Precept 6: Duration models

Soc 504: Advanced Social Statistics

Ian Lundberg

Princeton University

March 16, 2018

Outline

We've gotten some questions about our personal project workflows.

Rmarkdown is in some ways ideal:

- Fully reproducible
- Code and results in one place

Problem: If code is slow to run, Rmarkdown is slow to compile each time.

I more often use **R** and LaTeX:

- In RStudio, you can create a new R script. This is your code but does not produce a PDF.
- Save results (see ?save, ggsave, etc.)
- Produce final report in LaTeX
  - I use TexShop
  - You can also work in an online platform like Overleaf. They also provide great templates!

Outline

Duration models are useful when we are interested in the

time T to an event

but some observations are **censored**: the event has not occurred at the end of data collection

Think, pair, share:
Why can't we do OLS when some observations are censored?
Because for those observations we don't know T!

The time to death T is a random variable.
Its distribution is described by four critical functions:

1. **Density function** $f(t)$
   - Density of death at t
2. **CDF** $F(t) = P(T < t)$
   - Probability of death by t
3. **Survival function** $S(t) = P(T > t) = 1 - F(t)$
   - Probability of survival to t
4. **Hazard function** $h(t) = \frac{f(t)}{S(t)}$
   - Density of death at t given survival up to t

Question: Why isn't the hazard function a probability?

Photo credit: J Zamudio via
https://www.nps.gov/yose/planyourvisit/stargazing.htm

Exponential distribution $T \sim \text{Exponential}(\lambda)$

PDF

$$f(t) = \lambda e^{-\lambda t}$$

CDF

$$F(t) = 1 - e^{-\lambda t}$$

Survival function

$$S(t) = P(T > t) = 1 - P(T < t) = 1 - F_T(t) = e^{-\lambda t}$$

Hazard function:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

We just proved the memoryless property! **How?**
$h(t)$ is not a function of t. The hazard is **constant**.

Modeling with covariates

Suppose we want to allow the hazard to vary by some set of predictors.

Then, we can assume a **proportional hazards** model.

Hazard function given covariate set x

Baseline hazard

Hazard ratio

$$h(t \mid x) = h_0(t)e^{x\beta}$$

This changes for different families of hazard models

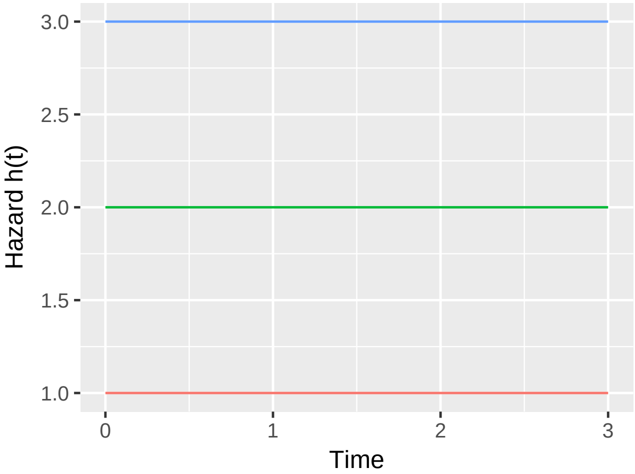
For the exponential, $h_0(t) = \lambda$

This is the same for all proportional hazard models

Why add covariates? It might be cloudy.

Photo credit: Hannah Lundberg

Exponential hazards



Question: If the green is the baseline hazard $h_0(t)$, what is the hazard ratio that produces the blue line? The red line?

