BST 210 Lab: Week 12-13
Survival Analysis

Introduction to Survival

In some studies, the outcome of interest is the time to event (survival time to failure). We can perform survival analysis to estimate the distribution of survival times, test equality of two or more survival distributions, or estimate the relationships between predictors and the survival time while control for other covariates.

Let T_i be the time to event for patient i. There exist four interrelated functions which categorize the distribution of T_i:

1. **Survival function:**

$$S(t) = P(T_i > t) = 1 - P(T_i \leq t) = 1 - F(t)$$

   This is the probability of not having event until time t. Note $S(0) = 1$, and $S(t) \in [0, 1], t \in [0, \infty)$.

2. **Cumulative distribution function:** $F(t) = P(T_i \leq t) = 1 - S(t)$.

3. **Probability density function:** $f(t) = F'(t) = -S'(t)$.

4. **Hazard function:**

$$h(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T_i \leq t + \triangle t | T_i \geq t)}{\triangle t} = \frac{f(t)}{S(t)}$$

   This instantaneous rate of failure at time t, given that a subject as survived to time t. Note that $h(t) \geq 0$ but does not have a upper bound because it is not a probability. [1]

Censoring

Theoretically, there is no reason why we cannot use analyses/tests/regression methods we learned previously in this class for time to event outcomes, since it is a continuous outcome and t. The unique feature of survival analysis (and why there is a field dedicated to this type of data) is how it handles censorship.

A patient is **censored** if we do not observe the time of their outcome. This could be due to many different factors but is essentially a missing data problem (we cannot observe T). Unlike most missing data problems, however, survival data is unique in that the analyses we have developed can incorporate information from censored patients. Let's denote the censoring time as C.

However, survival analysis is dependent on the assumption of **noninformative censoring**, which says that the event of censoring must be independent of the time of event.

Question: What situations would the noninformative censoring assumption be invalid?

Any case where censoring is associated with survival, for example people who have symptoms of a disease are more likely to drop out of the study.

[1] Hernán, Miguel A. "The hazards of hazard ratios." Epidemiology (Cambridge, Mass.) 21.1 (2010): 13.

Censoring comes in three flavors:

1. **Left censoring:** A patient is not observed before a certain time point, and may have experienced the event already, (i.e. $T < C$). This time point is usually their recruitment into the study.

2. **Interval censoring:** A patient experiences an event within an unobserved interval, but the exact time within the interval is unknown. Most survival data suffers from interval censoring, simply because it is difficult to collect information on the exact time an event occurs. The severity of interval censoring depends on the frequency of follow-up and effort made to find supplemental information (death certificates etc.) Commonly, statisticians can assume events all occur right after the last follow-up, right before the current follow-up or in the middle. This does not change the outcome of the analysis with regards to hypothesis testing (will change the outcome with regards to estimated survival time, especially if the intervals are large), as long as the assumptions are consistent.

3. **Right censoring:** A patient is not observed experiencing the event, but may have the event after their observation period is over (i.e. $T > C$). Right censoring is the most common "problem" in survival analysis because it cannot be designed or assumed away and must be dealt with through analysis. Common causes of right censoring include dropout, study conclusion and death.

Nonparametric survival analysis: Kaplan-Meier estimator

The **Kaplan-Meier** estimator (aka **product-limit** esimator) is a **non-parametric** estimator for survival curves. Using the K-M estimator, we do not have to make any distributional assumptions about survival times. This method is intuitive, but has limitations in that we cannot account for covariate data.

Assume that in your data, there were m distinct times when events occurred, which we call $t_1 < t_2 < \cdots < t_m$. Then we can go back and look at each time t_j to see the number of events that occurred d_j, as well as the total number of people present in the data at that time n_j. We say these people are in the 'risk set' at time t_j. Don't forget, the people who experienced events are still included in n_j, because they were at risk when they experienced the event!

Then, given that someone is in the risk set at time t_j, their probability of experiencing the event at that time is simply $\hat{p}_j = \frac{d_j}{n_j}$, which in turn means that given a person remained in the risk set until time t_j, their probability of survival beyond time t_j is $\hat{q}_j = 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j}$.

