Lecture 12: Introduction to Survival Analysis

In many biomedical studies, the outcome variable is a survival time, or more generally a time to an event. We will describe some of the standard tools for analyzing survival data.

Most studies of survival last a few years, and at completion many subjects may still be alive. For those individuals, the actual survival time is not known – all we know is how long they survived from their entry in the study. Similarly, certain individuals may drop out from the study or be lost to follow-up. Each of these cases is said to be *censored*, and the recorded time for such individuals is their time until the censoring event.

Example: HPA staining for breast cancer survival

We consider data from a retrospective study of 45 women who had surgery for breast cancer. Tumor cells, surgically removed from each woman, were classified according to the results of staining on a marker taken from the Roman snail, the *Helix pomatia agglutinin* (HPA). The marker binds to cancer cells associated with metastasis to nearby lymph nodes. Upon microscopic examination, the cancer cells stained with HPA are classified as positive, corresponding to a tumor with the potential for metastasis, or negative. It is of interest to determine the relationship of HPA staining and the survival of women with breast cancer.

The survival times in months T_i and staining results ($x_i = 0$ for negative and $x_i = 1$ for positive) for the 45 women are presented in the following table. Also included is a *censoring indicator* d_i. Contrary to the normal definition of an indicator variable, the censoring indicator is zero if the observation is right-censored, and one if the observation is uncensored. So it's really a *non-censoring* indicator! A woman's survival time was right censored if the woman was alive at the end of the study or if the woman died of causes unrelated to breast cancer.

T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d
23	0	1	47	0	1	69	0	1	70	0	0	71	0	0	100	0	0	101	0	0	148	0	1	181	0	1	198	0	0
208	0	0	212	0	0	224	0	0	5	1	1	8	1	1	10	1	1	13	1	1	18	1	1	24	1	1	26	1	1
26	1	1	31	1	1	35	1	1	40	1	1	41	1	1	48	1	1	50	1	1	59	1	1	61	1	1	68	1	1
71	1	1	76	1	0	105	1	0	107	1	0	109	1	0	113	1	1	116	1	0	118	1	1	143	1	1			
154	1	0	162	1	0	188	1	0	212	1	0	217	1	0	225	1	0												

This is the format the data should be in to work with it in **Stata**, but succinctly, the *sorted* survival times for the negative stained women are

$$23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*,$$

where $*$ denotes a right-censored observation. The survival times for the positive stained group are

$$5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50, 59, 61, 68, 71, 76^*, 105^*,$$

$$107^*, 109^*, 113, 116^*, 118, 143, 154^*, 162^*, 188^*, 212^*, 217^*, 225^*.$$

In the breast cancer study, 8 individuals in the negative stained group, and 11 in the positive stained group are censored. Although it is common for studies to have *right-censored* cases, such as we have here, left-censoring and interval-censoring are found in other clinical studies.

Survival Curves

A first step in survival analysis is often to estimate the *survival curve*, or *survival time distribution*. Suppose we are considering a single (homogeneous) population. Let T be the survival time (from some reference point) for a randomly selected individual from the population. Where t is any arbitrary positive value, the survival time distribution is defined to be

$$\begin{aligned} S(t) &= Pr(T \geq t) \\ &= \text{probability randomly selected individual survives at least until time } t \\ &= \text{proportion of population that survives at least until time } t. \end{aligned}$$

The function might look like Figure 1.