Fitting an Exponential with survreg

```
> library(survival)
> fit <- survreg(Surv(time, event) ~ age + sex,
+                dist = "exponential",
+                data = lung)
> summary(fit)

Call:
survreg(formula = Surv(time, event) ~ age + sex, data = lung,
    dist = "exponential")
              Value Std. Error     z        p
(Intercept)  6.3597    0.63547 10.01 1.41e-23
age         -0.0156    0.00911 -1.72 8.63e-02
sex          0.4809    0.16709  2.88 4.00e-03

Exponential distribution
Loglik(model)= -1156.1   Loglik(intercept only)= -1162.3
Chisq= 12.48 on 2 degrees of freedom, p= 0.002
Number of Newton-Raphson Iterations: 4
n= 228
```

Interpreting hazard ratios

$$h(t \mid x) = h_0(t)e^{-x\beta}$$

```
> exp(-coef(fit))
(Intercept)          age          sex
      0.002        1.016        0.618
```

Q: How would you interpret these?
A year increase in age is associated with a 1.6% increase in the hazard, holding sex constant.

There are some things demographers just memorize.

We recommend just looking these up when you need them.

For instance, this fact:

> The survival function is e to the minus cumulative hazard.

Hazard function $\rightarrow$ survival function

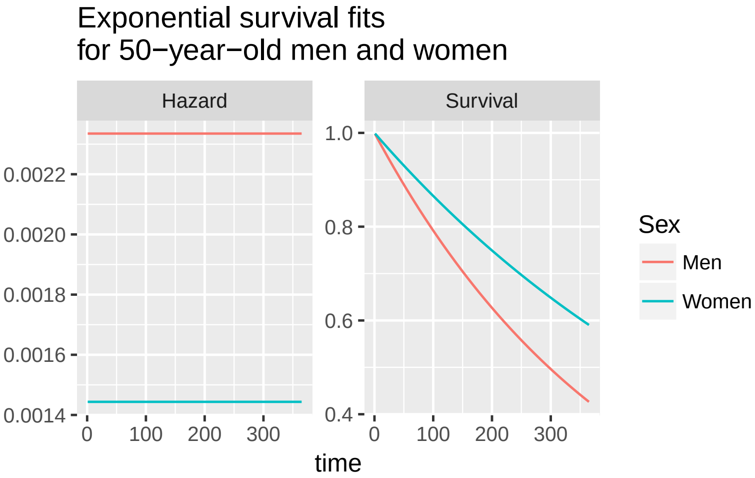
The derivative of the negative log of the survival function is

$$\begin{aligned}
\frac{\partial}{\partial t}\left(-\log\left[S(t)\right]\right) &= \frac{\frac{\partial}{\partial t}\left(-S(t)\right)}{S(t)} \\
&= \frac{\frac{\partial}{\partial t}\left(-[1-F(t)]\right)}{S(t)} \\
&= \frac{f(t)}{S(t)} = h(t)
\end{aligned}$$

Doing the reverse, we can go from $h(t)$ to $S(t)$

$$\begin{aligned}
\int_0^t \frac{\partial}{\partial t'}\left(-\log\left[S(t')\right]\right) dt &= \int_0^t h(t')dt' \\
-\log\left[S(t)\right] &= \int_0^t h(t')dt' \\
S(t) &= e^{-\int_0^t h(t')dt'}
\end{aligned}$$

Plotting survival curves



Plotting survival curves

How we made the previous slide:

```
data.frame(t = seq(.5,20,.5)) %>%
  mutate(Men.Hazard = lambda[1],
         Women.Hazard = lambda[2],
         Men.Survival = exp(-lambda[1]*t),
         Women.Survival = exp(-lambda[2]*t)) %>%
  melt(id = "t") %>%
  separate(variable, into = c("Sex","QOI")) %>%
  ggplot(aes(x = t, y = value, color = Sex)) +
  geom_line() +
  facet_wrap(~QOI, scales = "free") + ylab("") + xlab("time") +
  ggtitle("Exponential survival fits, for 50-year-old men and women") +
  ggsave("ExpoFit.pdf",
         height = 3, width = 5)
```

Scales and rates

The exponential is almost always parameterized with a **rate** λ.