The Kaplan-Meier estimator simply multiplies these running conditional survival probabilities together up through each event time. The empirical probabilities at each time t follow the formula

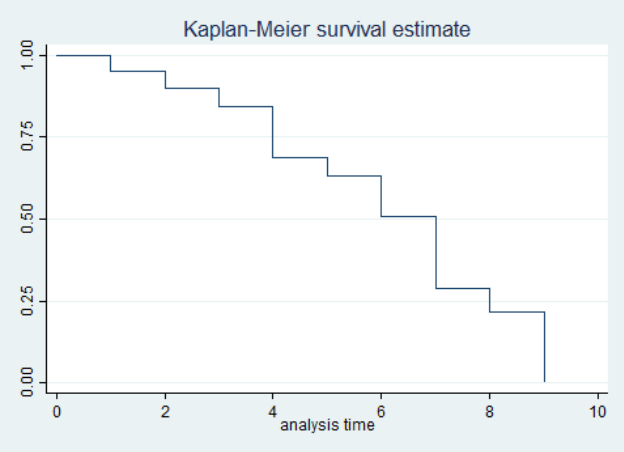
$$\hat{S}(t) = \prod_{j=1}^{k} \hat{q}_j = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}, \quad t_k \leq t < t_{k+1}$$

where $\hat{S}(0) = 1$.

Let's do an example: let the survival times of 20 patients be: 1, 1+, 2, 3, 4, 4, 4, 5, 5+, 5+, 6, 6, 6+, 7, 7, 7, 8, 8+, 8+, 9. Where '+' indicates censoring (if an observation is designated '1+' it means that they had their event after time period 1). Fill out the following table and graph this curve by hand.

k	**Risk Set n_k**	**Events d_k**	**Conditional Survival $\hat{q}_k = 1 - \frac{d_k}{n_k}$**	**Unconditional Survival $\hat{S}(t_k) = \prod_{j=1}^{k} \hat{q}_i$**
1	20	1	19/20	(19/20) = 0.95
2	18	1	17/18	(19/20)(17/18) = 0.897
3	17	1	16/17	(19/20)(17/18)(16/17) = 0.844
4	16	3	13/16	(19/20)...(13/16)=0.686
5	13	1	12/13	(19/20)...(12/13)=0.633
6	10	2	8/10	(19/20)...(8/10)=0.507
7	7	3	4/7	(19/20)...(4/7)=0.290
8	4	1	3/4	(19/20)...(3/4)=0.217
9	1	1	0	0

Draw a survival curve based on this table:



Survival Analysis in Stata

The Sorbinil Retinopathy Trial (SRT) was a randomized, double-blind, placebo controlled trial testing whether the aldose reductase inhibitor, sorbinil, would reduce rates of progression in complications of diabetes, with a primary focus on diabetic retinopathy (Arch Ophthalmol 1990; 108:1234-1244). Between August 1983 and October 1986, the SRT randomized 497 type I diabetic patients at 11 clinical centers. The primary outcome of the trial was worsening of a patient's retinopathy by two or more steps, based on grading fundus photographs. The file srt.dat contains data from 478 randomized participants with some follow-up and complete information on important baseline covariates.

Variables on the file srt.dat are: id (a subject id), sorb (randomized treatment assignment, 1=sorbinil, 0=placebo), tgh (total glycosylated hemoglobin in percent), dur (duration of diabetes in years since diagnosis), sex (2=female, 1=male), fup (duration of follow-up in years until progression of diabetic retinopathy or end of follow-up), and status (1=diabetic retinopathy progressed, 0=no progression).

