ .

So we can try and estimate  $\text{log}(H(t)/(1-H(t))) = a + b_1X_1 + b_1X_2 + e$  – that would be a regular logistic regression.

```
. logit care marriagel girl age
```

```
Iteration 0:  log likelihood = -546.12392
Iteration 1:  log likelihood = -535.40009
Iteration 2:  log likelihood = -534.76925
Iteration 3:  log likelihood = -534.76875
```

```
Logit estimates
```

```
Number of obs   =      1155
LR chi2(3)      =       22.71
Prob > chi2     =       0.0000
Pseudo R2      =       0.0208
```

```
Log likelihood = -534.76875
```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
marriagel	-.9371247	.2104167	-4.45	0.000	-1.349534 - .5247156
girl	.035769	.1571917	0.23	0.820	-.2723212 .3438591
age	-.0285349	.0137288	-2.08	0.038	-.0554428 -.0016269
_cons	.1544935	.4717539	0.33	0.743	-.7701272 1.079114

That's almost correct, but this assumes that all changes over time in the hazard rate that we observed are due to our covariates – other than that, the hazard is constant. We can't make that assumption before checking it first.

So we estimate instead:

$$\log(H(t)/(1-H(t))) = a(t) + b_1X_1 + b_2X_2 + e$$

```
. xi: logit care marriagel girl age i.t
i.t          _It_1-5          (naturally coded; _It_1 omitted)
Iteration 0:  log likelihood = -546.12392
Iteration 1:  log likelihood = -467.72339
Iteration 2:  log likelihood = -456.98615
Iteration 3:  log likelihood = -456.73624
Iteration 4:  log likelihood = -456.7353
Iteration 5:  log likelihood = -456.7353
Logit estimates
Log likelihood = -456.7353
Number of obs   =      1155
LR chi2(7)      =      178.78
Prob > chi2     =      0.0000
Pseudo R2      =      0.1637
```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
marriagel	-1.766389	.2480357	-7.12	0.000	-2.25253	-1.280248
girl	.0303724	.16995	0.18	0.858	-.3027235	.3634684
age	-.045422	.0150314	-3.02	0.003	-.074883	-.015961
_It_2	.8026075	.3305734	2.43	0.015	.1546956	1.450519
_It_3	1.670909	.3182897	5.25	0.000	1.047073	2.294745
_It_4	2.454864	.3216694	7.63	0.000	1.824403	3.085324
_It_5	3.219607	.3353651	9.60	0.000	2.562303	3.87691
_cons	-.256291	.539595	-0.47	0.635	-1.313878	.8012957

It appears that time has significant effects on the hazard rate above and beyond the contributions of our predictors. So we cannot assume that the hazard rate is constant. Those dummy variables represent our estimates of what's called **baseline hazard**. To see it more clearly, we could estimate a model without the constant:

```
. tab t, gen(tgroup)
```

spell year	Freq.	Percent	Cum.
1	300	25.97	25.97
2	283	24.50	50.48
3	250	21.65	72.12
4	195	16.88	89.00
5	127	11.00	100.00
Total	1,155	100.00	

```
. xi: logit care marriagel girl age tgroup1 tgroup2 tgroup3 tgroup4 tgroup5, nocons
Iteration 0:  log likelihood = -800.58499
Iteration 1:  log likelihood = -484.35629
Iteration 2:  log likelihood = -459.526
Iteration 3:  log likelihood = -457.07743
Iteration 4:  log likelihood = -457.02062
Iteration 5:  log likelihood = -457.02057
Logistic regression
Log likelihood = -457.02057
Number of obs   =      1155
LR chi2(8)      =      .
Prob > chi2     =      .
```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
marriage1	-1.762434	.2479184	-7.11	0.000	-2.248345	-1.276523
girl	.0329077	.1698892	0.19	0.846	-.300069	.3658845
age	-.0435857	.0149076	-2.92	0.003	-.072804	-.0143673
tgroup1	-.3118921	.5368338	-0.58	0.561	-1.364067	.7402827
tgroup2	.488564	.5379791	0.91	0.364	-.5658557	1.542984
tgroup3	1.355298	.5511774	2.46	0.014	.2750104	2.435586
tgroup4	2.139123	.5635852	3.80	0.000	1.034516	3.243729
tgroup5	2.902041	.5754741	5.04	0.000	1.774132	4.029949

We could also graph the baseline hazard using adjust command. Adjust command generates predicted values (here, predicted probabilities) holding certain variables constant and allowing other variables to vary:

```
. adjust marriage1 girl age, gen(prob) pr
```

```
-----
Dependent variable: care      Command: logit
Created variable: prob
Variables left as is: _It_2, _It_3, _It_4, _It_5
Covariates set to mean: marriage1 = .89264069, girl = .41038961, age = 30.784416
-----
```

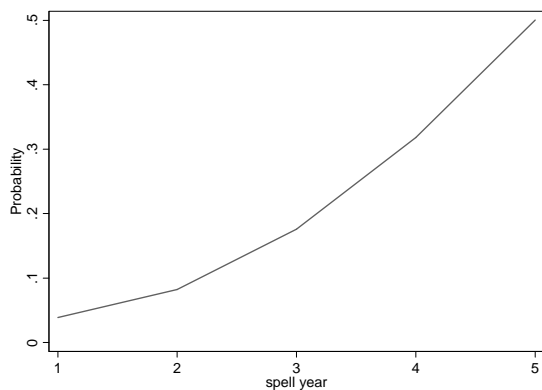
All	pr
	.131339

Key: pr = Probability

```
. tab prob t
```

Probability	spell year					Total
y	1	2	3	4	5	
.0386658	300	0	0	0	0	300
.0821935	0	283	0	0	0	283
.1756393	0	0	250	0	0	250
.318136	0	0	0	195	0	195
.5001417	0	0	0	0	127	127
Total	300	283	250	195	127	1,155

```
. line prob t, sort
```



Note that here, we don't make any assumptions as to how the baseline hazard rate changes over time. When we get to continuous-data models, you'll see that this is also the strategy that so-called semiparametric models (like Cox model) use.