But, it could just as well be defined in terms of a **scale** $\theta = \frac{1}{\lambda}$

Rate parameterization	**Scale parameterization**
$E(T) = \frac{1}{\lambda}$	$E(T) = \theta$
$f(T) = \lambda e^{-\lambda x}$	$f(T) = \frac{1}{\theta} e^{-\frac{1}{\theta} x}$
As rate grows, expected waiting time shrinks	As scale grows, expected waiting time grows

In general, you have to be careful with the parameterization of survival distributions.

What if we want the hazard to be a function of time?

Many options.

Weibull distribution

$$T \sim \text{Weibull}(\alpha, \lambda)$$

PDF [1]

$$f(t) = t^{\alpha-1}\alpha\lambda^{\alpha}e^{-(\lambda t)^{\alpha}}$$

CDF

$$F(t) = 1 - e^{-(\lambda t)^{\alpha}}$$

Survival function

$$S(t) = P(T > t) = 1 - P(T < t) = 1 - F_T(t) = e^{-(\lambda t)^{\alpha}}$$

Hazard function: Risk of event at t given survival up to t

$$h(t) = \frac{f(t)}{S(t)} = \frac{t^{\alpha-1}\alpha\lambda^{\alpha}e^{-(\lambda t)^{\alpha}}}{e^{-(\lambda t)^{\alpha}}} = t^{\alpha-1}\alpha\lambda^{\alpha}$$

[1] I have used the rate parameterization for λ; lecture slides use the scale parameterization.

Weibull distribution

$$h(t) = \frac{f(t)}{S(t)} = \frac{t^{\alpha-1}\alpha\lambda^{\alpha}e^{-(\lambda t)^{\alpha}}}{e^{-(\lambda t)^{\alpha}}} = t^{\alpha-1}\alpha\lambda^{\alpha}$$

The hazard increases with t when $\alpha > 1$
The hazard decreases with t when $\alpha < 1$
The hazard is constant over t when $\alpha = 1$
In that case, it's the exponential!

$$h(t \mid \alpha = 1) = t^{\alpha-1}\alpha\lambda^{\alpha} = t^{1-1}1\lambda^{1} = \lambda$$

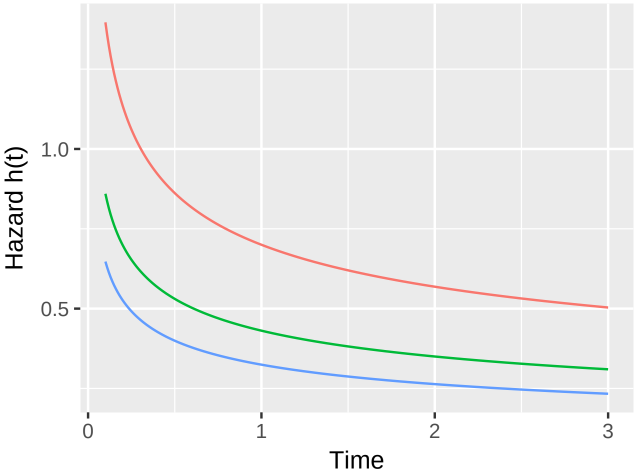
Discussion: If the Weibull contains the Exponential as a special case, why not always use the Weibull?

The Weibull is more **flexible**, which we like. If the world is Weibull but not Exponential, the Weibull is definitely better!

If the world is actually Exponential, we gain **efficiency** by making the assumption that the hazard is constant over time.

This is a **general theme** of statistics: Modeling assumptions buy us efficiency if they are correct.

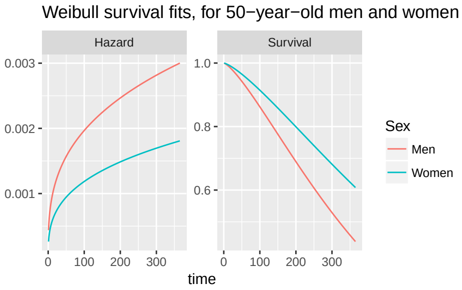
Weibull hazards



Fitting a Weibull model

```
## Fitting a Weibull model
fit <- survreg(Surv(time, event) ~ age + sex,
               dist = "weibull",
               data = lung)
```

Weibull results



Common question: The gap between those hazards clearly changes over time! Is this a violation of a modeling assumption?