Read the data into Stata, prepare them for survival analysis, and compare the distributions of patient characteristics by treatment group. Note that women were less likely to be included because sorbinil was teratogenic.

```
import delimited srt.csv

*Summarize by treatment group
sort sorb
by sorb: su
------------------------------------------------------------------------------
-> sorb = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |        241    3.46e+08    9.58e+07   1.80e+08   6.13e+08
        sorb |        241           0           0          0          0
         tgh |        241    11.81743    1.920966       8.35       18.1
         dur |        241    6.875259    3.676774          1         16
         sex |        241    1.261411    .4403177          1          2
-------------+---------------------------------------------------------
         fup |        241    2.974056    .8576834       .9747     4.4983
      status |        241    .2738589    .4468655          0          1

------------------------------------------------------------------------------
-> sorb = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |        237    3.51e+08    1.00e+08   1.80e+08   6.13e+08
        sorb |        237           1           0          1          1
         tgh |        237    11.93734    2.063994        8.2      23.05
         dur |        237    6.797732    3.346496      1.375    15.9375
         sex |        237    1.240506    .4282955          1          2
-------------+---------------------------------------------------------
         fup |        237    2.932503    .9030455       .9692     4.5366
      status |        237    .2700422     .444921          0          1
```

Note that we need to create a new variable to denote which observations are censored:

```
*Set up survival analysis in stata
stset fup, id(id) failure(status)
```

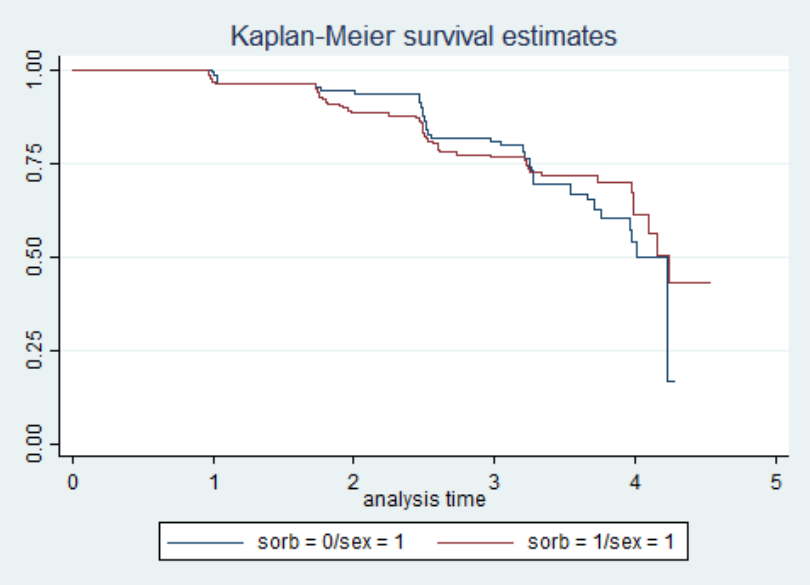
We can get compute the $\hat{S}(t)$, its standard error (Greenwood's method) and its confidence interval (complementary log-log transformation). If you use Greenwood's method to compute the confidence interval, make sure you truncate the results so that the range is within $[0, 1]$.

Let's first get the K-M estimates for the males in Stata:

```
. sts list if sex==1, by(sorb sex)

         failure _d:  status
   analysis time _t:  fup
                 id:  id

           Beg.          Net            Survivor      Std.
  Time    Total   Fail   Lost           Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
sorb=0 sex=1
 .9747      178      0      1             1.0000         .          .         .
 .9938      177      1      0             0.9944    0.0056     0.9606    0.9992
. . .
  4.23        2      1      0             0.1659    0.1375     0.0118    0.4833
  4.29        1      0      1             0.1659    0.1375     0.0118    0.4833
sorb=1 sex=1
 .9692      180      1      0             0.9944    0.0055     0.9612    0.9992
 .9719      179      1      0             0.9889    0.0078     0.9563    0.9972
. . .
 4.501        2      0      1             0.4321    0.0985     0.2395    0.6110
 4.537        1      0      1             0.4321    0.0985     0.2395    0.6110
-------------------------------------------------------------------------------
```

And we can plot the K-M curve for sorbinil and placebo groups separately (only for males):

```
*Plot Kaplan-Meier estimate
sts graph if sex==1, by(sorb sex)
```



Question: What conclusion can you draw from these two survival curves?

