

Logit regression diagnostics (continued)

We will use the dataset on child survival in Bangladesh. As before:

```
logit died prog moth fem y2 y3 y4
```

```
Iteration 0: Log Likelihood =-558.65643
Iteration 1: Log Likelihood =-534.89626
Iteration 2: Log Likelihood =-531.89359
Iteration 3: Log Likelihood =-531.78721
Iteration 4: Log Likelihood = -531.7867
```

```
Logit Estimates                                     Number of obs = 4434
                                                    chi2(6)         = 53.74
                                                    Prob > chi2     = 0.0000
Log Likelihood = -531.7867                       Pseudo R2      = 0.0481
```

died	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
progarea	-.5532452	.1952145	-2.834	0.005	-.9358587	-.1706318
mothed	-.4205078	.2163915	-1.943	0.052	-.8446273	.0036118
female	.6108467	.1911598	3.195	0.001	.2361803	.9855131
y2	-.0847862	.2167913	-0.391	0.696	-.5096893	.3401169
y3	-.5518401	.2535713	-2.176	0.030	-1.048831	-.0548495
y4	-1.64557	.3842616	-4.282	0.000	-2.398708	-.8924308
_cons	-3.193167	.2065652	-15.458	0.000	-3.598027	-2.788306

```
. logistic died prog moth fem y2 y3 y4
```

```
Logit Estimates                                     Number of obs = 4434
                                                    chi2(6)         = 53.74
                                                    Prob > chi2     = 0.0000
Log Likelihood = -531.7867                       Pseudo R2      = 0.0481
```

died	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
progarea	.5750805	.1122641	-2.834	0.005	.3922489	.843132
mothed	.6567133	.1421072	-1.943	0.052	.4297175	1.003618
female	1.84199	.3521146	3.195	0.001	1.266403	2.679186
y2	.9187087	.199168	-0.391	0.696	.6006822	1.405112
y3	.5758892	.146029	-2.176	0.030	.3503472	.9466276
y4	.1929026	.0741251	-4.282	0.000	.0908352	.4096587

The logit command gives us the coefficients and the logistic gives us the odds. (Note: we could also use logit and specify the option to get the odds ratio: logit y x1 x2... , or.) In any case, we see that mother's education is just barely above the significance level and age 2 is not significant (children aged 2 appear to have mortality rates no different from 1-year olds. We can test whether both can be dropped:

```
. test mothed y2
( 1) mothed = 0.0
```

```
( 2)  y2 = 0.0

      chi2( 2) =      3.94
      Prob > chi2 =      0.1391
```

However, before going on to fit models, we may want to see if some observations are influential

```
. lpredict pattern, number
. lpredict phat
. lpredict db, dbeta
. lpredict pear, dx2
. lpredict dev, ddeviance
```

The first assigns a number to all observations with a particular pattern of X values. The above commands give us the predicted probability of dying, and, were we to drop a particular **pattern**, the standardized change in the estimated parameters (db), the change in the sum of squared Pearson residuals (pear) and the change in the deviance residual (dev).

For convenience, I generate a variable, age, so that I don't have to look at the 3 dummy variables y2, y3, y4

```
. gen age=1
. replace age=2 if y2==1      (1135 real changes made)
. replace age=3 if y3==1      (1050 real changes made)
. replace age=4 if y4==1      (1002 real changes made)
```

I then use a trick that is extremely convenient. I am going to want to plot, for example, db vs phat -- but I don't want to plot all 4100 values. Since all phats are the same for observations of a given pattern, I could just plot the values from the **first** observation I encounter with that pattern. STATA offers a way of numbering the observations of a given pattern **and** of finding how many observations there are with that pattern. I have called these variables a and b:

```
. sort pattern age
. quietly by pattern: gen a=_n
. quietly by pattern: gen b=_N
```