A: No, they are still proportional!

(Also since these are fitted values, they necessarily agree with the modeling assumptions, so this was a trick question.)

You can fit a survival model using
any distribution
for which the support is
all positive numbers.

There are a huge number of options.

Lognormal distribution

$$T \sim \text{LogNormal}(\mu, \sigma^2) \sim e^Z \text{ (where } Z \sim N(\mu, \sigma^2)\text{)}$$

$$f(t) = \frac{1}{Y_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln(Y_i) - \mu)^2}{2\sigma^2}\right)$$

CDF

$$F(t) = \int_0^t f(x)dx = \text{ugly formula}$$

Survival function

$$S(t) = P(T > t) = 1 - P(T < t) = 1 - F_T(t) = \text{ugly formula}$$

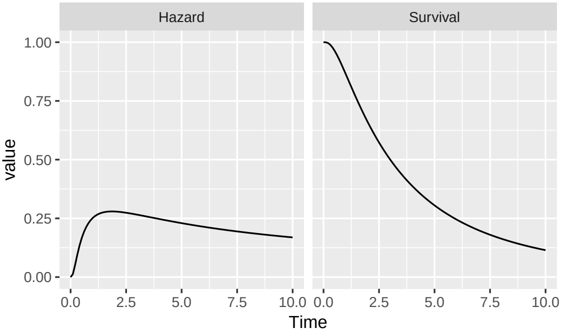
Hazard function: Risk of event at t given survival up to t

$$h(t) = \frac{f(t)}{S(t)} = \text{ugly formula}$$

Fitting a Lognormal

```
fit <- survreg(Surv(time, event) ~ age + sex,
               dist = "lognormal",
               data = lung)
```

Note: This figure doesn't correspond to the model above - just an example of a LogNormal



Gompertz distribution

$$
\begin{aligned}
f(t) &= b\eta e^{bt} e^{\eta} \exp(-\eta e^{bt}) \\
F(t) &= 1 - \exp\left(-\eta\left(e^{bt} - 1\right)\right) \\
h(t) &= \frac{f(t)}{S(t)} \\
&= \frac{b\eta e^{bt} e^{\eta} \exp(-\eta e^{bt})}{\exp\left(-\eta\left(e^{bt} - 1\right)\right)} \\
&= b\eta e^{bt} e^{\eta} \\
\log[h(t)] &= \underbrace{(\log(b) + \log(\eta) + \eta)}_{\text{Intercept}} + \underbrace{b}_{\text{Slope}} t \\
&= \alpha + \beta t
\end{aligned}
$$

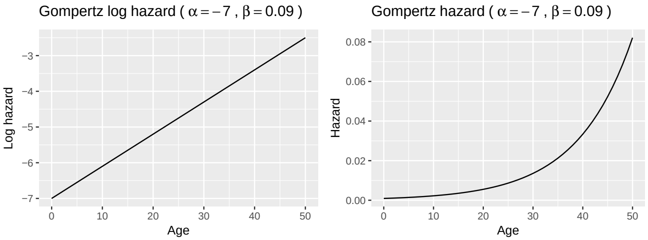
The log of the hazard function is linear in time!
This is why people like the Gompertz.

Gompertz distribution

Gompertz hazard with $\alpha = -7, \beta = .09$

$$\log[h(t)] = \alpha + \beta t, \quad h(t) = \exp(\alpha + \beta t)$$

Q: If the $\log[h(t)]$ increases linearly with t, what does $h(t)$ look like?



Q: For what questions would this be a good choice? Mortality

Note: Example motivated by U.S. mortality; see German Rodriguez's example here.

