台中榮總2023/3/16

STATA 入門體驗工作坊

昊青公司 蕭鎮宇

02-25050525

迴歸分析(regression;OLS)

- 基本指令: reg y x1 x2 x3
- 安裝外掛插件: ssc install …

- 搭配概念:
  - ➢虛擬變數設定
  - ➢產生預測值
  - ➢輸出報表

```
. do "C:\Users\YYChen\AppData\Local\Temp\STD2424_000000.tmp"

. regress bpsystol bmi height
```

Source	SS	df	MS	Number of obs	=	10,351
				F(2, 10348)	=	718.33
Model	686920.7	2	343460.35	Prob > F	=	0.0000
Residual	4947749.33	10,348	478.135807	R-squared	=	0.1219
				Adj R-squared	=	0.1217
Total	5634670.03	10,350	544.412563	Root MSE	=	21.866

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.653363	.043859	37.70	0.000	1.567391	1.739335
height	-.0234993	.0223247	-1.05	0.293	-.0672601	.0202615
_cons	92.59841	3.993783	23.19	0.000	84.76982	100.427

虛擬變數(dummy variable)

- 將類別變數轉換成一系列的0/1變數，好進行迴歸分析

Name	Class	d_classA	d_classB	d_classC
John	A	1	0	0
Marry	B	0	1	0
Lisa	B	0	1	0
Ted	C	0	0	1
Mike	A	1	0	0
Nancy	C	0	0	1

類別虛擬變數 (factor variable) 運算子

Operator	Description
`i.`	unary operator to specify indicators
`c.`	unary operator to treat as continuous
`o.`	unary operator to omit a variable or indicator
`#`	binary operator to specify interactions
`##`	binary operator to specify full-factorial interactions

```
. list group i.group in 1/5
```

	group	1b. group	2. group	3. group
1.	1	0	0	0
2.	1	0	0	0
3.	2	0	1	0
4.	2	0	1	0
5.	3	0	0	1

Factor specification	Result
`i.group`	indicators for levels of **group**
`i.group#i.sex`	indicators for each combination of levels of **group** and **sex**, a two-way interaction
`c.age#c.age`	**age** squared
`c.age#c.age#c.age`	**age** cubed

類別虛擬變數 (factor variable) 運算子

```
. regress bpsystol bmi i.sex
```

Source	SS	df	MS		
Model	725791.096	2	362895.548	Number of obs	= 10,351
Residual	4908878.93	10,348	474.379487	F(2, 10348)	= 764.99
				Prob > F	= 0.0000
				R-squared	= 0.1288
				Adj R-squared	= 0.1286
Total	5634670.03	10,350	544.412563	Root MSE	= 21.78

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.659014	.043559	38.09	0.000	1.57363	1.744398
sex						
Female	-3.907013	.4287053	-9.11	0.000	-4.747358	-3.066668
_cons	90.56624	1.153803	78.49	0.000	88.30457	92.82792

i.Sex

i. 可以創造類別虛擬變數

```
. do "C:\Users\YYChen\AppData\Local\Temp\STD2424_000000.tmp"

. regress bpsystol bmi i.sex i.region
```

Source	SS	df	MS		
Model	726569.077	5	145313.815	Number of obs	= 10,351
Residual	4908100.95	10,345	474.441851	F(5, 10345)	= 306.28
				Prob > F	= 0.0000
				R-squared	= 0.1289
				Adj R-squared	= 0.1285
Total	5634670.03	10,350	544.412563	Root MSE	= 21.782

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.658517	.0435675	38.07	0.000	1.573116	1.743918
sex						
Female	-3.904151	.4287719	-9.11	0.000	-4.744626	-3.063675
region						
MW	-.7519373	.630418	-1.19	0.233	-1.987678	.4838039
S	-.2434113	.626676	-0.39	0.698	-1.471817	.9849948
W	-.5055111	.637923	-0.79	0.428	-1.755964	.7449413
_cons	90.97439	1.230431	73.94	0.000	88.5625	93.38627

```
. do "C:\Users\YYChen\AppData\Local\Temp\STD2424_000000.tmp"

. regress bpsystol c.bmi##i.sex i.region
```