The last adjustment that we might want to do to our logit model is to adjust for the fact that in our dataset, multiple observations belong to the same person. Typically, the results don't change much, though:

```
. xi: logit care marriagel girl age i.t, cluster(id)
i.t          _It_1-5          (naturally coded; _It_1 omitted)
Iteration 0:  log pseudolikelihood = -546.12392
Iteration 1:  log pseudolikelihood = -467.96827
Iteration 2:  log pseudolikelihood = -457.26844
Iteration 3:  log pseudolikelihood = -457.0215
Iteration 4:  log pseudolikelihood = -457.02057
Iteration 5:  log pseudolikelihood = -457.02057
Logistic regression
Number of obs   =      1155
Wald chi2(7)    =      135.05
Prob > chi2     =      0.0000
Pseudo R2       =      0.1632
Log pseudolikelihood = -457.02057
                    (Std. Err. adjusted for 300 clusters in id)
```

care	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
marriagel	-1.762434	.3119892	-5.65	0.000	-2.373921	-1.150946
girl	.0329077	.1778872	0.18	0.853	-.3157448	.3815602
age	-.0435857	.0153134	-2.85	0.004	-.0735994	-.0135719
_It_2	.8004561	.3195566	2.50	0.012	.1741367	1.426776
_It_3	1.66719	.3038745	5.49	0.000	1.071607	2.262773
_It_4	2.451015	.3147851	7.79	0.000	1.834047	3.067982
_It_5	3.213933	.3389521	9.48	0.000	2.549599	3.878267
_cons	-.3118921	.5675233	-0.55	0.583	-1.424217	.800433

## Interpreting findings

How do we interpret the findings of our model? The same way we would with regular logistic regression. For example, we can obtain odds ratios to be able to interpret effect sizes:

```
. xi: logit care marriagel girl age i.t, cluster(id) or
i.t          _It_1-5          (naturally coded; _It_1 omitted)
Iteration 0:  log pseudolikelihood = -546.12392
Iteration 1:  log pseudolikelihood = -467.96827
Iteration 2:  log pseudolikelihood = -457.26844
Iteration 3:  log pseudolikelihood = -457.0215
Iteration 4:  log pseudolikelihood = -457.02057
Iteration 5:  log pseudolikelihood = -457.02057
Logistic regression
Number of obs   =      1155
Wald chi2(7)    =      135.05
Prob > chi2     =      0.0000
Pseudo R2       =      0.1632
Log pseudolikelihood = -457.02057
                    (Std. Err. adjusted for 300 clusters in id)
```

care	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
marriagel	.1716266	.0535457	-5.65	0.000	.0931149	.3163373
girl	1.033455	.1838384	0.18	0.853	.7292455	1.464568
age	.9573506	.0146603	-2.85	0.004	.9290438	.9865197
_It_2	2.226556	.7115107	2.50	0.012	1.190218	4.165247
_It_3	5.297263	1.609703	5.49	0.000	2.920069	9.609703
_It_4	11.60011	3.651543	7.79	0.000	6.259168	21.49848
_It_5	24.87673	8.43202	9.48	0.000	12.80197	48.34035

We can also use any other interpretation tools that can be used with logistic regression. For example, we can install the spostado package that contains some helpful programs for logistic regression:

```
. net search spostado
```

Click on:

spost9\_ado from <http://www.indiana.edu/~jslsoc/stata>

and install.

Using a command available in this package, we can also obtain odds ratios as well as standardized odds ratios:

```
. listcoef
logit (N=1155): Factor Change in Odds
Odds of: 1 vs 0
```

care	b	z	P> z	e^b	e^bStdX	SDofX
marriage1	-1.76243	-5.649	0.000	0.1716	0.5794	0.3097
girl	0.03291	0.185	0.853	1.0335	1.0163	0.4921
age	-0.04359	-2.846	0.004	0.9574	0.7773	5.7809
_It_2	0.80046	2.505	0.012	2.2266	1.4112	0.4303
_It_3	1.66719	5.486	0.000	5.2973	1.9875	0.4120
_It_4	2.45101	7.786	0.000	11.6001	2.5057	0.3748
_It_5	3.21393	9.482	0.000	24.8767	2.7343	0.3130

And we can express those odds ratios as percent change:

```
. listcoef, percent
logit (N=1155): Percentage Change in Odds
Odds of: 1 vs 0
```

care	b	z	P> z	%	%StdX	SDofX
marriage1	-1.76243	-5.649	0.000	-82.8	-42.1	0.3097
girl	0.03291	0.185	0.853	3.3	1.6	0.4921
age	-0.04359	-2.846	0.004	-4.3	-22.3	5.7809
_It_2	0.80046	2.505	0.012	122.7	41.1	0.4303
_It_3	1.66719	5.486	0.000	429.7	98.8	0.4120
_It_4	2.45101	7.786	0.000	1060.0	150.6	0.3748
_It_5	3.21393	9.482	0.000	2387.7	173.4	0.3130

We could examine predicted hazard value at different levels of independent variables:

```
. prvalue, x(age=25 girl=1 marriage1=1)
logit: Predictions for care
Confidence intervals by delta method
```

		95% Conf. Interval	
Pr(y=1 x):	0.1410	[ 0.1000,	0.1820]
Pr(y=0 x):	0.8590	[ 0.8180,	0.9000]

```

marriage1    girl    age    _It_2    _It_3    _It_4    _It_5
x=           1      1      25    .24502165    .21645022    .16883117    .10995671
. prtab marriage1 girl
```

logit: Predicted probabilities of positive outcome for care

```
-----  
marriage1 |      girl  
           |      0      1  
-----+-----  
           |  
0 | 0.4184  0.4264  
1 | 0.1099  0.1131  
-----
```

```
marriage1      girl      age      _It_2      _It_3      _It_4      _It_5  
x= .89264069 .41038961 30.784416 .24502165 .21645022 .16883117 .10995671
```

```
. prtab marriage1 girl, x( _It_2=0 _It_3=0 _It_4=0 _It_5=0)
```

logit: Predicted probabilities of positive outcome for care

```
-----  
marriage1 |      girl  
           |      0      1  
-----+-----  
           |  
0 | 0.1606  0.1651  
1 | 0.0318  0.0328  
-----
```

```
marriage1      girl      age      _It_2      _It_3      _It_4      _It_5  
x= .89264069 .41038961 30.784416          0          0          0          0
```

```
. prtab marriage1 girl, x( _It_2=1 _It_3=0 _It_4=0 _It_5=0)
```

logit: Predicted probabilities of positive outcome for care

```
-----  
marriage1 |      girl  
           |      0      1  
-----+-----  
           |  
0 | 0.2988  0.3057  
1 | 0.0681  0.0703  
-----
```

```
marriage1      girl      age      _It_2      _It_3      _It_4      _It_5  
x= .89264069 .41038961 30.784416          1          0          0          0
```

```
. prtab marriage1 girl, x( _It_2=0 _It_3=1 _It_4=0 _It_5=0)
```

logit: Predicted probabilities of positive outcome for care

```
-----  
marriage1 |      girl  
           |      0      1  
-----+-----  
           |  
0 | 0.5034  0.5116  
1 | 0.1482  0.1524  
-----
```

```
marriage1      girl      age      _It_2      _It_3      _It_4      _It_5  
x= .89264069 .41038961 30.784416          0          1          0          0
```

```
. prtab marriage1 girl, x( _It_2=0 _It_3=0 _It_4=1 _It_5=0)
```

logit: Predicted probabilities of positive outcome for care

```

-----
marriage1 |      girl
           |      0      1
-----+-----
           | 0 | 0.6894  0.6964
           | 1 | 0.2759  0.2825
-----

      marriage1      girl      age      _It_2      _It_3      _It_4      _It_5
x=    .89264069    .41038961  30.784416          0          0          1          0

