Modeling the probability of occurrence of events with the new `stpreg` command

Matteo Bottai, ScD[1] Andrea Discacciati, PhD[1] Giola Santoni, PhD[1]

[1]Karolinska Institutet
Stockholm, Sweden

26 September 2019

The probability function

Let T indicate the time to an event.
Let $S(t) = P(T > t)$ be its survival function.

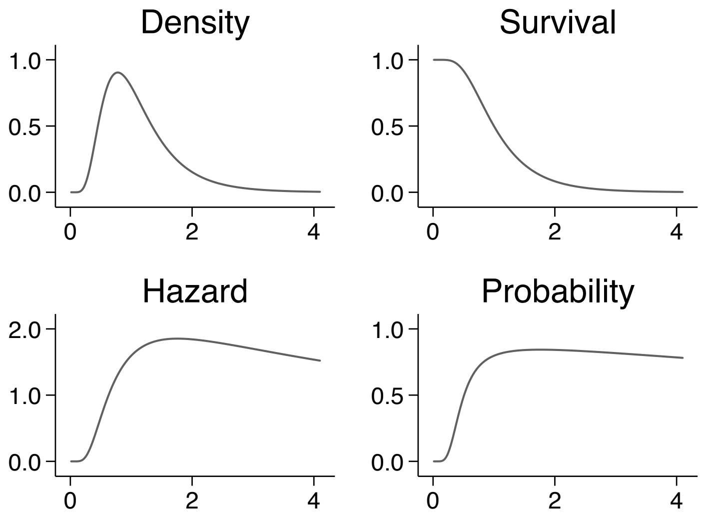
The *probability function* is (Bottai, 2017)

$$g(t) = 1 - \lim_{\delta \to 0} P(T > t + \delta \mid T > t)^{\frac{1}{\delta}} = 1 - \lim_{\delta \to 0} \left[\frac{S(t+\delta)}{S(t)} \right]^{\frac{1}{\delta}}$$

The above is the probability of an event at time t given $T > t$.

Suppose t is time to death in years and $g(t) = 0.25$.
Then 25% of the population is expected to die every year.

Log-normal time to event



A two-population example

The annual risk in two populations is

$$g_0(t) = 0.5 \quad \text{and} \quad g_1(t) = 0.9$$

The risk ratio, odds ratio, and hazard ratio are

$$\text{RR}(t) = 1.8 \qquad \text{OR}(t) = 9.0 \qquad \text{HR}(t) = 3.3$$

The hazard ratio is not a risk ratio.

The new `stpreg` command

- Estimates virtually any probability function model
- Allows time-dependent effects
- Has postestimation commands (`predict`, `test`, `lincom`, `estat`, ...)
- Stems from `stgenreg` by Crowther and Lambert (2013)

Download it with

```
. net from http://www.imm.ki.se/biostatistics/stata
. net install stpreg
```


Proportional-odds models

Let x denote a covariate.
We consider the proportional-odds model

$$\frac{g(t|x)}{1-g(t|x)} = \frac{g_0(t)}{1-g_0(t)} \exp(\beta_1 x)$$

The above can be written as

$$\text{logit } g(t|x) = \text{logit } g_0(t) + \beta_1 x$$

The baseline function can be anything, e.g.

$$\text{logit } g_0(t) = \theta_0 + \theta_1 t$$
$$\text{logit } g_0(t) = \theta_0 + \theta_1 \text{spline}_1(t) + \theta_2 \text{spline}_2(t)$$

The quantity $\exp(\beta_1)$ is the odds ratio per unit-increase in x.

Flexible proportional-odds model

We estimate a flexible proportional-odds model

```
. qui webuse brcancer, clear
. qui stset rectime, failure(censrec = 1) scale(3652.4)
. stpreg x4a, df(2) nolog
Event-probability regression                    Number of obs     =        686
Log likelihood = -667.42897
```