Time between breaks while hiking out of this valley.
You don't need a rest right away...

Donahue Pass, Yosemite. Photo credit: Riley Brian

...but after going for a while your hazard of resting increases.

Gompertz.

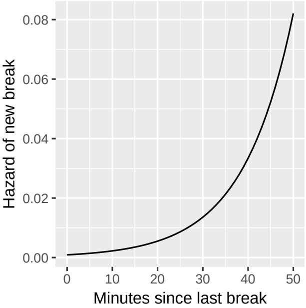


Photo credit: Riley Brian

As I said at the beginning, all of the survival models above have the form:

Hazard function given covariate set x

Baseline hazard

Hazard ratio

$$h(t \mid x) = h_0(t)e^{x\beta}$$

This changes for different families of hazard models

This is the same for all proportional hazard models

Different models allow different kinds of flexibility in the baseline hazard $h_0(t)$.

Can we model hazard ratios without any assumptions about $h_0(t)$?

Cox proportional hazards model

Then we can fit a Cox proportional hazards model!

To save time, I won't cover this here, but it's important and in lecture slides.

The Cox model is fit based on the order at which people die, rather than the times, so it does not assume a baseline hazard.

You can fit one with `coxph()`

Outline

Using distributions

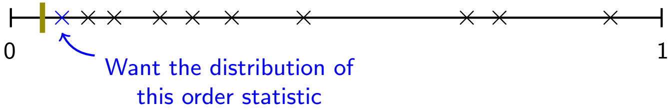
Most common question we are asked:
How do I know when to use a given distribution for a given problem?

When you know the **story of the distributions**, you can find one that **maps onto** your current problem.

An example we will answer by analogy

Suppose someone says to you, "I ran 10 hypothesis tests. What's the probability that the at least 1 p-values is less than 0.05 if all the null hypotheses are true?"

You draw this picture.

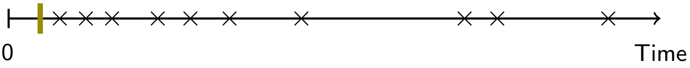


You reply:
"You want to know the distribution of the order statistic $U_{(3)}$. Let me take you to the wilderness. We will count shooting stars."

PC: `http://wilderness.org/30-prettiest-lakes-wildlands`

Imagine laying out on your pad on the granite, looking up at the sky.

We will count shooting stars and record the times we see them.[2]

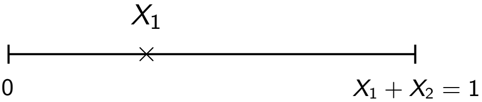


Shooting stars come at a **constant rate**.
The times between the arrivals are $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$.

[2]Thanks to William Chen for the shooting stars example. See more at `http://www.wzchen.com/probability-cheatsheet/`.

Suppose we saw the second star at time $X_1 + X_2 = 1$.

Q: What is the distribution of X_1 given this information?

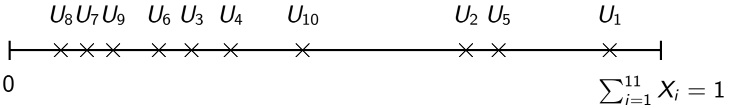


$$\frac{X_1}{X_1 + X_2} \sim \text{Uniform}(0, 1) \qquad \leftarrow \text{Same as } p\text{-value under } H_0!$$

Q: If I run one hypothesis test, what is the probability under the null that it falls below 0.05?

A: $P(U < .05) = P\left(\frac{X_1}{X_1+X_2} < .05\right) = 0.05$

Now suppose we observe $X_1, \ldots, X_{11}$ and we rescale so their sum is 1.

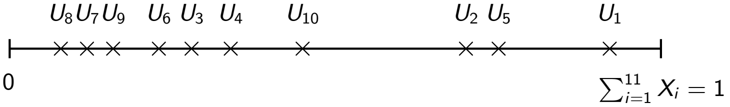


Let's re-label the $\times$ marks with U values with arbitrary indexes.