The control group has better survival until time 3, when the treatment group has a small gain before the two curves cross again at the end of the study. Additionally, the two treatments are very similar and cross at multiple points, leading us to conclude that the they likely do not result in different survival probabilities over time.

Testing differences: log-rank test

The **log-rank test** is a non-parametric test of the difference in survival curves:

$$H_0 : S_1(t) = S_2(t) \text{ for all } t$$
$$H_1 : S_1(t) \neq S_2(t) \text{ for at least one value of } t$$

It is a generalization of the Mantel-Haenszel test and requires the following steps:

1. Create a 2x2 table for each observed failure time $t_k, k = 1, 2, ..., m$, where m is the number of distinct failure times.
2. At each time point t_k, note the number of events (d_{jk}) and patients in the risk set (n_{jk}) for each group $j = 1, 2$. Let d_k be the total number of events at t_k and n_k be the total number of patients in the risk set.
3. Calculate $O = \sum_{k=1}^{m} d_{1k}$, $E = \sum_{k=1}^{m} \frac{n_{1k} d_k}{n_k}$ and $V = \sum_{k=1}^{m} \frac{n_{1k} n_{2k} d_k (n_k - d_k)}{n_k^2}$.
4. Compare $\frac{(O-E)^2}{V}$ to a chi-square distribution with 1 degree of freedom (χ_1^2).

The restrictions of this test are that V need to be larger than 5. Also, the test is the most powerful when the proportional hazards assumption holds (i.e. the survival/hazard curves of the two groups do not cross, with $h_2(t)/h_1(t) = c$, c is some constant). If this assumption is violated, the test is still valid but became underpowered.

We will perform an log-rank test on our Stata dataset as an example, but keep in mind that the survival curves we are comparing do cross.

```
. sts test sorb if sex==1

         failure _d:  status
   analysis time _t:  fup
                 id:  id


Log-rank test for equality of survivor functions

      |   Events         Events
sorb  |  observed       expected
------+-------------------------
0     |        50          47.49
1     |        49          51.51
------+-------------------------
Total |        99          99.00

            chi2(1) =       0.26
            Pr>chi2 =     0.6108
```

Also, you can perform stratified log-rank tests to adjust for sex:

```
. sts test sorb, strata(sex)

         failure _d:  status
   analysis time _t:  fup
                 id:  id


Stratified log-rank test for equality of survivor functions

      |   Events         Events
sorb  |  observed    expected(*)
------+-------------------------
0     |        66          64.32
1     |        64          65.68
------+-------------------------
Total |       130         130.00

(*) sum over calculations within sex

            chi2(1) =       0.09
            Pr>chi2 =     0.7672
```

Parametric survival analysis

Kaplan-Meier survival curves and Log-Rank test compares the survival data in two groups using completely non-parametric methods. But what if we wanted to measure the change in survival across a continuous covariate? Or across multiple covariates at once?

Exponential survival regression model

Let $X_{i1}, ..., X_{ip}$ denote p covariates for patient i. If we assume that $T_i|X_{i1}, ..., X_{ip}$ follows a exponential distribution with parameter λ:

$$f(t) = \lambda \exp(-\lambda t)$$

The corresponding hazard function is:

$$h(t) = \lambda$$

which is independent of t. Then we can build a regression model:

$$h(t|X_1, ..., X_p) = \exp\{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p\}$$

We define the baseline hazard function as hazard function when all covariates are 0. In exponential survival regression, $h_0 = h_0(t) = h(t|X_1 = 0, ..., X_p = 0) = exp(\beta_0)$, then the model can be rewritten as:

$$h(t|X_1, ..., X_p) = h_0 \exp\{\beta_1 X_1 + ... + \beta_p X_p\}$$

or equivalently,

$$\log\left\{\frac{h(t|X_1, ..., X_p)}{h_0}\right\} = \beta_1 X_1 + ... + \beta_p X_p$$

Question: Can you find any connection between this model and a Poisson regression model?

The setting was quite similar: when we fit a Poisson model, we were building a model for the number of events Y based on an event rate λ and a followup time t, both of which we also have here in some form: in the Exponential model, we also have a rate λ, and followup time T.