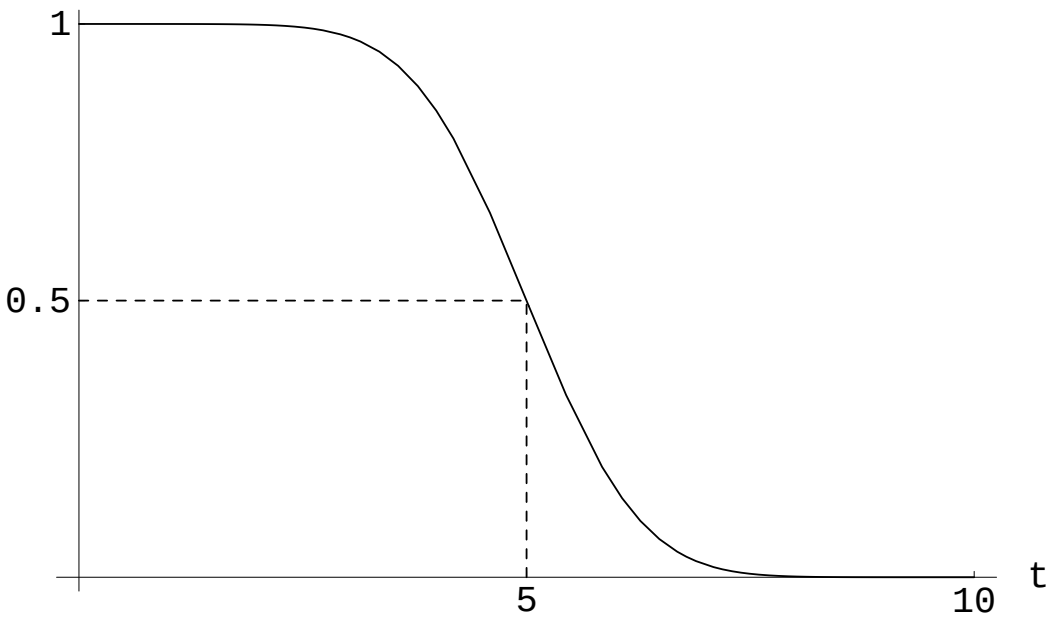

Figure 1: $S(t)$ versus t; median survival time for population is 5.

Estimating the Survival Curve

Case I: No censoring

If we have a random sample from the population, we use the *empirical survival function*:

$$\hat{S}(t) = \text{sample proportion that survive at least until time t}$$

to estimate $S(t)$. This is easy to compute and plot as a function of t.

Suppose we have a sample of 5 survival times (in days): 5, 8, 20, 30, and 33. $\hat{S}(t)$ has "jumps" of size 1/5 (i.e. 1 divided by the sample size) at each survival time; see Figure 2.

Case II: Right censoring

Recall the data on the survival of women with breast cancer whose cells were negatively stained with HPA:

$$23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*,$$

where the $*$ superscript identifies a right-censored observation.

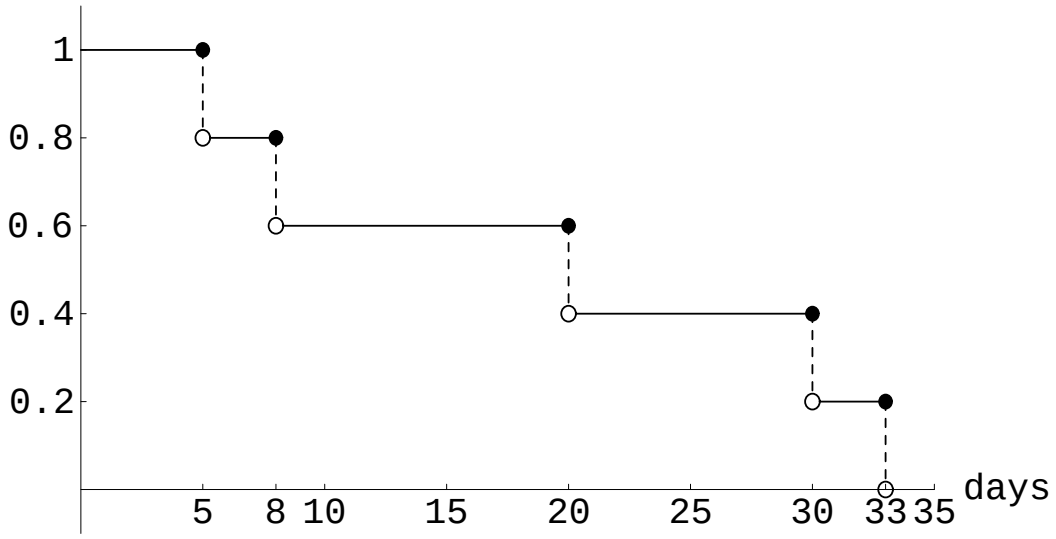

Figure 2: Empirical survival function $\hat{S}(t)$ for the data 5, 8, 20, 30, and 33.

The following algorithm describes the Kaplan-Meier (KM) method for estimating the survival curve (Kaplan-Meier product-limit estimate).

1. Identify times for non-censored cases $0 = t_0 < t_1 < t_2 < \cdots < t_r$. That is, t_1 is the smallest non-censored survival time, t_2 is the second smallest, et cetera. For the example $r = 5$ and $t_0 = 0$, $t_1 = 23$, $t_2 = 47$, $t_3 = 69$, $t_4 = 148$, and $t_5 = 181$.