做完全的效應修飾因子

c.連續 dummy ## i.類別dummy

Source	SS	df	MS		
Model	731872.976	6	121978.829	Number of obs	= 10,351
Residual	4902797.05	10,344	473.974966	F(6, 10344)	= 257.35
				Prob > F	= 0.0000
				R-squared	= 0.1299
				Adj R-squared	= 0.1294
Total	5634670.03	10,350	544.412563	Root MSE	= 21.771

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.445272	.0772005	18.72	0.000	1.293944	1.5966
sex						
Female	-11.88881	2.42508	-4.90	0.000	-16.64244	-7.135184
sex#c.bmi						
Female	.312816	.0935124	3.35	0.001	.1295137	.4961184
region						
MW	-.7805466	.6301658	-1.24	0.216	-2.015793	.4547001
S	-.2998277	.6265946	-0.48	0.632	-1.528074	.9284188
W	-.529543	.6376495	-0.83	0.406	-1.779459	.7203733
_cons	96.44331	2.04579	47.14	0.000	92.43317	100.4535

女性是BMI對SBP的效應修飾因子

```
. regress bpsystol i.sex##i.region
```

懷疑sex和region有交互作用

Source	SS	df	MS		
Model	42826.5815	7	6118.08307	Prob > F	0.0000
Residual	5591843.44	10,343	540.640379	R-squared	0.0076
Total	5634670.03	10,350	544.412563	Root MSE	23.252

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
sex						
Female	-5.318304	1.016171	-5.23	0.000	-7.310195	-3.326413
region						
MW	-1.263899	.9714852	-1.30	0.193	-3.168198	.6404002
S	-1.881623	.9679706	-1.94	0.052	-3.779033	.0157863
W	-1.505315	.9807497	-1.53	0.125	-3.427774	.4171445
sex#region						
Female#MW	.8301454	1.347069	0.62	0.538	-1.81037	3.470661
Female#S	3.302631	1.33938	2.47	0.014	.6771872	5.928076
Female#W	1.452027	1.362776	1.07	0.287	-1.219277	4.123331
_cons	134.1189	.728753	184.04	0.000	132.6904	135.5474

安裝外掛插件

```
. ssc install outreg2
checking outreg2 consistency and verifying not already installed...
installing into c:\ado\plus\...
installation complete.
```

* outreg2
*安裝外掛插件 ssc install outreg2
regress bpsystol bmi
outreg2 using test, replace cttop(full) excel
regress bpsystol bmi i.sex
outreg2 using test, cttop(full) excel
regress bpsystol bmi i.sex i.region
outreg2 using test, cttop(full) excel
regress bpsystol c.bmi##i.sex i.region
outreg2 using test, cttop(full) excel

```
. regress bpsystol bmi
```

Source	SS	df	MS		
Model	686390.93	1	686390.93	Number of obs	= 10,351
Residual	4948279.1	10,349	478.140796	F(1, 10349)	= 1435.54
				Prob > F	= 0.0000
				R-squared	= 0.1218
				Adj R-squared	= 0.1217
Total	5634670.03	10,350	544.412563	Root MSE	= 21.866

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.656894	.0437307	37.89	0.000	1.571173	1.742615
_cons	88.56855	1.137272	77.88	0.000	86.33928	90.79783

```
. outreg2 using test, replace cttop(full) excel
test.xml
dir : seeout

. seeout using
Hit Enter to

. seeout using
Hit Enter to
```

Command

全院教育課程規劃-20
(C:)
磁碟 (D:)
已選取 1 個項目

Data Editor (Browse) - [Untitled]

File Edit View Data Tools

var5[12]

	v1	v2	Notes_Titles
1		(1)	
2		full	Standard errors in parentheses
3	VARIABLES	bpsystol	*** p<0.01, ** p<0.05, * p<0.1
4			
5	bmi	1.657***	
6		(0.0437)	
7	Constant	88.57***	
8		(1.137)	
9			
10	Observations	10,351	
11	R-squared	0.122	

```
. regress bpsystol c.bmi##i.sex i.region
```

