Linear Regression

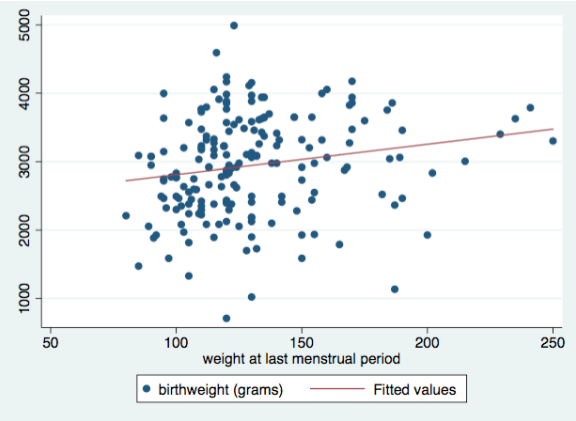
use lbw1.dta, clear

. regress bwt lwt

bwt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lwt	4.429993	1.713244	2.59	0.010	1.050222	7.809763
_cons	2369.184	228.4671	10.37	0.000	1918.479	2819.888

. gen lwt130=lwt-130
. regress bwt lwt130

twoway lfit calculates the prediction for yvar from a linear regression of yvar on xvar and plots the resulting line
. twoway (scatter bwt lwt) (lfit bwt lwt)



Postestimation commands after **regress** (type **help regress postestimation** for a full list of available commands):

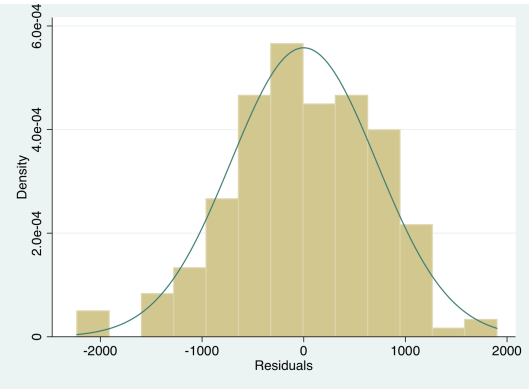
predict can be used to obtain predictions, residuals, influence statistics, etc.
. regress bwt lwt age

You can get fitted values using **predict** *newvar*, **xb**
. predict predval, xb

or store residuals using **predict** *newvar*, **residual**

. predict rbwt, residual

. hist rbwt, normal
. qnorm rbwt



There are other postestimation diagnostic plots like **rvfplot** (to plot residuals versus predicted values) and **rvpplot** (plot residuals versus explanatory variable)

. rvfplot, yline(0)
. rvpplot lwt, yline(0)

Use **lincom** to calculate linear combinations of the model parameters**.** For example, to estimate the expected birthweight of a child born to a 25 year old women who weighed 125lbs: [bwt=β_0 + β_1*lwt+ β_2*age+ error]

```
. lincom _cons + lwt*125 + age*25

 ( 1)  125*lwt + 25*age + _cons = 0
```

bwt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	2938.177	56.33194	52.16	0.000	2827.045	3049.308

Categorical Predictors (factor variables):

. tab race

race	Freq.	Percent	Cum.
1. white	96	50.79	50.79
2. black	26	13.76	64.55
3. other	67	35.45	100.00

```
. regress bwt i.race, noheader
```

bwt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
race						
2	-383.3181	157.8914	-2.43	0.016	-694.8064	-71.82985
3	-298.9955	113.6899	-2.63	0.009	-523.2829	-74.70811
_cons	3103.01	72.88956	42.57	0.000	2959.214	3246.807

```
. regress bwt b2.race, noheader
```

bwt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
race						
1	383.3181	157.8914	2.43	0.016	71.82985	694.8064
3	84.32262	165.0131	0.51	0.610	-241.2152	409.8604
_cons	2719.692	140.0601	19.42	0.000	2443.382	2996.003

* the b2. prefix specifies to use the 2nd category as the base (or reference)