2. For the j^{th} interval, where $t_{j-1} \leq t < t_j$, evaluate

$$\begin{aligned} n_j &= \text{number at risk (of dying) at beginning of interval,} \\ d_j &= \text{number of deaths in interval,} \\ \frac{n_j - d_j}{n_j} &= \text{estimated probability of surviving past } t_{j-1}, \\ &\quad \text{given you are at risk at time } t_{j-1} \\ &= \hat{P}(T \geq t_{j-1} | T \geq t_{j-2}). \end{aligned}$$

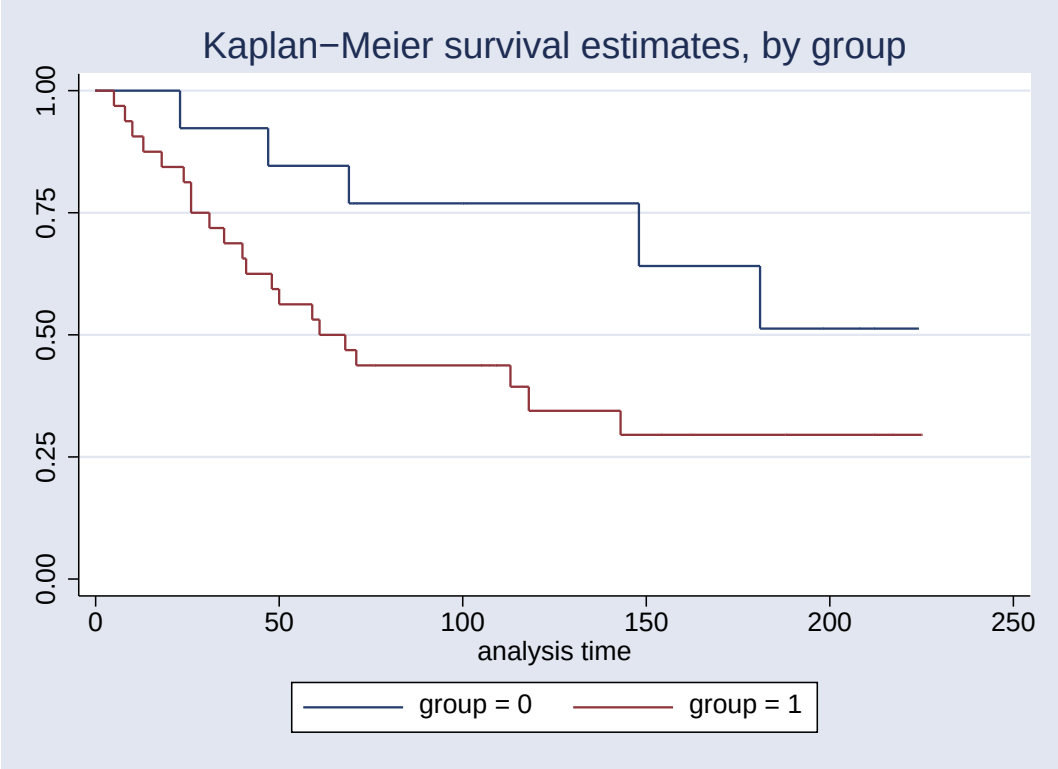

Figure 3: KM survival curves for positively and negatively stained groups.

3. For $t_{j-1} \leq t < t_j$,

$$\begin{aligned}
\hat{S}(t) &= \hat{P}(T \geq t) \\
&= \hat{P}(T \geq t_{j-1} | T \geq t_{j-2}) \times \\
&\quad \hat{P}(T \geq t_{j-2} | T \geq t_{j-3}) \times \cdots \times \\
&\quad \hat{P}(T \geq t_1 | T \geq t_0) \\
&= \frac{n_j - d_j}{n_j} \times \frac{n_{j-1} - d_{j-1}}{n_{j-1}} \times \cdots \times \frac{n_1 - d_1}{n_1}.
\end{aligned}$$

Remark: Censored observations are taken into account by being treated as cases at risk at the beginning of the interval in which they fail.