```

```

. prtab marriage1 girl, x( _It_2=0 _It_3=0 _It_4=0 _It_5=1)

logit: Predicted probabilities of positive outcome for care

```

```

-----
marriage1 |      girl
           |      0      1
-----+-----
           | 0 | 0.8264  0.8311
           | 1 | 0.4496  0.4578
-----

      marriage1      girl      age      _It_2      _It_3      _It_4      _It_5
x=    .89264069    .41038961  30.784416          0          0          0          1

```

We could also see how the predicted hazard changes with the change in predictors:

```

. prchange

logit: Changes in Probabilities for care

      marriage1      min->max      0->1      -+1/2      -+sd/2      MargEfct
marriage1      -0.3104      -0.3104      -0.2084      -0.0625      -0.2011
girl            0.0038      0.0038      0.0038      0.0018      0.0038
age            -0.1523      -0.0101      -0.0050      -0.0288      -0.0050
_It_2          0.1062      0.1062      0.0921      0.0394      0.0913
_It_3          0.2629      0.2629      0.1965      0.0788      0.1902
_It_4          0.4461      0.4461      0.2974      0.1059      0.2796
_It_5          0.6294      0.6294      0.4005      0.1162      0.3667

```

```

           0      1
Pr(y|x)  0.8687  0.1313

      marriage1      girl      age      _It_2      _It_3      _It_4      _It_5
x=    .892641      .41039      30.7844      .245022      .21645      .168831      .109957
sd(x)= .309704      .492118      5.7809      .430286      .412003      .374765      .312971

```

Finally, we can also interpret our findings graphically by comparing predicted hazard functions. For example, we can use adjust command to see hazard functions for those married and those single:

```

. adjust girl age, gen(prob2) pr

-----
Dependent variable: care      Command: logit
Created variable: prob2
Variables left as is: marriage1, _It_2, _It_3, _It_4, _It_5
Covariates set to mean: girl = .41038961, age = 30.784416
-----

All |      pr

```

```

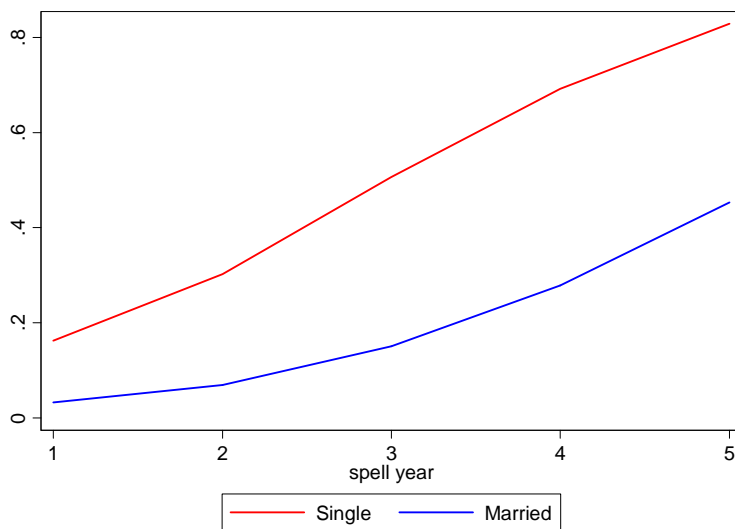
-----+-----
      | .131339
-----+-----
      Key: pr = Probability
      . separate prob2, by(marriage1)
      variable name      storage  display      value
                        type      format      label      variable label
-----+-----
prob20                  float    %9.0g
prob21                  float    %9.0g
                        prob2, marriage1 == 0
                        prob2, marriage1 == 1

      . lab var prob20 "Single"

      . lab var prob21 "Married"

      . line prob20 prob21 t, sort lcolor(red blue)

```



To learn more about logistic regression interpretation tools in Stata, please see my SC704 notes at <http://www.sarkisian.net/sc704/notes.html>.

## Modeling time

So far, we allowed a separate coefficient for each time point. We could instead make some assumption about how hazard changes over time. We could decide that the hazard rate is constant – that’s what we did above, before including time into the model. We could decide that it increases in a linear fashion and include time as a single independent variable:

```

      . logit care married girl age t

Iteration 0:  log likelihood = -546.12392
Iteration 1:  log likelihood = -464.68047
Iteration 2:  log likelihood = -456.92323
Iteration 3:  log likelihood = -456.78476
Iteration 4:  log likelihood = -456.78464

Logit estimates
Number of obs   =      1155
LR chi2(4)      =      178.68

```



```

. xi: logit care marriagel girl age t1 t2

Iteration 0:  log likelihood = -546.12392
Iteration 1:  log likelihood = -467.9148
Iteration 2:  log likelihood = -457.29096
Iteration 3:  log likelihood = -457.0387
Iteration 4:  log likelihood = -457.03794
Iteration 5:  log likelihood = -457.03794
Logistic regression
Log likelihood = -457.03794
Number of obs   =      1155
LR chi2(5)      =      178.17
Prob > chi2     =      0.0000
Pseudo R2       =      0.1631

```

	care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
marriagel		-1.761036	.2475446	-7.11	0.000	-2.246215 -1.275858
girl		.0327509	.1698877	0.19	0.847	-.300223 .3657247
age		-.0435775	.0149082	-2.92	0.003	-.072797 -.0143581
t1		.8090363	.0739495	10.94	0.000	.664098 .9539746
t2		-.0133846	.0533435	-0.25	0.802	-.1179359 .0911666
_cons		1.339906	.5376753	2.49	0.013	.2860818 2.39373

Clearly, there is no quadratic relationship here. To approximately assess the shape of the relationship between our hazard rate and time, we can graphically evaluate the hazard rate function on the graph that we estimated with `adjust` command (see above). In addition to various transformations of time, we could do piecewise estimation – break time into intervals – e.g. group times 1-3 and times 4-5. We’ll examine that later on.

It is especially important to explore the shape of the relationship with time when:

- there are many discrete periods – too many dummy variables
- the hazard is expected to be near zero in some time periods. When few or no events occur in a specific period, a model with dummy variables for time may fail to converge or may present unrealistic estimates.
- some time periods have very small risk sets

Note: we don’t actually think that time has an effect on its own – rather, time here is used as a proxy for all sorts of unmeasured covariates. In a perfect world, if we had all relevant information that affects mothers’ decisions to place a child to the child care, we wouldn’t need to have time in the models (i.e. we could explain all the reasons as to why the hazard is non-constant over time).