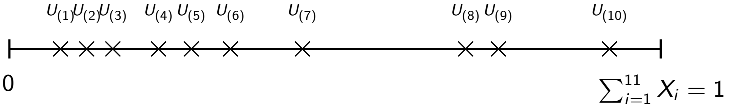
Q: What is the distribution of the $U_1, \ldots, U_{10}$?
A:
$U_1, \ldots, U_{10} \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$ $\leftarrow$ Same as 10 p-values under H_0!

There is a connection between p-values and shooting stars.

Order statistics



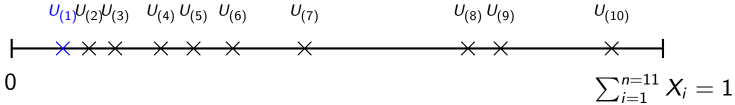
Let's denote the k-th order statistic by $U_{(k)}$.



$$U_{(k)} = \frac{\sum_{i=1}^{k} X_i}{\sum_{i=1}^{11} X_i}$$

A new distribution: The **Beta**

If $X_1, \ldots, X_n \sim$ Exponential, then $\frac{\sum_{i=1}^{k} X_i}{\sum_{i=1}^{n} X_i} \sim$ Beta(k,n - k - 1)

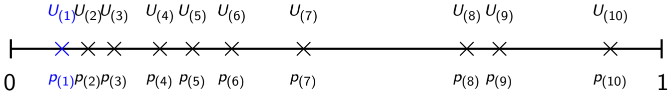


So $U_{(1)} \sim \text{Beta}(1, 10)$.

Q: Can you reason about the expected value of a Beta(1,10)?

A: There are 11 white space that we would expect to be of equal size, so we might expect that $\text{E}(U_{(1)}) = \frac{1}{11}$. **This is right!**

Q: Given what we know about shooting stars, what distribution do you think the smallest p-value takes?



$$p_{(1)} \sim \text{Beta}(1, 9)$$

Q: What is the probability that the smallest p-value is less than 0.05?

$$P(p_{(1)} < .05) = P(\text{Beta}(1, 10) < .5) = F_{\text{Beta}(1,9)}(.05) = 0.37$$

It is very easy to get a false positive by running 10 hypothesis tests!

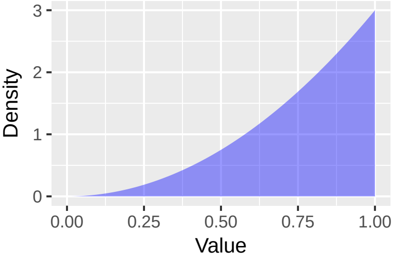
Key takeaways

We've taught you the stories of many distributions.

To use them, try to fit your problem into one of these **known stories**!

Photo credit: Hannah Lundberg

In my own research, I wanted to choose a prior distribution on a correlation that I expected to be near 1. I chose Beta(3,1).



I chose that by thinking:

- I want the distribution of the highest of 3 uniform draws.
- I want the distribution of the proportion of time spent waiting for 3 shooting stars, out of a total time spend waiting for 4.

Plugging your problem into a **known story** can help you find a solution.

Generalizing that story

Suppose someone says to you, "I ran 100 hypothesis tests. What's the probability that the 7th-smallest p-value is less than 0.05 if the null hypotheses are true?"

You say...let me take you to the wilderness. We will count shooting stars.

That is the proportion of time spent waiting for the 7th shooting star:

$$U_{(7)} \sim \text{Beta}(7, 93)$$

$$P(U_{(7)} < .05) = F_{\text{Beta}(7,93)}(.05) = 0.23$$

So, it's not that strange to see 7 p-values less than 0.05. And we learned this all from shooting stars!

One other story you might use

What if we wanted a distribution for the time until the kth star comes?

$$X_1, \ldots, X_k \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$$
$$G_k \sim X_1 + \cdots + X_k$$

Then we say

$$G_k \sim \text{Gamma}(k, \lambda)$$

The **Gamma distribution** characterizes the wait time until the kth star.