Source	SS	df	MS		
Model	731872.976	6	121978.829	Number of obs	= 10,351
Residual	4902797.05	10,344	473.974966	F(6, 10344)	= 257.35
				Prob > F	= 0.0000
				R-squared	= 0.1299
				Adj R-squared	= 0.1294
Total	5634670.03	10,350	544.412563	Root MSE	= 21.771

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.445272	.0772005	18.72	0.000	1.293944	1.5966
sex						
Female	-11.88881	2.42508	-4.90	0.000	-16.64244	-7.135184
sex#c.bmi						
Female	.312816	.0935124	3.35	0.001	.1295137	.4961184
region						
MW	-.7805466	.6301658	-1.24	0.216	-2.015793	.4547001
S	-.2998277	.6265946	-0.48	0.632	-1.528074	.9284188
W	-.529543	.6376495	-0.83	0.406	-1.779459	.7203733
_cons	96.44331	2.04579	47.14	0.000	92.43317	100.4535

```
* outreg2
*安裝外掛插件  ssc install outreg2
regress bpsystol bmi
outreg2 using test, replace cttop(full) excel
regress bpsystol bmi i.sex
outreg2 using test, cttop(full) excel
regress bpsystol bmi i.sex i.region
outreg2 using test, cttop(full) excel
regress bpsystol c.bmi##i.sex i.region
outreg2 using test, cttop(full) excel
```

從我開始 輸出一系列的
regression model 排列輸出

Data Editor (Browse) - [Untitled]

File Edit View Data Tools

var8[13]

regress bpsystol c.bmi##i.sex i.region
outreg2 using test, cttop(full) excel

	v1	v2	v3	v4	Notes_Titles
1		(1)	(2)	(3)	
2		full	full	full	Standard errors in parentheses
3	VARIABLES	bpsystol	bpsystol	bpsystol	*** p<0.01, ** p<0.05, * p<0.1
4					
5	bmi	1.657***	1.445***	1.445***	
6		(0.0437)	(0.0772)	(0.0772)	
7	2.sex		-11.89***	-11.89***	
8			(2.425)	(2.425)	
9	1b.sex#co.bmi		0	0	
10			(0)	(0)	
11	2.sex#c.bmi		0.313***	0.313***	
12			(0.0935)	(0.0935)	
13	2.region		-0.781	-0.781	
14			(0.630)	(0.630)	
15	3.region		-0.300	-0.300	
16			(0.627)	(0.627)	
17	4.region		-0.530	-0.530	
18			(0.638)	(0.638)	
19	Constant	88.57***	96.44***	96.44***	
20		(1.137)	(2.046)	(2.046)	
21					
22	Observations	10,351	10,351	10,351	
23	R-squared	0.122	0.130	0.130	

robust: 在variance不一致的情況下，算出較穩定的std

```
. regress bpsystol bmi i.sex, robust

Linear regression                               Number of obs     =     10,351
                                                F(2, 10348)       =     627.58
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1288
                                                Root MSE          =      21.78
```

bpsystol	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.659014	.0500087	33.17	0.000	1.560988	1.757041
sex						
Female	-3.907013	.4255359	-9.18	0.000	-4.741146	-3.072881
_cons	90.56624	1.304872	69.41	0.000	88.00844	93.12404

```
. do "C:\Users\YYChen\AppData\Local\Temp\STD2424_000000.tmp"

. regress bpsystol bmi i.sex
```

Source	SS	df	MS
Model	725791.096	2	362895.548
Residual	4908878.93	10,348	474.379487
Total	5634670.03	10,350	544.412563

```
                                                Number of obs     =     10,351
                                                F(2, 10348)       =     764.99
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1288
                                                Adj R-squared     =     0.1286
                                                Root MSE          =      21.78
```

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.659014	.043559	38.09	0.000	1.57363	1.744398
sex						
Female	-3.907013	.4287053	-9.11	0.000	-4.747358	-3.066668
_cons	90.56624	1.153803	78.49	0.000	88.30457	92.82792

```
. regress bpsystol bmi
```