### Time-varying predictors

Next, we’ll try using time-variant predictors. For that, we will return to the marriage variable and combine it with the dataset we are currently using.

```

. use "M:\childcare.dta", clear

. drop childcare girl age

. gen id = _n

. reshape long marriage, i(id)
(note: j = 1 2 3 4 5)

```

```
Data                                wide  ->  long
-----
Number of obs.                      300  ->  1500
Number of variables                   7    ->    4
j variable (5 values)                ->  _j
xij variables:
    marriagel marriage2 ... marriage5 ->  marriage
-----
```

```
. by id: drop if _n>time
(345 observations deleted)
```

```
. rename _j t
. sort id t
. save "M:\marriage.dta"
```

Going back to the file that we've been working with, let's merge them together:

```
. sort id t
. merge id t using "M:\marriage.dta"
. tab _merge
```

_merge	Freq.	Percent	Cum.
3	1,155	100.00	100.00
Total	1,155	100.00	

```
. drop _merge
```

Note: `_merge` indicates from where the cases in the merged file came. The value of 3 means that the case came from both datasets; the value of 1 means it only existed in the master dataset; the value of 2 means it only existed in the dataset we added. Here, as expected, all cases have the value of 3.

Now we can run a model with marriage as a time variant predictor:

```
. xi: logit care marriage girl age i.t
i.t                _It_1-5                (naturally coded; _It_1 omitted)

Iteration 0:  log likelihood = -546.12392
Iteration 1:  log likelihood = -462.10619
Iteration 2:  log likelihood = -450.6915
Iteration 3:  log likelihood = -450.41222
Iteration 4:  log likelihood = -450.41105
Iteration 5:  log likelihood = -450.41105

Logistic regression                                Number of obs =          1155
                                                    LR chi2(7)      =          191.43
                                                    Prob > chi2     =           0.0000
Log likelihood = -450.41105                        Pseudo R2      =           0.1753
-----
```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
marriage	-1.917234	.242222	-7.92	0.000	-2.39198	-1.442488
girl	.0206707	.1713888	0.12	0.904	-.3152451	.3565865
age	-.0379315	.014949	-2.54	0.011	-.0672309	-.008632
_It_2	.8181404	.3324918	2.46	0.014	.1664686	1.469812
_It_3	1.679057	.3200576	5.25	0.000	1.051756	2.306358
_It_4	2.473043	.3234156	7.65	0.000	1.83916	3.106926
_It_5	3.264627	.3373709	9.68	0.000	2.603392	3.925861
_cons	-.3885422	.531908	-0.73	0.465	-1.431063	.6539782

The findings are similar, although the effect of time appears stronger. The interpretation is somewhat different here: For a dichotomous time-varying predictor like marriage, the change in risk only kicks in during those time periods when the event denoted by that predictor happens and the subsequent time periods. That's different from how we interpret time-invariant predictors' effects.

There are actually a few different ways how we can use time-variant predictors:

- Lasting effect of the new status – that's what we did here, the new status is assumed to have an effect when the event happens and in the subsequent time periods
- Contemporaneous effect of transition – the new status is only assumed to have an effect during the period when the transition happens
- Delayed effect (lagged time-variant predictor) – the effect is assumed to take place in the period that follows the one when the transition happened (if the lag=1; it is possible to envision longer lags as well)
- Anticipatory effect – the effect is assumed to start before the transition happened, i.e. in anticipation of that transition.

All of these assumptions will affect how we code the time-variant predictor.

Time	Marital status	Lasting effect	Contemporaneous effect	Delayed effect	Anticipatory effect
1	0	0	0	0	0
2	0	0	0	0	1
3	1	1	1	0	1
4	1	1	0	1	1
5	1	1	0	1	1

Note that these types of coding are also possible if time-variant predictor is continuous.

### Time-varying effects of predictors

So far, we assumed that our independent variables have the same effects over time – that is, that there are no interactions between our independent variables and some unmeasured covariates. Note that this assumption of no interaction between our independent variables and time is called the assumption of proportionality of hazards. That is, the baseline hazard shape is assumed to be the same for all values of predictors. But that isn't necessarily the case, as we can see below. Note that varying gaps between hazard estimates in plots do not mean that there is an interaction!

To check whether there is some variation in effects of predictors over time, we could include interactions of the variables we have in the model with time:

```
. xi: logit care marriage girl age i.t*marriage
i.t          _It_1-5          (naturally coded; _It_1 omitted)
i.t*marriage _ItXmarri_#      (coded as above)
```

note: marriage dropped due to collinearity

```
Iteration 0: log likelihood = -546.12392
Iteration 1: log likelihood = -460.92504
Iteration 2: log likelihood = -446.23802
Iteration 3: log likelihood = -445.49342
Iteration 4: log likelihood = -445.47544
Iteration 5: log likelihood = -445.47542
```

```
Logistic regression          Number of obs   =      1155
                             LR chi2(11)       =      201.30
                             Prob > chi2       =      0.0000
Log likelihood = -445.47542   Pseudo R2        =      0.1843
```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
marriage	-2.420353	.5384762	-4.49	0.000	-3.475747 -1.364959
girl	.0412374	.1716412	0.24	0.810	-.2951731 .377648
age	-.0379674	.0149161	-2.55	0.011	-.0672025 -.0087323
_It_2	.7740498	.4837381	1.60	0.110	-.1740594 1.722159
_It_3	1.6582	.5297884	3.13	0.002	.6198339 2.696566
_It_4	1.346123	.6774849	1.99	0.047	.0182768 2.673969
_It_5	.8290567	.9847911	0.84	0.400	-1.101098 2.759212
_ItXmarri_2	.2100144	.6899006	0.30	0.761	-1.142166 1.562195
_ItXmarri_3	.2459365	.6987408	0.35	0.725	-1.12357 1.615443
_ItXmarri_4	1.469699	.8097527	1.81	0.070	-.1173866 3.056786
_ItXmarri_5	2.785965	1.083826	2.57	0.010	.6617042 4.910225
_cons	-.1669092	.5547287	-0.30	0.764	-1.254157 .9203391

```
. xi: logit care marriage girl age i.t*girl
i.t          _It_1-5          (naturally coded; _It_1 omitted)
i.t*girl     _ItXgirl_#      (coded as above)
```