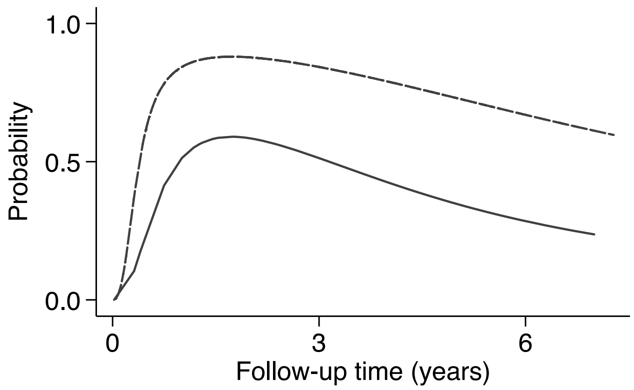
	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x4a	5.082306	1.795635	4.60	0.000	2.542856	10.1578
_eq1_cp2_rcs1	1.415463	.1732414	2.84	0.005	1.113572	1.799197
_eq1_cp2_rcs2	2.369021	.3778431	5.41	0.000	1.733037	3.238395
_cons	.7311249	.2436484	-0.94	0.347	.3804761	1.404933

```
Note: _cons estimates baseline odds.
```

The odds are 5.1 times greater in the larger tumor grade group.

Predicted event probabilities

```
. predict predicted, probability
. gen years = rectime/365.24
. tw line predict years if x4a==0, sort || line predict years if x4a==1, sort
```



Probability-power models

Let x denote a covariate.
We consider the probability-power model

$$\bar{g}(t|x) = \bar{g}_0(t)^{\exp(\beta_1 x)}$$

where $\bar{g}(t) = 1 - g(t)$.

The above can be written as

$$\log\{-\log[\bar{g}(t|x)]\} = \log\{-\log[\bar{g}_0(t)]\} + \beta_1 x$$

The baseline probability function $\bar{g}_0(t)$ can be anything.

The power parameter $\exp(\beta_1)$ is a measure of association.
It corresponds to the hazard ratio per unit-increase in x.

Flexible probability-power model

We estimate a flexible probability-power model

```
. stpreg x4a, power df(2) nolog
Event-probability regression                    Number of obs     =        686
Log likelihood = -668.30844
```

	Power param.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
x4a	2.584105	.6285316	3.90	0.000	1.604252	4.162439
_eq1_cp2_rcs1	1.207611	.0890262	2.56	0.011	1.045143	1.395335
_eq1_cp2_rcs2	1.692367	.1611357	5.53	0.000	1.404264	2.039577
_cons	.5631365	.1349343	-2.40	0.017	.3520915	.9006827

The power parameter (hazard ratio) is 2.6.

Semi-parametric probability-power model

We estimate a semi-parametric probability-power model

```
. stcox x4a, nolog noshow
Cox regression -- Breslow method for ties
No. of subjects =          686                 Number of obs    =        686
No. of failures =          299
Time at risk    =  211.2035922
                                               LR chi2(1)       =      19.92
Log likelihood  =   -1778.2134                 Prob > chi2      =     0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
x4a	2.566048	.6241802	3.87	0.000	1.592993	4.133481

The power parameter (hazard ratio) is 2.6.

The probability and the hazard function

The probability and the hazard functions are (Bottai, 2017)

$$g(t) = 1 - \lim_{\delta \to 0} P(T > t + \delta \mid T > t)^{\frac{1}{\delta}} = 1 - \lim_{\delta \to 0} \left[\frac{S(t+\delta)}{S(t)} \right]^{\frac{1}{\delta}}$$

$$h(t) = \lim_{\delta \to 0} P(T \leq t + \delta \mid T > t) \frac{1}{\delta} = \lim_{\delta \to 0} \left[1 - \frac{S(t+\delta)}{S(t)} \right] \frac{1}{\delta}$$

It can be shown that (Bottai, 2017)

$$g(t) = 1 - \exp[-h(t)]$$

The probability is always smaller than the hazard

$$g(t) < h(t)$$

Conclusions

- Hazards are often mistaken for probabilities.
- For example, *"the risk increases by 68% (HR = 1.68)"*.
- This problem is consequential (Sutradhar & Austin, 2018).
- `stpreg` makes modeling probability functions simple.

References

Bottai, M. (2017). A regression method for modelling geometric rates. *Statistical Methods in Medical Research* 26, 2700-2707.

Bottai, M., Discacciati, A. and Santoni, G. (submitted). Modeling the probability of occurrence of events.

Crowther, M. and Lambert, P. (2013). stgenreg: A stata package for general parametric survival analysis. *Journal of Statistical Software* 53, 1-17.

Discacciati, A. and Bottai, M. (2017). Instantaneous geometric rates via generalized linear models. *Stata Journal* 17, 358-371.

Sutradhar, R. and Austin, P. C. (2018). Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology* 28, 54-57.