Source	SS	df	MS		
Model	686390.93	1	686390.93	Number of obs	= 10,351
Residual	4948279.1	10,349	478.140796	F(1, 10349)	= 1435.54
				Prob > F	= 0.0000
				R-squared	= 0.1218
				Adj R-squared	= 0.1217
Total	5634670.03	10,350	544.412563	Root MSE	= 21.866

bpsystol	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
bmi	1.656894	.0437307	37.89	0.000	1.571173	1.742615
_cons	88.56855	1.137272	77.88	0.000	86.33928	90.79783

```
. test bmi=1

 ( 1)  bmi = 1

       F(  1, 10349) =  225.64
            Prob > F =    0.0000

. test bmi=1.65

 ( 1)  bmi = 1.65

       F(  1, 10349) =    0.02
            Prob > F =    0.8747

.
```

Logistic Regression(羅吉斯迴歸)

- 基本指令: logit y x1 x2 x3

- 只適用於outcome為0/1變數
- 可附加**or**，改為勝算比呈現
- 非線性關係，預測機率需要使用predict指令較方便

```
. use lbw, clear
(Hosmer & Lemeshow data)

. logit low age lwt

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -113.61062
Iteration 2:   log likelihood = -113.56933
Iteration 3:   log likelihood = -113.56932

Logistic regression                                     Number of obs =    189
                                                        LR chi2(2)    =   7.53
                                                        Prob > chi2   = 0.0231
Log likelihood = -113.56932                             Pseudo R2     = 0.0321

------------------------------------------------------------------------------
         low | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -.0398125   .0322857    -1.23   0.218    -.1030913    .0234664
         lwt |  -.0127544   .0062101    -2.05   0.040     -.024926   -.0005827
       _cons |   1.746786   .9970917     1.75   0.080    -.2074774    3.70105
------------------------------------------------------------------------------
```

用媽媽的特徵 來預測小孩是否會體重過輕

可附加**or**，改為勝算比呈現

```
. logit low age lwt i.race, or

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -111.44695
Iteration 2:   log likelihood = -111.33851
Iteration 3:   log likelihood = -111.33847
Iteration 4:   log likelihood = -111.33847

Logistic regression                                     Number of obs =    189
                                                        LR chi2(4)    =  12.00
                                                        Prob > chi2   = 0.0174
Log likelihood = -111.33847                             Pseudo R2     = 0.0511

------------------------------------------------------------------------------
         low | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .9747731   .0324118    -0.77   0.442      .913273    1.040415
         lwt |   .9857717   .0064287    -2.20   0.028     .9732518    .9984526
             |
        race |
       Black |   2.727539   1.358248     2.01   0.044     1.027764    7.238499
       Other |   1.558693   .5614898     1.23   0.218     .7693628    3.157838
             |
       _cons |   3.685916   3.942892     1.22   0.223     .4528972    29.99793
------------------------------------------------------------------------------
```

logit low age lwt i.race, or

predict low_prob

Data Editor (Browse) - [lbw]

File Edit View Data Tools

1C .31309325

	id	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt	low_prob
1	85	0	19	182	Black	Nonsmoker	0	0	1	0	2523	.3130932
2	86	0	33	155	Other	Nonsmoker	0	0	0	3	2551	.2114791
3	87	0	20	105	White	Smoker	0	0	0	1	2557	.3293344
4	88	0	21	108	White	Smoker	0	0	1	2	2594	.3143761
5	89	0	18	107	White	Smoker	0	0	1	0	2600	.3343096
6	91	0	21	124	Other	Nonsmoker	0	0	0	0	2622	.3623497
7	92	0	22	118	White	Nonsmoker	0	0	0	1	2637	.2791675
8	93	0	17	103	Other	Nonsmoker	0	0	0	1	2637	.4595792
9	94	0	29	123	White	Smoker	0	0	0	1	2663	.2316351
10	95	0	26	113	White	Smoker	0	0	0	0	2665	.2730616
11	96	0	19	95	Other	Nonsmoker	0	0	0	0	2722	.4753965
12	97	0	19	150	Other	Nonsmoker	0	0	0	1	2733	.291797
13	98	0	22	95	Other	Nonsmoker	0	1	0	0	2750	.4563253
14	99	0	30	107	Other	Nonsmoker	1	0	1	2	2750	.3655128
15	100	0	18	100	White	Smoker	0	0	0	0	2769	.3569921