note: girl dropped due to collinearity

```
Iteration 0: log likelihood = -546.12392
Iteration 1: log likelihood = -461.09444
Iteration 2: log likelihood = -449.29777
Iteration 3: log likelihood = -448.98333
Iteration 4: log likelihood = -448.98144
Iteration 5: log likelihood = -448.98144
```

```
Logistic regression          Number of obs   =      1155
                             LR chi2(11)       =      194.28
                             Prob > chi2       =      0.0000
Log likelihood = -448.98144   Pseudo R2        =      0.1779
```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
marriage	-1.930473	.2426513	-7.96	0.000	-2.40606 -1.454885
girl	-.5248036	.5407896	-0.97	0.332	-1.584732 .5351245
age	-.0391333	.0149969	-2.61	0.009	-.0685267 -.0097398
_It_2	.5552088	.4297523	1.29	0.196	-.2870903 1.397508
_It_3	1.400652	.4014228	3.49	0.000	.6138777 2.187426
_It_4	2.136317	.3991044	5.35	0.000	1.354087 2.918547
_It_5	3.180871	.4108078	7.74	0.000	2.375703 3.98604
_ItXgirl_2	.6387143	.6790202	0.94	0.347	-.6921408 1.969569

```

    _ItXgirl_3 | .6875975 .6418705 1.07 0.284 -.5704456 1.945641
    _ItXgirl_4 | .8302397 .6319372 1.31 0.189 -.4083345 2.068814
    _ItXgirl_5 | .2053064 .6563081 0.31 0.754 -1.081034 1.491647
    _cons      | -.1238275 .5692771 -0.22 0.828 -1.23959 .991935
-----

```

```

. xi: logit care marriage girl age i.t*age
i.t      _It_1-5      (naturally coded; _It_1 omitted)
i.t*age  _ItXage_#    (coded as above)

```

note: age dropped due to collinearity

```

Iteration 0: log likelihood = -546.12392
Iteration 1: log likelihood = -457.42967
Iteration 2: log likelihood = -441.73561
Iteration 3: log likelihood = -439.9021
Iteration 4: log likelihood = -439.72262
Iteration 5: log likelihood = -439.71966
Iteration 6: log likelihood = -439.71966

```

```

Logistic regression              Number of obs   =      1155
                                LR chi2(11)        =      212.81
                                Prob > chi2         =      0.0000
Log likelihood = -439.71966     Pseudo R2      =      0.1948
-----

```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
marriage	-1.909204	.2486807	-7.68	0.000	-2.396609 -1.421799
girl	-.0131134	.1730587	-0.08	0.940	-.3523022 .3260753
age	-.2160995	.0648173	-3.33	0.001	-.3431391 -.0890599
_It_2	-1.705507	2.039787	-0.84	0.403	-5.703415 2.292402
_It_3	-3.557638	1.922417	-1.85	0.064	-7.325507 .2102306
_It_4	-4.561333	1.901381	-2.40	0.016	-8.287971 -.8346955
_It_5	-2.056172	1.959712	-1.05	0.294	-5.897137 1.784793
_ItXage_2	.0998928	.0763826	1.31	0.191	-.0498145 .2496
_ItXage_3	.1923058	.071474	2.69	0.007	.0522193 .3323923
_ItXage_4	.2499162	.0705957	3.54	0.000	.1115513 .3882812
_ItXage_5	.1948238	.0723811	2.69	0.007	.0529595 .3366881
_cons	4.423321	1.678297	2.64	0.008	1.133919 7.712723

If we would want to use time as a continuous variable in an interaction term (i.e. if we would observe that effect of age increases or decreases with time in a linear fashion), we would have to generate interaction terms before entering them in the model – and don't forget to mean-center the variables if a continuous predictor (like age) is involved in that interaction term:

```

. sum age
Variable |      Obs      Mean   Std. Dev.   Min     Max
-----+-----
age      |    1155    30.78442   5.780901     18     52

```

```
. gen agem=age-r(mean)
```

```
. gen agemt1=agem*t1
```

```
. xi: logit care marriage1 girl agem t1 agemt1, cluster(id)
```

```

Iteration 0: log pseudolikelihood = -546.12392
Iteration 1: log pseudolikelihood = -462.33246
Iteration 2: log pseudolikelihood = -452.1663
Iteration 3: log pseudolikelihood = -451.69758
Iteration 4: log pseudolikelihood = -451.69522
Iteration 5: log pseudolikelihood = -451.69522

```

```

Logistic regression              Number of obs   =      1155
                                Wald chi2(5)     =      127.20

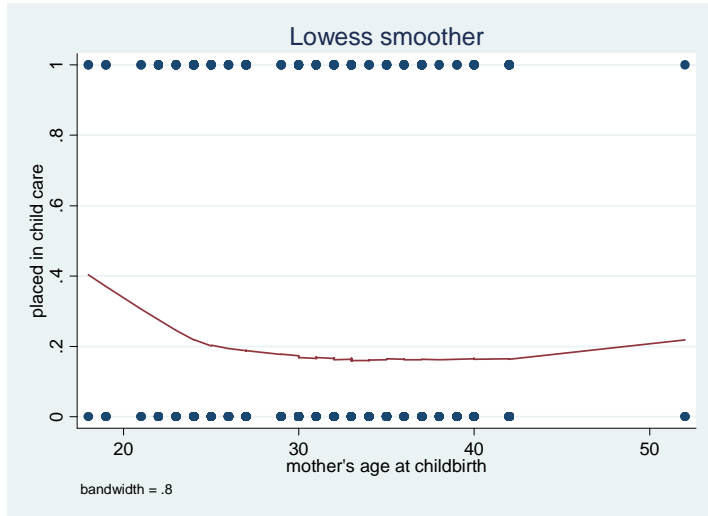
```



Linearity:

First, let's examine linearity in bivariate context:

.lowess care age



Now a test in a multivariate model – Box-Tidwell test:

```
. xi: boxtid logit care marriagel girl age t, cluster(id)
Iteration 0: Deviance = 910.9368
Iteration 1: Deviance = 907.1644 (change = -3.772437)
Iteration 2: Deviance = 906.9721 (change = -.192309)
Iteration 3: Deviance = 906.9701 (change = -.0019165)
Iteration 4: Deviance = 906.9701 (change = -7.43e-06)
-> gen double Iage__1 = X^-5.1296-.0031264986 if e(sample)
-> gen double Iage__2 = X^-5.1296*ln(X)-.0035155084 if e(sample)
      (where: X = age/10)
-> gen double It__1 = t^0.8632-2.299846638 if e(sample)
-> gen double It__2 = t^0.8632*ln(t)-2.218874222 if e(sample)
[Total iterations: 8]
Box-Tidwell regression model
Logistic regression
Number of obs = 1155
Wald chi2(6) = 137.88
Prob > chi2 = 0.0000
Pseudo R2 = 0.1696
Log pseudolikelihood = -453.48507
(Std. Err. adjusted for 300 clusters in id)
```