The main difference is how we conceive of our model: in Poisson regression we are estimating the number of events that occur over a fixed amount of time, whereas here we are estimating the amount of time until an event occurs. Notice that both are rooted in understanding the *rate* at which the events occur, and in fact if we wrote out the likelihood functions for each setting, letting Y be the total number of events we observed in our Exponential model, they would only differ by contants unrelated to λ, so in fact our maximum likelihood estimation of βs for covariates associated with these rates would be the same!

This is nuanced, and will not be tested on, but is interesting to realize that there is a relationship between understanding 'time until an event' and 'number of events in a time interval'.

Let's fit a exponential survival regression model in Stata:

```
. streg i.sorb i.sex tgh dur, distribution(exponential)

        failure _d:  status
  analysis time _t:  fup
                id:  id

Iteration 0:   log likelihood = -326.78183
Iteration 1:   log likelihood = -312.36888
Iteration 2:   log likelihood = -309.78515
Iteration 3:   log likelihood = -309.77591
Iteration 4:   log likelihood = -309.77591

Exponential regression -- log relative-hazard form

No. of subjects =          478                 Number of obs    =         478
No. of failures =          130
Time at risk    =  1411.750697
                                               LR chi2(4)       =       34.01
Log likelihood  =   -309.77591                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.sorb |   .9624412   .1693349    -0.22   0.828      .681728    1.358743
       2.sex |   .7256609   .1520984    -1.53   0.126      .481196    1.094323
         tgh |   1.237219   .0457712     5.75   0.000     1.150685    1.330262
         dur |   1.083351   .0275497     3.15   0.002     1.030678    1.138715
       _cons |   .0043087   .0023062   -10.18   0.000     .0015092     .012301
------------------------------------------------------------------------------
```

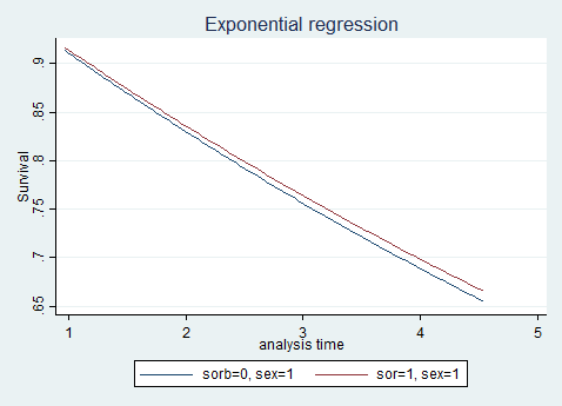
Question: Interpret the intercept and the coefficient for `sorb`.

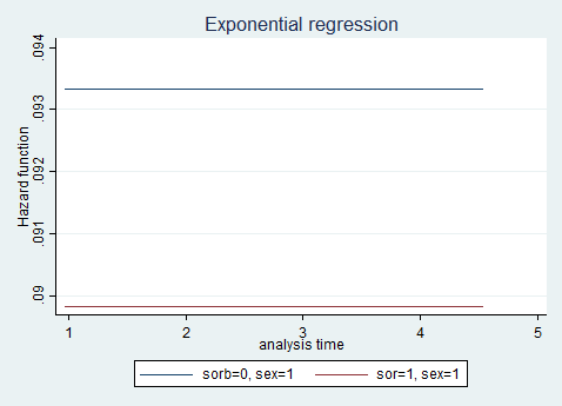
$h_0 = \exp(\beta_0)$: The baseline hazard rate is 0.0043 among the non-diabetic males who don't have any total glycosylated hemoglobin. Note this combination of covariate values is not realistic!

$\exp(\beta_{sorb})$: Within the population under study, patients who received sorbinil experienced 0.96 times (or, 4% lower) the hazard rate of those who received placebo for experiencing a worsening of retinopathy by two or more steps after controlling for sex, total glycosylated hemoglobin, and duration of diabetes in years since diagnosis.