Cox Proportional Hazard Model

- 存活分析半參數模型，需要先宣告存活資料
- 存活資料重要三元素:存活時間、事件發生、設限

- **宣告存活資料**:stset *timevar*, failure(*failvar*)

  stset 存活時間, failure(failvar)
- 存活資料型態差異甚大，有些甚至需要進行資料處理，無法一個指令打遍天下
- 除了Cox Model還有Life Table, Kaplan Meier法可使用

use drugtr, clear
stset studytime, failure(died)
stcox drug age 因為已經宣告存活資料，所以不用放Y
stphplot , by(drug)

```
. stcox drug age

        Failure _d: died
  Analysis time _t: studytime

Iteration 0:   log likelihood = -99.911448
Iteration 1:   log likelihood = -83.551879
Iteration 2:   log likelihood = -83.324009
Iteration 3:   log likelihood = -83.323546
Refining estimates:
Iteration 0:   log likelihood = -83.323546

Cox regression with Breslow method for ties

No. of subjects =  48                                 Number of obs =     48
No. of failures =  31
Time at risk    = 744
                                                      LR chi2(2)    =  33.18
Log likelihood = -83.323546                           Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
        drug |   .1048772   .0477017    -4.96   0.000     .0430057    .2557622
         age |   1.120325   .0417711     3.05   0.002     1.041375     1.20526
------------------------------------------------------------------------------

. stphplot , by(drug)

        Failure _d: died
  Analysis time _t: studytime
```

Cox Model

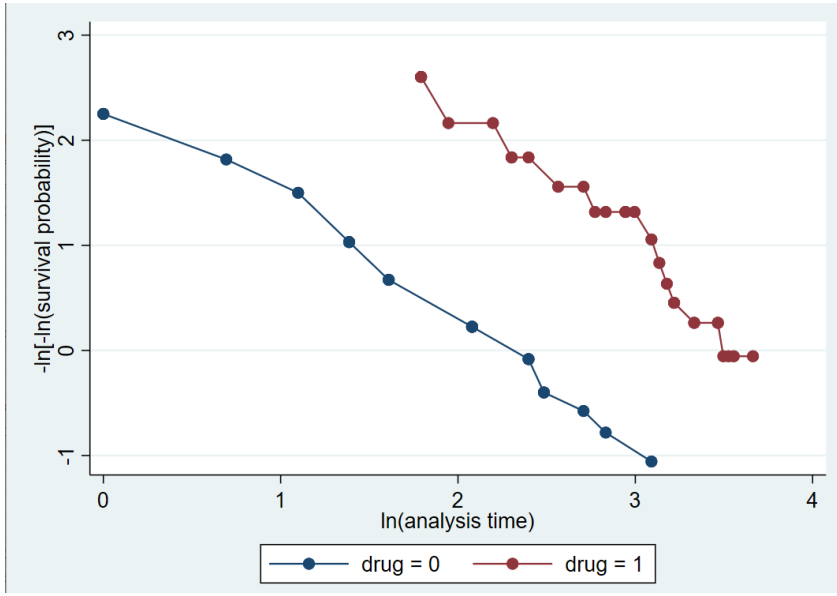
- 估計指令: stcox x1 x2 x3
- 係數取指數後為hazard ratio之倍數，小於1表示該因子可預防event發生，大於1表示該因子有助於event發生
- 非線性關係，不可直接倍數推論

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\hat{h}_0(t) \exp\left[\sum \hat{\beta}_i X_i^*\right]}{\hat{h}_0(t) \exp\left[\sum \hat{\beta}_i X_i\right]} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right]$$

where $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_p^*)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ denote the set of X's for two individuals.