care	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
Iage__1	52.9969	66.45506	0.80	0.425	-77.25263 183.2464
Iage_p1	-.0190863	99.32435	-0.00	1.000	-194.6912 194.6531
It__1	1.10129	.8434839	1.31	0.192	-.5519079 2.754488
It_p1	-1.38e-06	.3794896	-0.00	1.000	-.7437874 .7437846
marriagel	-1.738695	.3095441	-5.62	0.000	-2.345391 -1.132
girl	.0045901	.1785553	0.03	0.979	-.3453719 .3545521
_cons	-.4377707	.2908542	-1.51	0.132	-1.007834 .1322929
-----					
age	-.0436651	.0153872	-2.838	Nonlin. dev. 7.140	(P = 0.008)
p1	-5.129599	1.865607	-2.750		
-----					
t	.8200891	.0749768	10.938	Nonlin. dev. 0.149	(P = 0.700)
p1	.8632353	.344608	2.505		
-----					

Deviance: 906.970.

Clearly, the relationship between age and the hazard of child care is not linear. Let's try quadratic:

```
. gen agem2=agem^2

. xi: boxtid logit care marriagel girl agem agem2 t, cluster(id)

Iteration 0: Deviance = 906.6065
Iteration 1: Deviance = 906.2935 (change = -.3130011)
Iteration 2: Deviance = 906.2845 (change = -.0089953)
Iteration 3: Deviance = 906.2829 (change = -.0016762)
Iteration 4: Deviance = 906.2828 (change = -.0000117)
-> gen double Iagem__1 = X^-0.3450-.8951951401 if e(sample)
-> gen double Iagem__2 = X^-0.3450*ln(X)-.2873160111 if e(sample)
    (where: X = (agem+13.78441476821899)/10)
-> gen double Iagama_1 = X^1.5411-.1844356707 if e(sample)
-> gen double Iagama_2 = X^1.5411*ln(X)+.2023106398 if e(sample)
    (where: X = agem2/100)
-> gen double It__1 = t^0.8711-2.317295155 if e(sample)
-> gen double It__2 = t^0.8711*ln(t)-2.235708416 if e(sample)
```

[Total iterations: 12]

Box-Tidwell regression model

```
Logistic regression                               Number of obs   =       1155
                                                    Wald chi2(8)    =       145.33
                                                    Prob > chi2     =       0.0000
Log pseudolikelihood = -453.14142                Pseudo R2      =       0.1703
```

(Std. Err. adjusted for 300 clusters in id)

care	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
Iagem__1	1.692929	2.388666	0.71	0.478	-2.988769 6.374628
Iagem__p1	.0001906	.6930259	0.00	1.000	-1.358115 1.358496
Iagama_1	.1424417	.2884084	0.49	0.621	-.4228285 .7077118
Iagemap1	-4.88e-06	.1842325	-0.00	1.000	-.361094 .3610843
It__1	1.084449	.824601	1.32	0.188	-.5317396 2.700637
It__p1	2.65e-06	.3728073	0.00	1.000	-.7306862 .7306915
marriagel	-1.755784	.3099625	-5.66	0.000	-2.3633 -1.148269
girl	-.0089506	.179582	-0.05	0.960	-.3609249 .3430236
_cons	-.3946894	.2863801	-1.38	0.168	-.9559841 .1666054
-----					
agem	-.0524316	.015206	-3.448	Nonlin. dev. 2.574	(P = 0.109)
p1	-.3449517	.4148221	-0.832		
-----					
agem2	.0022278	.0016275	1.369	Nonlin. dev. 0.279	(P = 0.597)
p1	1.541096	1.290353	1.194		
-----					
t	.8217855	.0750748	10.946	Nonlin. dev. 0.132	(P = 0.716)
p1	.8710691	.3437421	2.534		

Deviance: 906.283.

```
. xi: logit care marriagel girl agem agem2 t, cluster(id)
```

```
Iteration 0: log pseudolikelihood = -546.12392
Iteration 1: log pseudolikelihood = -462.50683
Iteration 2: log pseudolikelihood = -454.34034
Iteration 3: log pseudolikelihood = -454.18326
Iteration 4: log pseudolikelihood = -454.1831
```

```
Logistic regression                               Number of obs   =       1155
                                                    Wald chi2(5)    =       140.88
                                                    Prob > chi2     =       0.0000
Log pseudolikelihood = -454.1831                Pseudo R2      =       0.1684
```

(Std. Err. adjusted for 300 clusters in id)

-----

care	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
marriagel	-1.75122	.3024024	-5.79	0.000	-2.343918	-1.158523
girl	-.0015946	.1794128	-0.01	0.993	-.3532372	.350048
agem	-.0502396	.0146979	-3.42	0.001	-.0790469	-.0214323
agem2	.0048779	.001604	3.04	0.002	.001734	.0080217
t	.8164373	.0743302	10.98	0.000	.6707529	.9621218
_cons	-2.640422	.2855574	-9.25	0.000	-3.200104	-2.08074

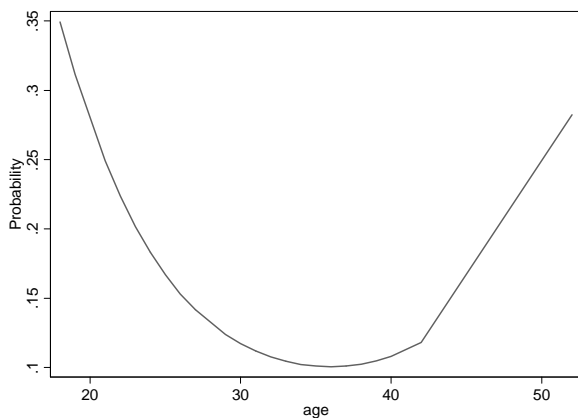
```
. adjust marriagel girl t, gen(pred1) pr
```

```
-----
Dependent variable: care      Command: logit
Created variable: pred1
Variables left as is: agem, agem2
Covariates set to mean: marriagel = .89264069, girl = .41038961, t = 2.6242424
-----
```