Plot hazard function:

```
stcurve, survival at1(sorb=0, sex=1) at2(sor=1, sex=1)
stcurve, hazard at1(sorb=0, sex=1) at2(sor=1, sex=1)
```





Question: If you adjust for multiple covariates in your regression, you need to specify the values for those covariates when you plot the survival/hazard/cumulative probability functions. What values would you use?

There are many reasonable ways. For example, Stata generates plots at the mean value of covariates. However, taking the mean is not meaningful for categorical data, so we can instead specify the category (e.g. baseline group). You can also use the mode.

Weibull survival regression model

In exponential survival regression, we assumed the hazard is constant over time, which is hardly true in real life. Weibull survival regression model is one of the more flexible model where we assume T follows a Weibull distribution and the hazard function varies over time. The Weibull model can be written as:

$$h(t|X_1, ..., X_p) = h_0(t) \exp(\beta_1 X_1 + ... + \beta_p X_p)$$

where

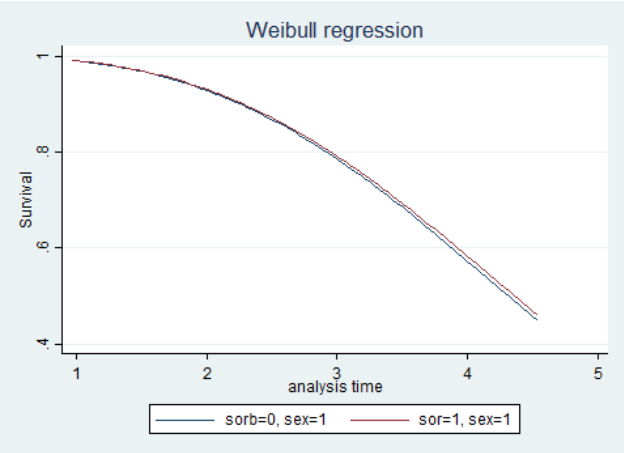
$$h_0(t) = \exp(\beta_0)\gamma t^{\gamma-1} = \lambda\gamma t^{\gamma-1}$$

λ is called the scale parameter, and γ is called the shape parameter. Specifically, the baseline hazard increases over time when $\gamma > 1$, decreases when $0 < \gamma < 1$, and remain constant when $\gamma = 1$ (the model reduces to a exponential survival regression model).

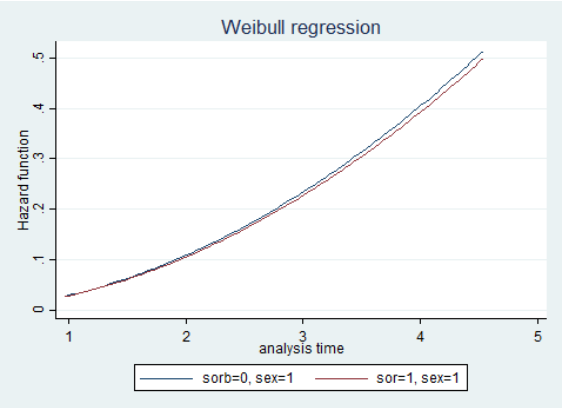
```
. streg i.sorb i.sex tgh dur, distribution(weibull)

         failure _d:  status
   analysis time _t:  fup
                 id:  id

Fitting constant-only model:

Iteration 0:   log likelihood = -326.78183
Iteration 1:   log likelihood = -275.67984
Iteration 2:   log likelihood = -267.73983
Iteration 3:   log likelihood = -267.61483
Iteration 4:   log likelihood = -267.61481
Iteration 5:   log likelihood = -267.61481

Fitting full model:
```