If there are 3 observations of pattern 1,

a	b
1	3
2	3
3	3

So a=1 indicates the observation is the first with this pattern, and b tells us how many observations have this pattern.

```

. list pattern prog mothed fem age b if a==1
      pattern  progarea  mothed  female  age  b
  1.         1         0         0         0         1  233
 234.        2         0         0         0         4  185
 419.        3         0         0         0         3  199
 618.        4         0         0         0         2  180
 798.        5         0         0         1         1  240
1038.        6         0         0         1         4  195
1233.        7         0         0         1         3  184
1417.        8         0         0         1         2  206
1623.        9         0         1         0         1  105
1728.       10         0         1         0         4  103
1831.       11         0         1         0         3   93
1924.       12         0         1         0         2  111
2035.       13         0         1         1         1   95
2130.       14         0         1         1         4   85
2215.       15         0         1         1         3   74
2289.       16         0         1         1         2   86
2375.       17         1         0         0         1  202
2577.       18         1         0         0         4  165
2742.       19         1         0         0         3  188
2930.       20         1         0         0         2  175
3105.       21         1         0         1         1  160
3265.       22         1         0         1         4  128
3393.       23         1         0         1         3  160
3553.       24         1         0         1         2  188
3741.       25         1         1         0         1  121
3862.       26         1         1         0         4   72
3934.       27         1         1         0         3   87
4021.       28         1         1         0         2   97
4118.       29         1         1         1         1   91
4209.       30         1         1         1         4   69
4278.       31         1         1         1         3   65
4343.       32         1         1         1         2   92

```

```
. list pattern db pear dev b if a==1
      pattern      db      pear      dev      b
  1.         1  1.596836  2.644484  2.352185  233
 234.        2  .5693872  2.107923  1.611329  185
 419.        3  .2476039  .6165161  .5631102  199
 618.        4  .0303126  .0672187  .0690777  180
 798.        5  .4893375  .4595934  .4762161  240
1038.       6  2.777249  4.571358  9.076692  195
1233.        7  .0160244  .0238062  .0234929  184
1417.        8  .3458719  .394589  .3805942  206
1623.        9  .0058144  .0270165  .0262726  105
1728.       10  .0585308  .5887592  1.174468  103
1831.       11  .2344227  1.648762  3.272194   93
1924.       12  .0109799  .0475254  .0457958  111
2035.       13  .0257099  .0760695  .0788712   95
2130.       14  .3062587  2.050527  1.452026   85
2215.       15  .5013344  2.534262  4.99742   74
2289.       16  .0586837  .2028914  .2172979   86
2375.       17  .2726476  .8077694  .9226949  202
2577.       18  .3287776  2.395278  1.63781   165
2742.       19  .2622082  1.143512  1.481591   188
2930.       20  .0078776  .0286646  .0279715   175
3105.       21  .6926788  1.595751  1.853486   160
3265.       22  .0009076  .0047067  .004804   128
3393.       23  .5363835  1.546059  1.331996   160
3553.       24  .1679513  .3253155  .3445735   188
3741.       25  .0739111  .4579827  .5490509  121
3862.       26  .0091237  .2240451  .4474216   72
3934.       27  .0675907  .8391352  1.670824   87
4021.       28  .1842583  1.545908  3.070008   97
4118.       29  1.260211  5.954288  4.327653   91
4209.       30  .0289487  .4070447  .8118558   69
4278.       31  .4326979  4.061503  2.714949   65
4343.       32  .2798082  1.419796  1.177303   92
```

I then looked at the 2 patterns that had db>1 and the pear and dev>4.

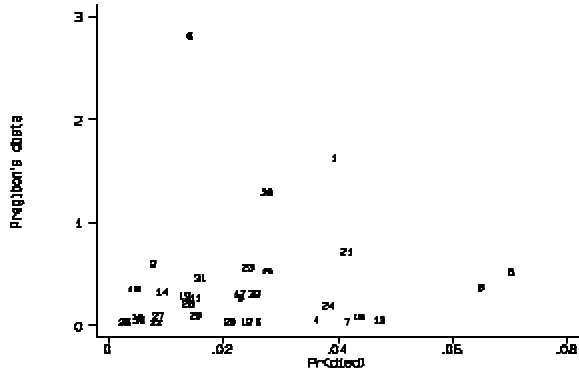
```
. list phat if (pattern==6|pattern==29) & a==1
      phat
1038.  .0143735
4118.  .0277582
```

```
. tab pattern died if pattern==6|pattern==29, r
```