Interactions in regression models

Interaction between continuous (lwt130) and categorical (race) variables:

```
. regress bwt i.race c.lwt130 i.race#c.lwt130, noheader
```

bwt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
race						
2	-413.8986	167.1201	-2.48	0.014	-743.6285	-84.16868
3	-227.6167	117.5805	-1.94	0.054	-459.6043	4.370945
lwt130	4.985665	2.487707	2.00	0.047	.0773892	9.89394
race#c.lwt130						
2	-2.55752	4.3425	-0.59	0.557	-11.12532	6.010284
3	1.147422	4.258696	0.27	0.788	-7.255035	9.54988
_cons	3092.779	72.17968	42.85	0.000	2950.368	3235.191

note: the c. prefix specifies lwt130 as a continuous variable for the interaction. Also the same analysis could be done using the shorter command: regress bwt b2.race##c.lwt130

Interaction between two categorical variables:
. tab smoke race

```
. regress bwt lwt130 b0.smoke##b2.race, noheader
```

bwt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lwt130	3.565503	1.708612	2.09	0.038	.1942672	6.936739
1.smoke	-326.1209	273.234	-1.19	0.234	-865.2345	212.9928
race						
1	612.0322	198.5243	3.08	0.002	220.3271	1003.737
3	67.67792	199.1864	0.34	0.734	-325.3336	460.6894
smoke#race						
1 1	-230.5573	306.36	-0.75	0.453	-835.0313	373.9167
1 3	251.8071	348.5279	0.72	0.471	-435.8677	939.4819
_cons	2785.196	172.5243	16.14	0.000	2444.791	3125.601

. testparm i.smoke#i.race

This tests the hypothesis of additivity between race and smoking (i.e., no interaction)

Notes on prefixes for interaction models:

- a categorical variable is often specified with a i. or b#. prefix, where # is the integer value you want to use as the reference (base) level
- continuous variables are specified with the c. prefix in interaction terms
- an interaction term (without main effects) is specified by one #
- an interaction term with main effects is specified by ##
- higher order interactions are allowed, for example i.smoke##i.race##c.lwt130

Logistic Regression

There are two main commands for logistic regression: **logistic** and **logit**
Logistic returns odds ratios by default, while the default for **logit** is log odds-ratios (but you can get odds ratios by using the **or** option)

. tabodds low race, or

. logistic low i.race
. logit low i.race
. logit low i.race, or

. logistic low i.race lwt130

low	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
race						
2	2.946799	1.438078	2.21	0.027	1.132277	7.669169
3	1.61775	.5769647	1.35	0.177	.8041433	3.254538
lwt130	.9849141	.0063409	-2.36	0.018	.9725641	.9974208
_cons	.3093892	.0754679	-4.81	0.000	.1918119	.4990394

Adjusted for weight at last menstrual period, the odds ratio for blacks compared with whites is 2.95 (1.13, 7.67). The OR for a 1 lb difference in lwt is 0.98 (0.97, 0.99). To get the OR for a 10 lb difference use **lincom**

```
. lincom 10*lwt130

 ( 1)  10*[low]lwt130 = 0
```

low	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.8589807	.0553018	-2.36	0.018	.7571509	.9745057

To get estimated probabilities for each subject in the dataset:
. predict predprob

Other logistic models:

Conditional logistic regression for matched case control using **clogit**
Multinomial logistic regression using **mlogit** and ordinal logistic regression using **ologit**
In the case of sparse data you can fit a logistic model by exact models using **exlogistic**

Incidence Rate Data

. use compliance2.dta, clear

. codebook, compact

```
Variable   Obs Unique   Mean        Min         Max   Label

id         555    555  6199.575    30          12630 Person number
particip   555      2  .7405405    0           1     Participated
birthdate  555    508  -12484.96   -14235      -9502 Date of birth
randate    555     12  12707.7     12512       14122 Date of randomization
enddate    555    101  14441.18    12597       14609 Date of last observation
died       555      2  .1855856    0           1     Died
ranage     555    494  68.97373    64.27       73.79 Age at randomization
ranagr     555      5  67.94595    64          72    Age group at randomization
```

This is data on a cohort of 555 men invited to participate in a screening trial for aneurysms. The men were followed until death or end of follow up.