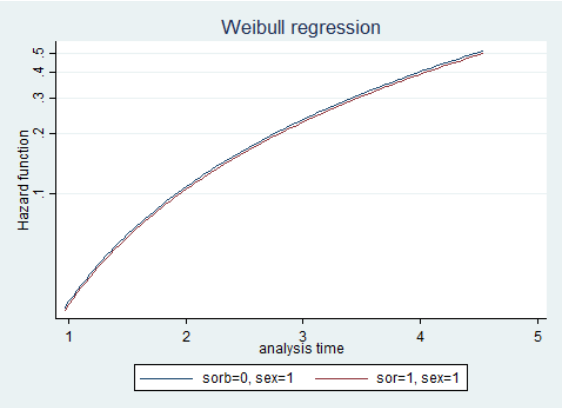
```
Iteration 0:   log likelihood = -267.61481
Iteration 1:   log likelihood = -248.51025
Iteration 2:   log likelihood = -243.26574
Iteration 3:   log likelihood = -243.21947
Iteration 4:   log likelihood = -243.21945
Iteration 5:   log likelihood = -243.21945

Weibull regression -- log relative-hazard form

No. of subjects =          478                 Number of obs    =         478
No. of failures =          130
Time at risk    =  1411.750697
                                               LR chi2(4)       =       48.79
Log likelihood  =   -243.21945                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.sorb |    .970657   .1705529    -0.17   0.865     .6878634    1.369712
       2.sex |   .6384986    .134474    -2.13   0.033     .4225603    .9647865
         tgh |    1.31143   .0506003     7.03   0.000     1.215913    1.414451
         dur |   1.096829   .0295709     3.43   0.001     1.040376    1.156346
       _cons |   .0002114   .0001371   -13.05   0.000     .0000593    .0007535
-------------+----------------------------------------------------------------
       /ln_p |   1.066101    .074717    14.27   0.000     .9196587    1.212544
-------------+----------------------------------------------------------------
           p |   2.904036   .2169808                      2.508434    3.362026
         1/p |   .3443484   .0257287                      .2974397    .3986551
------------------------------------------------------------------------------
```

```
stcurve, survival at1(sorb=0, sex=1) at2(sor=1, sex=1)
```



```
stcurve, hazard at1(sorb=0, sex=1) at2(sor=1, sex=1)
stcurve, hazard at1(sorb=0, sex=1) at2(sor=1, sex=1) yscale(log)
```





Both exponential and Weibull models are parametric because $h_0(t)$ has a specified function. Also, you can choose other distributions to build regression models (e.g. gamma distribution).

Semi-parametric survival analysis: Cox regression model

The **Cox regression model** (aka Cox proportional hazard model) is a semi-parametric regression model for survival data. It is semi-parametric in that it makes parametric assumptions about a linear combination of covariates, but the baseline hazard is not required to be a known distribution:

$$h(t|X_1, ..., X_p) = h_0(t) \exp(\beta_1 X_1 + ...\beta_p X_p)$$

To avoid estimating $h_0(t)$, the βs are computed by maximizing the partial likelihood. When there are tied events (more than 1 events happen at an observed failure time t_k), the computation of partial likelihood becomes tricky. There are a few different ways to handle ties, such as exact partial likelihood, Efron method and Breslow method. Among these three, exact partial likelihood is the most accurate but computationally expensive, while Breslow's method tend to be the least accurate but computationally efficient.

Let's fit a full Cox regression with our variables of interest in Stata.

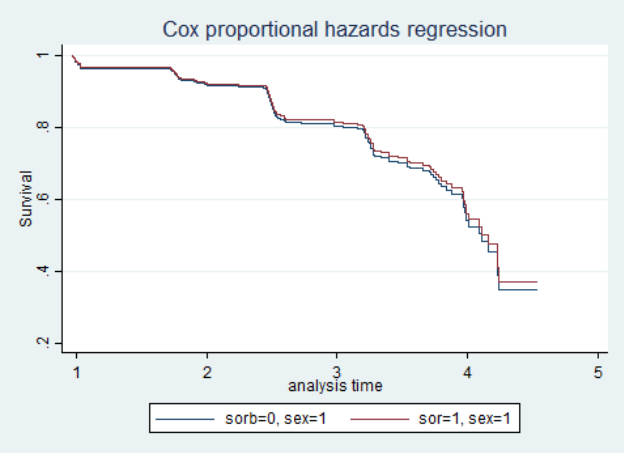
```
*Full Cox Model
stcox i.sorb i.sex tgh dur
estimates store fullmodel

failure _d:  status
analysis time _t:  fup
id:  id

Iteration 0:   log likelihood = -710.06399
Iteration 1:   log likelihood =  -687.1304
Iteration 2:   log likelihood = -685.80968
Iteration 3:   log likelihood =   -685.803
Iteration 4:   log likelihood =   -685.803
Refining estimates:
Iteration 0:   log likelihood =   -685.803
```

```
Cox regression -- Breslow method for ties

No. of subjects =          478                 Number of obs    =         478
No. of failures =          130
Time at risk    =  1411.750697
                                               LR chi2(4)       =       48.52
Log likelihood  =     -685.803                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.sorb |   .9436001   .1662714    -0.33   0.742     .6680332     1.33284
       2.sex |   .6489158   .1361722    -2.06   0.039     .4300979      .97906
         tgh |     1.3102   .0504697     7.01   0.000     1.214923    1.412949
         dur |   1.097906   .0300962     3.41   0.001     1.040476    1.158507
------------------------------------------------------------------------------
```