All	pr
	.130235

Key: pr = Probability

```
. line pred1 age, sort
```



### Additivity:

```
. xi: fitint logit care marriagel girl age t, factor(t) twoway(marriagel girl age t)
i.t          _It_1-5          (naturally coded; _It_1 omitted)
i.t*marriagel  _ItXmarri_#    (coded as above)
i.t*girl      _ItXgirl_#      (coded as above)
i.t*age       _ItXage_#       (coded as above)
```

```
Logistic regression          Number of obs   =      1155
                             LR chi2(22)         =      223.20
                             Prob > chi2         =      0.0000
Log likelihood = -434.52407   Pseudo R2      =      0.2043
```

care	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
marriagel	-4.750847	1.537215	-3.09	0.002	-7.763733	-1.737961
age	-.2791842	.0746135	-3.74	0.000	-.4254239	-.1329444
_It_3	-2.577874	2.323963	-1.11	0.267	-7.132758	1.977009
_It_4	-4.612167	2.517146	-1.83	0.067	-9.545683	.3213489
_5_5	.0970526	.5411913	0.18	0.858	-.963663	1.157768
_6_5	.08691	.0539182	1.61	0.107	-.0187676	.1925877

_ItXmarri_2	.08097	.7818918	0.10	0.918	-1.45151	1.61345
_ItXmarri_3	-.1139739	.8528188	-0.13	0.894	-1.785468	1.55752
_ItXmarri_4	1.158266	1.053697	1.10	0.272	-.9069428	3.223475
_ItXmarri_5	20.22958	2.278662	8.88	0.000	15.76349	24.69568
_8_5	.034147	.0313593	1.09	0.276	-.0273162	.0956101
_It_5	-21.35083	.	.	.	.	.
girl	-1.859943	1.049371	-1.77	0.076	-3.916672	.1967872
_ItXgirl_2	.8379581	.7814496	1.07	0.284	-.693655	2.369571
_ItXgirl_3	.8826185	.7741906	1.14	0.254	-.6347673	2.400004
_ItXgirl_4	.9903349	.7875135	1.26	0.209	-.5531631	2.533833
_ItXgirl_5	.6290018	.8147433	0.77	0.440	-.9678657	2.225869
_It_2	-1.784019	2.283833	-0.78	0.435	-6.260249	2.692211
_ItXage_2	.0918636	.0809994	1.13	0.257	-.0668923	.2506195
_ItXage_3	.155709	.0781309	1.99	0.046	.0025753	.3088428
_ItXage_4	.207087	.0793431	2.61	0.009	.0515773	.3625967
_ItXage_5	.1631896	.0811237	2.01	0.044	.0041901	.3221891
_cons	6.725603	1.989165	3.38	0.001	2.826911	10.62429

Note: 4 failures and 0 successes completely determined.

Fitting and testing any interactions and any main effects not included in interaction terms using the change in deviance from the full model when each term is removed in turn to obtain the likelihood ratio chi square statistic

#### Model summary

Number of observations used in estimation: 1155  
 Regression command: logit  
 Dependent variable: care  
 Full model deviance: 869.05  
 degrees of freedom: 22

Term	Model deviance	Chi-square	df	P>Chi
marriage1*girl	869.08	0.03	1	0.8578
marriage1*age	871.74	2.69	1	0.1011
i.t*marriage1	881.52	12.47	3	0.0059
girl*age	870.23	1.19	1	0.2761
i.t*girl	871.12	2.07	4	0.7232
i.t*age	879.04	10.00	4	0.0405

We confirmed once again that effects of marriage and age vary over time. There are no other significant interactions. In general, we need to exercise caution when deciding what interactions to include in a logit model, as significance tests can be somewhat misleading. To learn more about interactions in logit models, you can consult

- [http://www.stata.com/support/faqs/stat/mfx\\_interact.html](http://www.stata.com/support/faqs/stat/mfx_interact.html)
- <http://www.unc.edu/~enorton/NortonWangAi.pdf>
- [http://www.ats.ucla.edu/stat/stata/seminars/stata\\_vibl/](http://www.ats.ucla.edu/stat/stata/seminars/stata_vibl/)
- Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients Across Groups." *Sociological Methods and Research*, 28: 186-208.
- Hoetker, Glenn. 2004. "Confounded Coefficients: Extending Recent Advances in the Accurate Comparison of Logit and Probit Coefficients Across Groups."
- [http://www.business.uiuc.edu/Working\\_Papers/papers/03-0100.pdf](http://www.business.uiuc.edu/Working_Papers/papers/03-0100.pdf)
- Long, Scott. 2006. Comparing Group Effects in Logit and Probit Models.
- <http://www.umass.edu/family/conference/Long.htm>

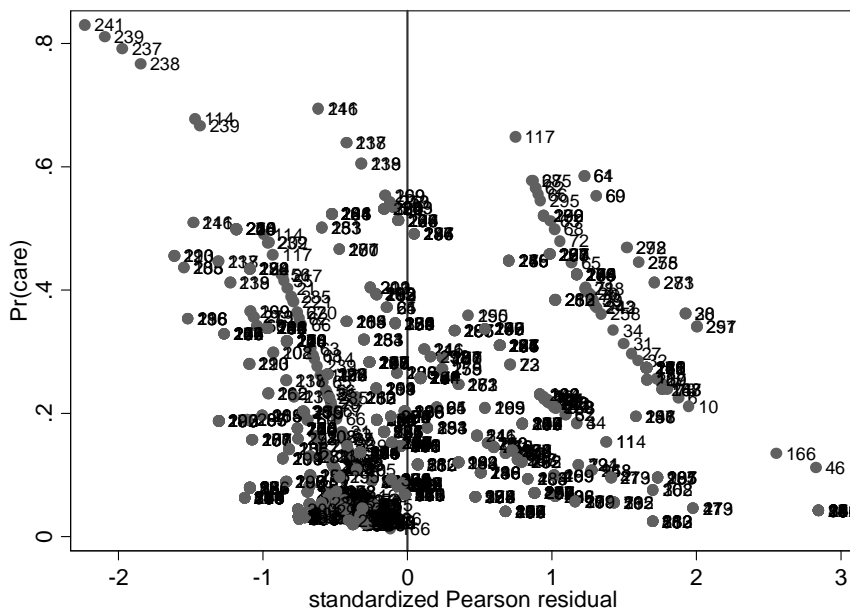
### Outliers and Influential Observations

To detect influential observations and outliers, there are a few statistics you can obtain using predict command after logit:

```
p          predicted probability of a positive outcome; the default
xb        linear prediction
stdp      standard error of the linear prediction
dbeta     Pregibon (1981) Delta-Beta influence statistic
deviance  deviance residual
dx2       Hosmer and Lemeshow (2000) Delta chi-squared infl. stat.
ddeviance Hosmer and Lemeshow (2000) Delta-D influence statistic
hat       Pregibon (1981) leverage
number    sequential number of the covariate pattern
residuals Pearson residual (adj. for # sharing covariate pattern)
rstandard standardized Pearson residual (adj. for # sharing covariate
pattern)
```