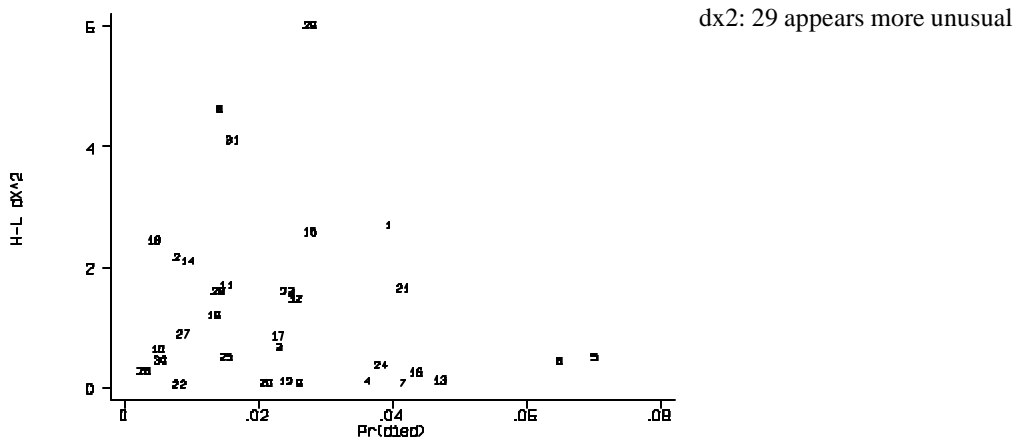
covariate pattern	Died during year		Total
	0	1	
6	195	0	195
	100.00	0.00	100.00
29	85	6	91
	93.41	6.59	100.00

These are quite different -- there are no deaths observed for children of pattern 6, while those in 29 have a much higher than predicted death rate. But we want to consider further whether these are really influential.

We turn to graphs: . graph db phat if a==1, xlabel ylabel symbol([pattern])
 dbeta: 6 stands out

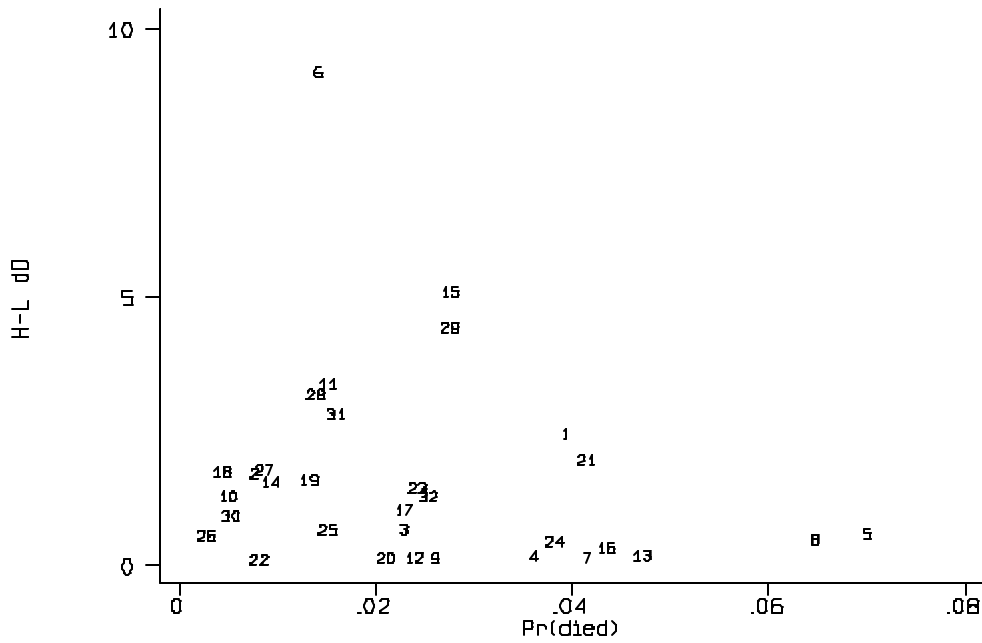


. graph pear phat if a==1, xlabel ylabel symbol([pattern])



. graph dev phat if a==1, xlabel ylabel symbol([pattern])