Question: Interpret the coefficient for sex.

Within the population under study, women experienced 0.65 times (or, 35% lower) the hazard rate of men for experiencing a worsening of retinopathy by two or more steps after controlling for treatment taken. Note that you don't have estimates for baseline hazard any more.

```
stcurve, survival at1(sorb=0, sex=1) at2(sor=1, sex=1)
```

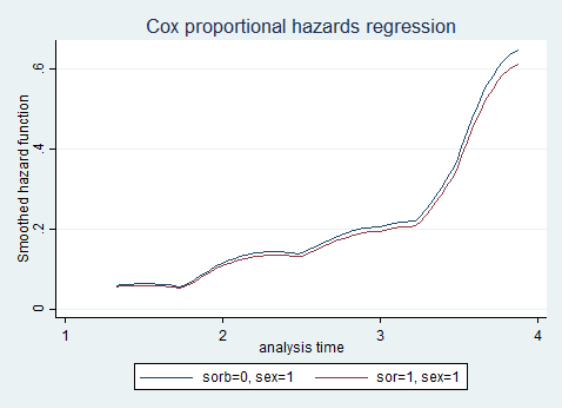


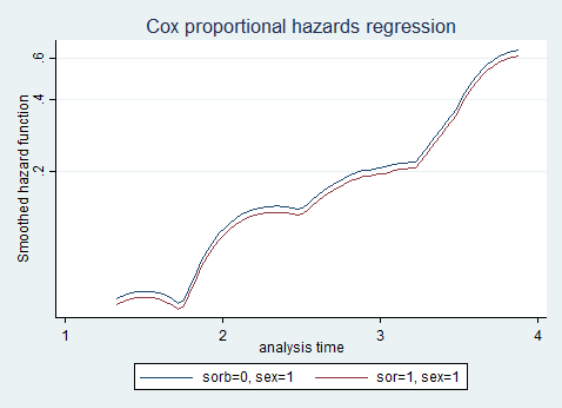
Question: Can you graph the hazard functions for males in each of two treatment groups soly based on the Cox regression outputs shown above? Why? Can you do that for exponential or Weibull survival regression?

We cannot directly graph the hazard functions for Cox model based on the outputs shown above, because we don't know the baseline hazard function. Note that we can anyways generate hazard function plots in Stata because it took additional steps (not part of the Cox regression itself) to estimate the baseline hazard function.

We can do that for exponential and Weibull (or any other parametric models) because we know the specific form of baseline hazard function.

```
stcurve, hazard at1(sorb=0, sex=1) at2(sor=1, sex=1)
stcurve, hazard at1(sorb=0, sex=1) at2(sor=1, sex=1) yscale(log)
```





All of exponential, Weibull and Cox regression models assumes proportional hazards (hazard ratio does not change over time):

$$\frac{h(t|X_1 = x_1, ..., X_i = x_i + 1, ...X_p = x_p)}{h(t|X_1 = x_1, ..., X_i = x_i, ...X_p = x_p)} = \exp(\beta_i) \qquad \text{(constant over t)}$$

However, this assumption does not always hold, so we will need to perform formal statistical tests – we will talk about such tests later.