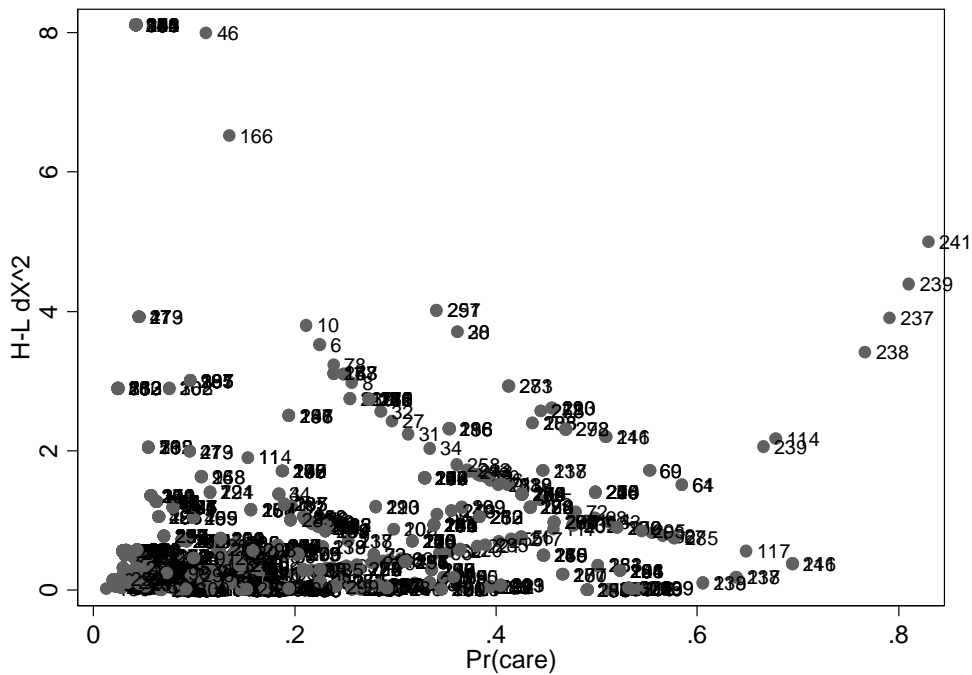
To examine residuals, it is recommended to use standardized Pearson residual that accounts for in-built heteroscedasticity of residuals in the logit model:

```
. predict rstandard, rs
. predict prob
(option p assumed; Pr(care))
. scatter prob rstandard, mlabel(id) xline(0)
```



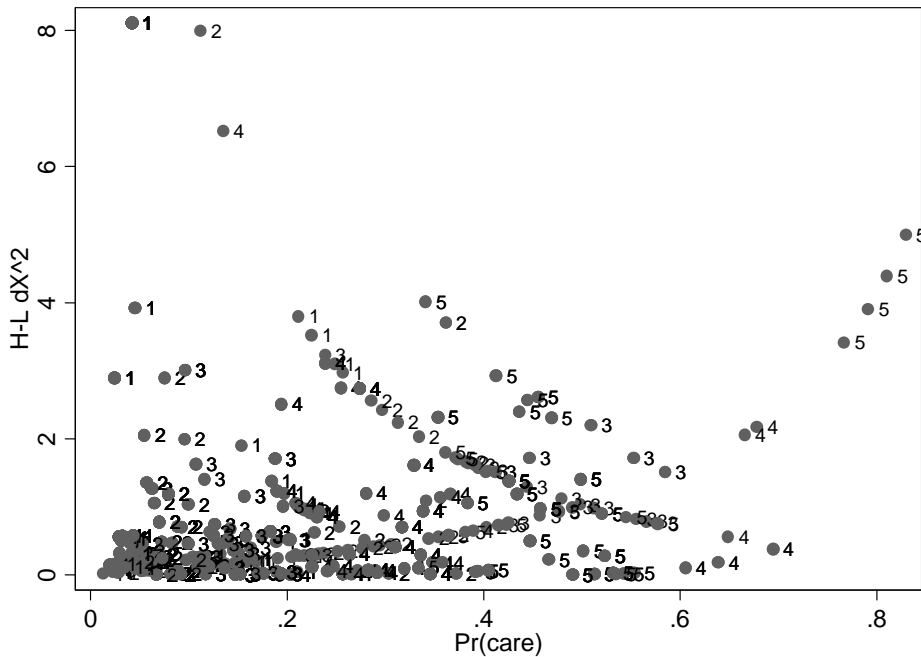
Let's also look at one of the influence statistics:

```
. predict dxt2, dx2
. scatter dxt2 prob, mlabel(id)
```



But what time points are these?

```
. scatter dxt2 prob, mlabel(t)
```



Since we can't see those observations with high dbeta well, let's list them:

```
. list id dxt2 care t marriagel girl age if dxt2>8 & dxt2~=. .
```

id	dxt2	care	t	marria~1	girl	age
----	------	------	---	----------	------	-----

12.	12	8.110425	1	1	1	0	24
14.	14	8.110425	1	1	1	0	24
15.	15	8.110425	1	1	1	0	24
18.	18	8.110425	0	1	1	0	24
58.	38	8.110425	0	1	1	0	24
-----							
183.	84	8.110425	0	1	1	0	24
222.	97	8.110425	0	1	1	0	24
265.	110	8.110425	0	1	1	0	24
353.	132	8.110425	0	1	1	0	24
393.	142	8.110425	0	1	1	0	24
-----							
571.	184	8.110425	0	1	1	0	24
606.	191	8.110425	0	1	1	0	24
671.	204	8.110425	0	1	1	0	24
781.	226	8.110425	0	1	1	0	24
891.	248	8.110425	0	1	1	0	24
-----							
906.	251	8.110425	0	1	1	0	24
956.	261	8.110425	0	1	1	0	24
-----							

We could then investigate how omitting certain covariate patterns like this one affects the coefficients in the model.

#### Article example:

Gupta, Sanjiv, Pamela J. Smock and Wendy D. Manning. 2004. Moving Out: Transition to Nonresidence among Resident Fathers in the United States, 1968-1997. *Journal of Marriage and Family*, 66, August, 627-638.

Questions to answer about the article:

1. Who is included in the sample? When do individuals enter and exit the sample? What constitutes a failure? A censored case? What does Table 1 tell you about the sample?
2. Does the analysis deal with repeatable or non-repeatable events? Single event type of multiple types?
3. What are the dependent and the independent variables in this analysis? Are the independent variables time-invariant or time-variant?
4. What kind of model is used? How is time represented in this model? Do the effects of predictors vary by time?
5. What is reported in Table 2? How can we interpret these results? How do the authors discuss these results in the text?
6. In addition to what the authors chose to present, how else could they have presented their results?
7. What measures of model fit and model diagnostics are presented? What do the authors report in the section titled "How Robust Are the Findings?"
8. What diagnostics and potential problems did the authors not address?