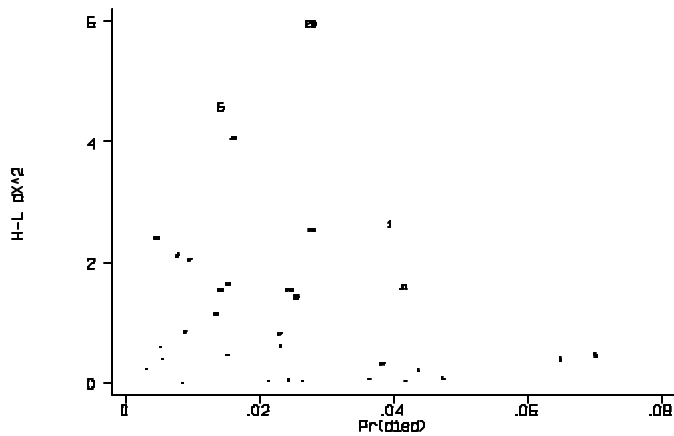
dev: 6 stands out



The next set of graphs *weights* the points by dbeta:

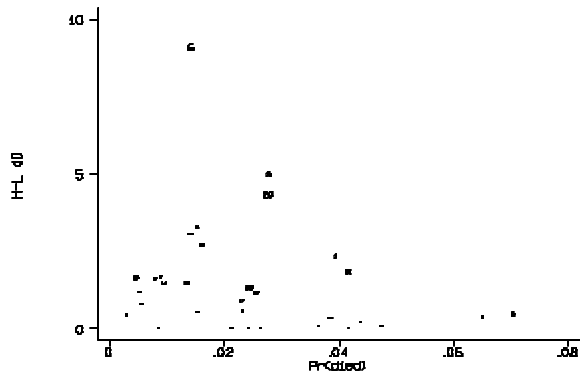
```
. graph pear phat [iweight=db] if a==1, xlabel ylabel symbol([pattern])
```

6 stands out as much larger than 29



```
. graph dev phat [iweight=db] if a==1, xlabel ylabel symbol([pattern])
```

Again 6 stands out



It appears that there's good reason to drop 6 -- it gives no info because ALL survive, *and* it is very influential.

I therefore repeated the analysis, dropping all observations with pattern 6.

```
. drop if pattern ==6                (195 observations deleted)
```

```
. logit died prog moth fem y2 y3 y4
Iteration 0:  Log Likelihood =-553.09083
Iteration 1:  Log Likelihood =  -530.768
Iteration 2:  Log Likelihood =-528.25227
Iteration 3:  Log Likelihood = -528.2081
Iteration 4:  Log Likelihood =-528.20802
```

```
Logit Estimates                                Number of obs =  4239
                                                chi2(6)         =  49.77
                                                Prob > chi2     = 0.0000
Log Likelihood = -528.20802                    Pseudo R2      = 0.0450
```

died	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
progarea	-.6022955	.1955202	-3.080	0.002	-.985508	-.2190829
mothed	-.4678065	.2168744	-2.157	0.031	-.8928725	-.0427404
female	.6688518	.1920838	3.482	0.000	.2923745	1.045329
y2	-.084159	.2169541	-0.388	0.698	-.5093811	.3410632
y3	-.5511517	.2537156	-2.172	0.030	-1.048425	-.0538783
y4	-1.186247	.3883747	-3.054	0.002	-1.947447	-.4250464
_cons	-3.202444	.2060215	-15.544	0.000	-3.606239	-2.79865

```
. logistic died prog mo fem y2 y3 y4
```

```
Logit Estimates                                Number of obs =  4239
                                                chi2(6)         =  49.77
                                                Prob > chi2     = 0.0000
Log Likelihood = -528.20802                    Pseudo R2      = 0.0450
```

died	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
progarea	.5475533	.1070577	-3.080	0.002	.3732496	.8032551
mothed	.6263747	.1358447	-2.157	0.031	.4094778	.9581601
female	1.951995	.3749466	3.482	0.000	1.339605	2.844335
y2	.9192851	.1994426	-0.388	0.698	.6008673	1.406442
y3	.5762857	.1462127	-2.172	0.030	.3504893	.9475474
y4	.3053652	.1185961	-3.054	0.002	.1426378	.6537395

There is relatively little change in the coefficients, except now all are significant except y2.