To illustrate the calculation for our data, consider the table:

j	Interval	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
1	$0 \leq t < 23$	13	0	$\frac{13-0}{13} = 1$	1.0
2	$23 \leq t < 47$	13	1	$\frac{13-1}{13} = \frac{12}{13} \doteq 0.923$	$1.0 \times 0.923 = 0.923$
3	$47 \leq t < 69$	12	1	$\frac{12-1}{12} = \frac{11}{12} \doteq 0.917$	$0.923 \times 0.917 = 0.846$
4	$69 \leq t < 148$	11	1	$\frac{10}{11} \doteq 0.909$	$0.846 \times 0.909 = 0.769$
5	$148 \leq t < 181$	6	1	$\frac{5}{6} \doteq 0.833$	$0.769 \times 0.833 = 0.641$
6	$181 \leq t$	5	1	$\frac{4}{5} = 0.8$	$0.641 \times 0.8 = 0.513$

To obtain the KM estimate in **Stata** we must declare the data we are working with to be survival data. **Stata** then uses the survival time variable and the censoring variable together in analyses. For the breast cancer data we first read in the variables using something like `infile time group cens using c:/breast.txt`. We declare the data to be survival data using `stset time, failure(cens)`. Finally we obtain the KM survival curve estimates across the two groups with the command `sts graph, by(group)`. In Figure 3 we have a picture of $\hat{S}(t)$ from the negatively stained group as well as the estimate from the positively stained group. Note that the negatively stained group tends to live longer, as we would expect. The estimated quartiles for survival across the two groups are obtained by `stsum, by(group)`. Annotated output follows; for example, we see that the median survival in the positive stained group is estimated to be 68 months.

```
          |        no. of    |------ Survival time -----|
group     |        subjects     25%       50%       75%
----------+-------------------------------------------
        0 |        13           148         .         .
        1 |        32            31        68         .
----------+-------------------------------------------
    total |        45            40       113         .
```

Some remarks:

- The estimated survival curve "drops to zero" only if the last case is not censored.
- The KM curve allows us to estimate percentiles of the survival distribution, with a primary interest being the median survival time (50^{th} percentile). In the example

above, the 90^{th} percentile is approximately 47 months (i.e. we estimate that 90% of the population will survive at least 47 months). The median cannot be estimated here – all we can say is that we estimate the median to be at least 181 months.

- The KM estimate is the usual empirical estimate if no cases are censored.
- Statistical methods are available to
  - Estimate the mean survival time.
  - Get a C.I. for the survival curve.
  - Compare survival curves across groups – you can think of this as the censored data analogue of (non-parametric) ANOVA.

The Cox Proportional Hazards Model

The risk of failing at time t is defined to be the probability of an individual dying in the "next instant" (e.g. in a time frame of length Δ) given this individual has survived at least until time t:

$$P(t \leq T < t + \Delta | t \leq T).$$

We define the *hazard function* $h(t)$ such that for small enough Δ,

$$P(t \leq T < t + \Delta | t \leq T) = h(t)\Delta.$$

The hazard function is proportional to the instantaneous "risk of failing" at any time t, given that an individual has lived at least to time t.

Now consider two individuals, 1 and 2, each with their own hazard functions $h_1(t)$ and $h_2(t)$. If we assume that one individual's instantaneous rate of failing is a constant multiple of the other's, i.e. $h_2(t) = ah_1(t)$ for some constant a, then these two individuals have *proportional hazard functions*. Figure 4 shows an example of this phenomenon where the hazard ratio is 1/2.

Proportional hazards may or may not be a reasonable assumption to make. For example, consider two people, roughly the same age and demographic except that at the age of 20,