To study the association between participation and mortality for incidence-rate data we can use **ir** (an epitab command like **cs** and **cc**). The syntax is **ir** *var_case var_exposed var_time*

Define a variable for person time at risk (in years):
. gen pyr=(enddate-randate)/365

. ir died particip pyr
. ir died particip pyr, level(99)

To adjust for age group by stratification:
. ir died particip pyr, by(ranagr)

```
. ir died particip pyr, by(ranagr)

Age group at ran |      IRR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
              64 |   .3140617      .0744762   1.513131     3.307495 (exact)
              66 |   .4926495      .1420931   2.151985     3.341521 (exact)
              68 |   .5236575       .228229    1.24793     8.088149 (exact)
              70 |   .7504398      .2471127   2.512317     4.137178 (exact)
              72 |   .4302365      .2117449   .8811187     12.27482 (exact)
-----------------+------------------------------------------------
           Crude |   .4490994      .2982485   .6820111              (exact)
    M-H combined |   .4913825      .3313807   .7286386
-----------------------------------------------------------------
 Test of homogeneity (M-H)    chi2(4) =      1.34  Pr>chi2 = 0.8549
```

The regression model corresponding to **ir** is poisson:
. poisson died particip i.ranagr, exposure(pyr) irr

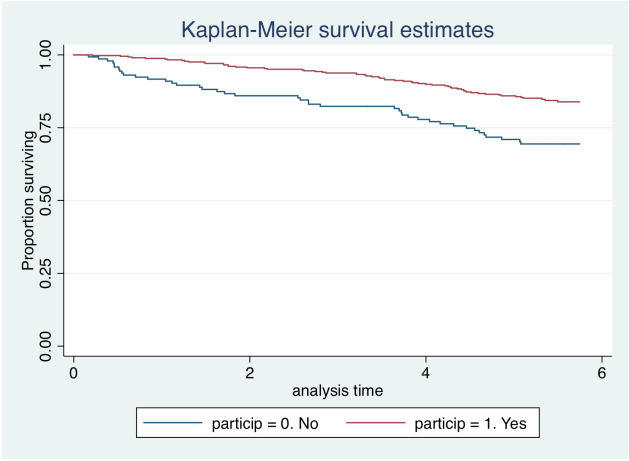
Time to Event (Survival) Models

With time to event data, it is typical to set the time scale and event using **stset.** Stata remembers this information so you do not have to repeat it each time you issue a survival command.

```
. gen timeatrisk=(enddate-randate)/365

. stset timeatrisk, failure(died)

     failure event:  died != 0 & died < .
obs. time interval:  (0, timeatrisk]
 exit on or before:  failure

------------------------------------------------------------------------------
      555  total obs.
        0  exclusions
------------------------------------------------------------------------------

      555  obs. remaining, representing
      103  failures in single record/single failure data
 2635.836  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  5.745205
```

Kaplan-Meier survival function:

. sts graph, by(particip) ytitle("Proportion surviving")



Log-rank test for equality of survivor functions:
. sts test particip

To get the incidence rate from data that is stset can use **stir** (instead of using **ir**)
. stir particip, by(ranagr)

Cox proportional Hazards regression

. stcox particip

```
Cox regression -- Breslow method for ties

No. of subjects =          555                     Number of obs    =       555
No. of failures =          103
Time at risk    =  2635.835618
                                                   LR chi2(1)       =     15.03
Log likelihood  =   -626.22599                     Prob > chi2      =    0.0001
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
particip	.4463335	.089556	-4.02	0.000	.3012086	.6613808

. stcox particip i.ranagr
. stcurve, survival at1(ranagr=72 particip=0) at2(ranagr=72 particip=1)