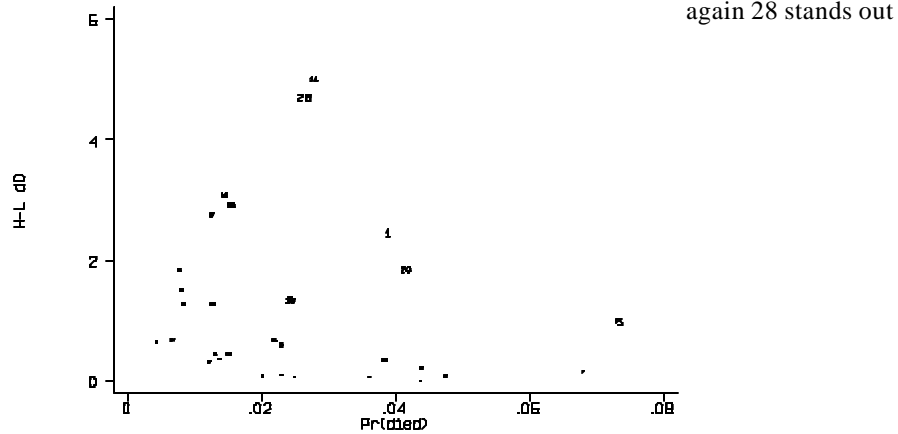
```
. lpredict phat
. lpredict pattern, number
. sort pattern age
. quietly by pattern: gen a=_n
```

```
. quietly by pattern: gen b=_N
. lpredict db, dbeta
. lpredict pear, dx2
. lpredict dev, ddeviance
. list pattern phat db pear dev if a==1
```

	pattern	phat	db	pear	dev
1.	1	.0390738	1.629929	2.759535	2.446634
234.	2	.0122647	.1767586	.3566391	.3239355
419.	3	.0228968	.2582012	.6530921	.5947352
618.	4	.0360336	.0240711	.0544807	.0558422
798.	5	.0735365	1.064688	.923676	.9697531
1038.	6	.0437409	.000379	.0005233	.0005244
1222.	7	.0680047	.138585	.1464232	.1433316
1428.	8	.0248375	.0150787	.0729383	.0696223
1533.	9	.0077177	.1326995	.9170251	1.826955
1636.	10	.0144657	.2106704	1.550529	3.078518
1729.	11	.0228786	.0230861	.1043549	.0986997
1840.	12	.0473627	.0259429	.0776437	.0805366
1935.	13	.0149549	.1397455	.535026	.455973
2020.	14	.0278534	.4981057	2.536556	5.001794
2094.	15	.0437069	.0591023	.2061119	.2208739
2180.	16	.0217801	.1955382	.6027817	.6776432
2382.	17	.006753	.1713634	.8515482	.6884965
2547.	18	.0126684	.2169858	.9897967	1.265544
2735.	19	.0200573	.0232547	.0881966	.084436
2910.	20	.0416509	.6795614	1.587176	1.843379
3070.	21	.0130977	.1289597	.3723953	.4351287
3198.	22	.0244341	.5299857	1.540212	1.327136
3358.	23	.0384182	.1656677	.3249607	.3442278
3546.	24	.0137544	.0476966	.3101797	.3616702
3667.	25	.0042406	.016837	.3226278	.6438854
3739.	26	.0079729	.0570282	.7522293	1.498445
3826.	27	.0126583	.1570298	1.384629	2.751657
3923.	28	.0265015	1.342359	6.600091	4.702913
4014.	29	.0082444	.0663437	.6336491	1.26206
4083.	30	.0154459	.4541816	4.442593	2.908468
4148.	31	.0244147	.3166427	1.669894	1.360223

```
. graph pear phat [iweight=db] if a==1, xlabel ylabel symbol([pattern])
(not shown)
```

```
. graph dev phat [iweight=db] if a==1, xlabel ylabel symbol([pattern])
```



Because death is a rare event at these young ages, many patterns contribute zero deaths. These patterns aren't identified as such because it's not a VARIABLE that completely identifies with the category survived -- it's a combination of variables. These findings indicate that we really need a much larger number of cases to do serious research. Our final analysis had about 45,000 child-years of observation - more than 10 times this sample size.

We might drop pattern 28, but I wouldn't do it yet.