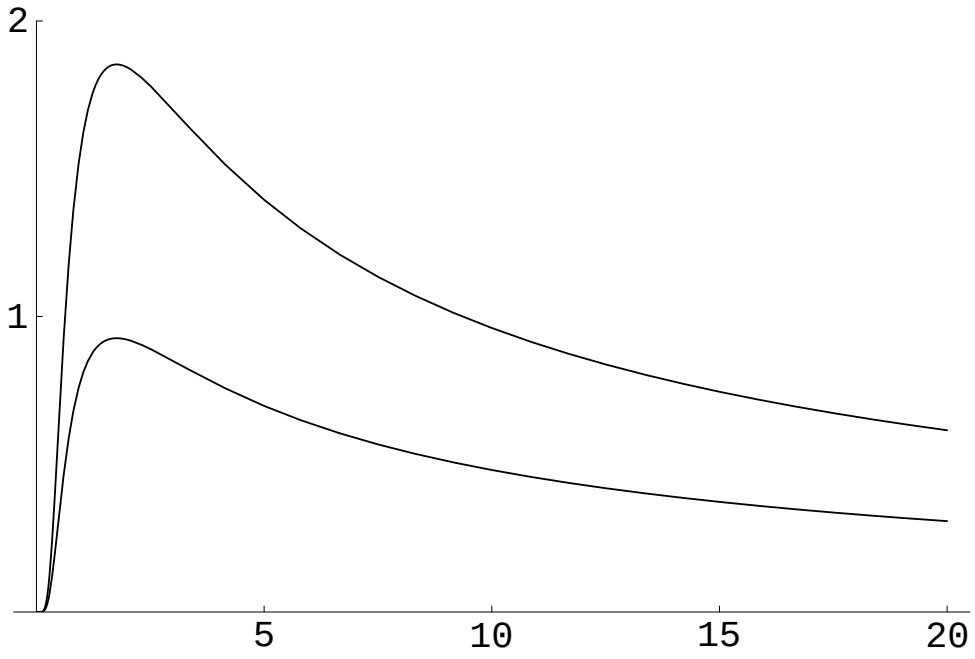

Figure 4: An example of proportional hazard functions; here the constant of proportionality is 0.5.

person 2 takes up smoking while person 1 does not. You will hopefully agree with me that initially, the smoker and the non-smoker will most likely have *identical* hazards. As the years roll by, and smoking takes its toll, we would think that the smoker's instantaneous rate of failing, which is proportional to the probability of dying in the next minute, say, will increase relative to the hazard for the non-smoker. In this example proportional hazards probably is an unreasonable assumption.

The proportional hazards *model* generalizes the above concept for n individuals, each with their own covariate value x_i or set of p covariate values $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. In the case where the n individuals only have one covariate, the model stipulates for individuals i and j, with a hazard functions $h_i(t)$ and $h_j(t)$ respectively, that

$$h_i(t)e^{-\beta x_i} = h_j(t)e^{-\beta x_j}.$$

Note that this implies

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta x_i}}{e^{\beta x_j}} = e^{\beta(x_i - x_j)}.$$

Here, $e^{\beta(x_i - x_j)}$ is the relative risk of instantaneous failure at *any time* t for individuals i and j. That is the power of the proportional hazards assumption: the relative risk of dying for two individuals is a simple function of the model parameters and holds for all t, independent of the value of t. If individual i has covariate value $x + 1$ and individual j has covariate value x, i.e. their covariate values only differ by 1 unit on the covariate measurement scale, then

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta(x+1)}}{e^{\beta x}} = e^{\beta}.$$

Thus, e^{β} is the relative risk of failing in the next instant when we increase the covariate by one unit. Note that if x_i is a simple zero/one variable denoting which group individual i falls into, then e^{β} is the relative risk of failing in the next instant for the group denoted by $x_i = 1$ versus $x_i = 0$.

The breast cancer data are loaded with the commands `infile time group cens using c:/breast.txt` and the Cox PH model is fit via `cox time group, dead(cens)`. The survival time, followed by the predictor variable(s) is specified. The non-censoring indicator is included in the subcommand `dead`. We obtain the following output:

```
------------------------------------------------------------------------------
  time |
  cens |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
 group |   .9080157   .5009228     1.81   0.070    -.0737749    1.889806
------------------------------------------------------------------------------
```

We have an estimate of $\hat{\beta} = 0.908$ and the estimated relative risk is $e^{\hat{\beta}} = e^{0.908} \stackrel{\bullet}{=} 2.5$. That is, those with positive staining are estimated to have a risk of dying in the next instant about 2.5 times as great as those with negative staining. Note that the p-value for $H_0 : \beta = 0$ is small but not significant at the 5% level. There is definitely *some* indication that staining affects survival, with positive staining decreasing survival. A 95% C.I. for the risk may be obtained by exponentiating the endpoints for the C.I. for β. Here, we estimate the relative risk of expiring (for positive compared to negative staining) to be within $(e^{-0.073}, e^{1.89}) = (0.93, 6.62)$ with 95% confidence.