Survival Function Graph

- stphplot , by(*varname*)
- 在Cox Model中又稱為**Log-log Plot**
- 理論上兩線不得相交，否則即違反”proportional”假設



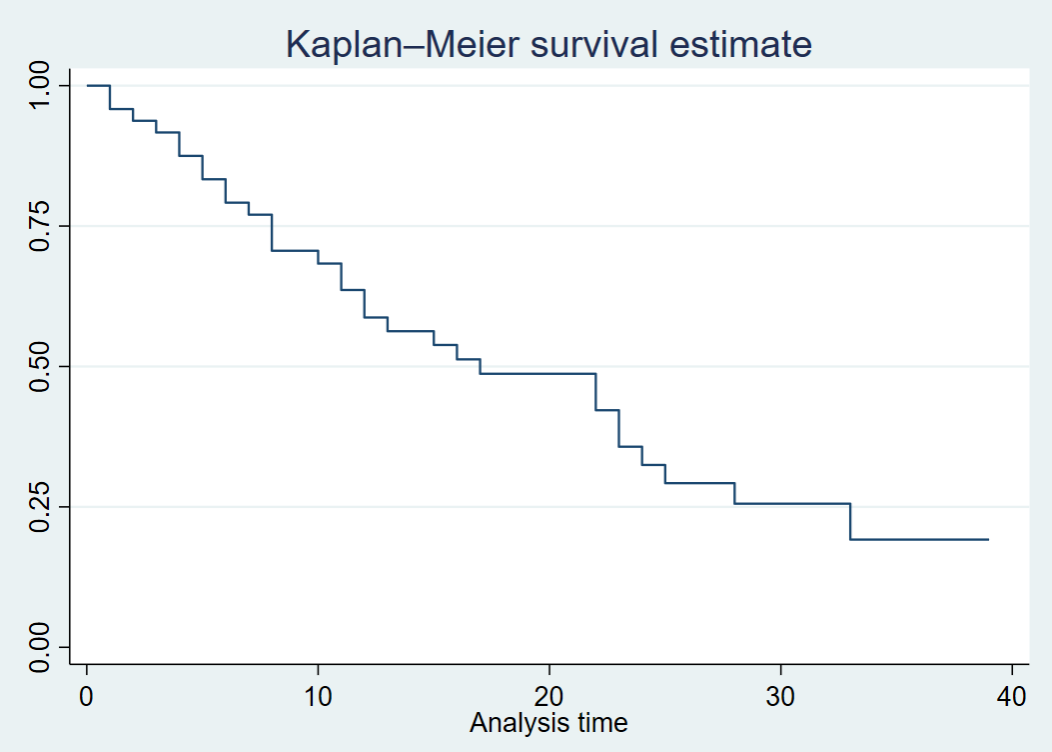
sts list

```
  Analysis time _t: studytime

. sts list

      Failure _d: died
  Analysis time _t: studytime

Kaplan-Meier survivor function
```

Time	At risk	Fail	Lost	Survivor function	Std. error	[95% conf.	int.]
1	48	2	0	0.9583	0.0288	0.8435	0.9894
2	46	1	0	0.9375	0.0349	0.8186	0.9794
3	45	1	0	0.9167	0.0399	0.7930	0.9679
4	44	2	0	0.8750	0.0477	0.7427	0.9418
5	42	2	0	0.8333	0.0538	0.6943	0.9129
6	40	2	1	0.7917	0.0586	0.6474	0.8820
7	37	1	0	0.7703	0.0608	0.6236	0.8656
8	36	3	1	0.7061	0.0661	0.5546	0.8143
9	32	0	1	0.7061	0.0661	0.5546	0.8143
10	31	1	1	0.6833	0.0678	0.5302	0.7957
11	29	2	1	0.6362	0.0708	0.4807	0.7564
12	26	2	0	0.5872	0.0733	0.4304	0.7145
13	24	1	0	0.5628	0.0742	0.4060	0.6931
15	23	1	1	0.5383	0.0749	0.3821	0.6712
16	21	1	0	0.5127	0.0756	0.3570	0.6483
17	20	1	1	0.4870	0.0761	0.3326	0.6249
19	18	0	2	0.4870	0.0761	0.3326	0.6249
20	16	0	1	0.4870	0.0761	0.3326	0.6249

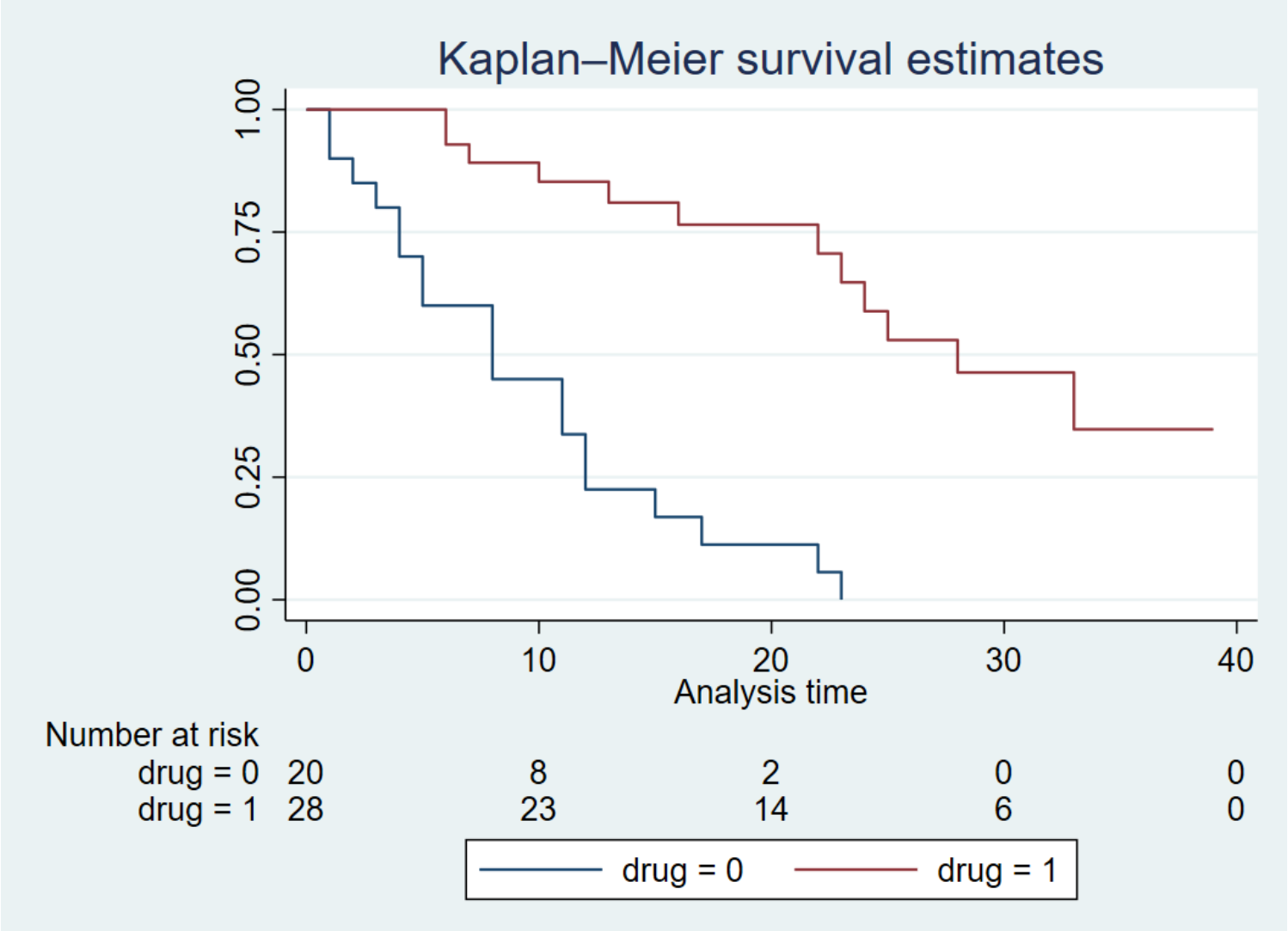
sts graph

Failure _d: died
Analysis time _t: studytime



. sts graph, by(drug) risktable

Failure _d: died
Analysis time _t: studytime



|| 後面的指令和前面的放在同一張圖上

```
twoway scatter bwt lwt || lfit bwt lwt , title(Mother's Weight &
Newborn's Weight) ytitle(Newborn's Weight (gram))
xtitle(Mother's Weight (lb)) by(smoke)
```