Remark: The hazard function for individual i can be defined to be a scale multiple $e^{x_i\beta}$ of a *baseline* hazard function denoted $h_0(t)$. The model may be recast as

$h_i(t) = h(t|x_i) = e^{x_i\beta}h_0(t)$. This baseline hazard function $h_0(t)$ and β thus completely determine the model. The baseline hazard $h_0(t)$ may be estimated from the data as well as survival curves, median and mean survival, et cetera, for any covariate value x. These sorts of inferences are quite easy to get out of **Stata** but a bit beyond what is comfortable to cover in this class.

A final example

We examine a data set consisting of the time spent running on a treadmill for 14 people aged 15 and older. Each subject's gender and age were recorded. It is of interest to the experimenter how age and gender affects ones endurance.

We define a numeric indicator variable for the gender variable by taking g to be 0 for a male subject and 1 for a female subject. When fitting the PH model with gender and age as main effects,

$$h(t|age, g) = e^{age\beta_1+g\beta_2}h_0(t),$$

the baseline group (i.e. those with covariates $age = 0$ and $g = 0$, and thus a hazard function of $e^{0\beta_1+0\beta_2} = e^0h_0(t) = h_0(t)$) consists of males of age zero, which is not interpretable in this context. Observations were censored due to a subject having to leave the treadmill for reasons other than being tired. The data follow:

Obs	gender	age	minutes	cens	weight	g
1	male	34	16	1	215	0
2	male	15	35	0	135	0
3	female	22	55	0	145	1
4	female	18	95	1	97	1
5	male	18	55	0	225	0
6	female	32	55	1	185	1
7	female	37	25	1	155	1
8	female	67	15	1	142	1
9	female	55	22	1	132	1
10	male	55	13	1	183	0
11	male	62	13	1	168	0
12	female	33	57	0	132	1
13	female	17	52	0	112	1
14	male	24	54	1	175	0

The fit of the model with only gender $h(t|g) = e^{g\beta_1}h_0(t)$:

minutes	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
g	-.6786811	.7161483	-0.95	0.343	-2.082306	.7249439

The test for a gender effect yields a p-value of 0.343. We would accept at any reasonable significance level that there is not a gender effect. The estimate of β_1 is $\hat{\beta}_1 = -0.679$ so the fitted model is $h(t|g) = e^{-0.678g}h_0(t)$ implying that $h(t|g = 1) = 0.507h(t|g = 0)$ and finally $h(t|g = 1)/h(t|g = 0) = 0.507$ for all t. That is, the probability of a randomly picked woman failing (stepping off the treadmill) in the next second is estimated be half the probability of a randomly picked male.

Rephrased, we see that, assuming proportional hazards is reasonable, females are about half (the hazard ratio is $e^{-0.679} = 0.507$) as likely to step off the treadmill at any instant versus males. We obtain an approximate 95% C.I. for this ratio by first considering the 95% C.I. for the regression effect: (-2.08, 0.72). Exponentiate both endpoints to obtain a 95% C.I. for the hazard ratio: (0.12, 2.07). The hazard ratio interval includes one (no difference in the hazard functions for males and females) because the regression effect interval includes zero.

Let's look at the model fit with only age $h(t|age) = e^{age\beta_1}h_0(t)$:

minutes	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.1116606	.0385688	2.90	0.004	.0360672	.187254

A year from now, a randomly selected individual will be $e^{0.1117} = 1.118$ times as likely to step off the treadmill after 15 minutes (or any amount of time) than now. In ten years it will be $1.118^{10} = 3.05$ times as likely. When we fit the model with both of these predictors $h(t|age, g) = e^{age\beta_1+g\beta_2}h_0(t) = e^{age\beta_1}e^{g\beta_2}h_0(t)$ we see that estimated regression effects, and therefore model interpretation, change somewhat:

minutes	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
g	-3.551859	1.57856	-2.25	0.024	-6.645779	-.4579388
age	.2186267	.0855601	2.56	0.011	.050932	.3863214

At a given age, a random male running alongside a random female is about $1/e^{-3.55} = 1/0.029 = 35$ times as likely to step off the treadmill at any time. A woman 20 years older than another woman is about $e^{0.218\times 20} = 80$ times as likely to step off compared to

the younger woman. Note that in the presence of age, gender is now significant, although marginally, gender is not a significant factor. In this case age is said to be a *suppressor* variable. The **Stata** commands for this analysis are:

```
infile age minutes cens weight g1 using c:/running.txt
cox minutes g1, dead(cens)
cox minutes age, dead(cens)
cox minutes g1 age, dead(cens)
```

In the model fit that included an interaction between age and gender, the interaction term was not significant.