Outline

Side note: Zelig

Zelig is an R package designed to make everything we do in class easier.

Note the Zelig workflow overview.

We will use the Zelig-Exponential.

Zelig example: Lung cancer survival

We will walk through the example using data on lung cancer survival

```
> library(survival)
> data(lung)
> head(lung)
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
4    5  210      2  57   1       1       90        60     1150      11
5    1  883      2  60   1       0      100        90       NA       0
6   12 1022      1  74   1       1       50        80      513       0
lung <- mutate(lung, event = as.numeric(status == 2))
```

Variable definitions: Lung cancer survival

```
?lung
inst: Institution code
time: Survival time in days
status: censoring status 1=censored, 2=dead
age: Age in years
sex: Male=1 Female=2
ph.ecog: ECOG performance score (0=good 5=dead)
ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
pat.karno: Karnofsky performance score as rated by patient
meal.cal: Calories consumed at meals
wt.loss: Weight loss in last six months
```

Zelig step 1: Fit a model

```
fit <- zelig(Surv(time, event) ~ age + sex,
             model = "exp",
             data = lung)
```

Zelig step 1: Fit a model

```
> summary(fit)
Model:

Call:
z5$zelig(formula = Surv(time, event) ~ age + sex, data = lung)
              Value Std. Error     z        p
(Intercept)  6.3597    0.63547 10.01 1.41e-23
age         -0.0156    0.00911 -1.72 8.63e-02
sex          0.4809    0.16709  2.88 4.00e-03

Scale fixed at 1

Exponential distribution
Loglik(model)= -1156.1   Loglik(intercept only)= -1162.3
Chisq= 12.48 on 2 degrees of freedom, p= 0.002
Number of Newton-Raphson Iterations: 4
n= 228

Next step: Use 'setx' method
```

Zelig step 2: Use `setx` to set covariates of interest

```
men <- setx(fit, age = 50, sex = 1)
women <- setx(fit, age = 50, sex = 2)
```

Zelig step 2: Use `setx` to set covariates of interest

```
> men
setx:
  (Intercept) age sex
1           1  50   1

Next step: Use 'sim' method
> women
setx:
  (Intercept) age sex
1           1  50   2

Next step: Use 'sim' method
```

Zelig step 3: Use `sim` to simulate quantities of interest

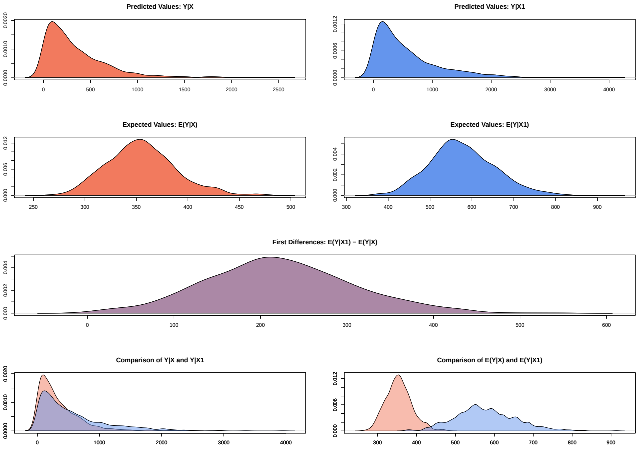
```
> sims <- sim(obj = fit, x = men, x1 = women)
> summary(sims)

 sim x :
 -----
ev
    mean       sd      50%     2.5%   97.5%
1 355.086 33.63733 353.5258 296.6169 428.758
pv
       mean       sd     50%     2.5%    97.5%
[1,] 351.414 361.6174 242.511 7.082744 1357.005

 sim x1 :
 -----
ev
     mean      sd     50%     2.5%    97.5%
1 577.5684 78.5113 571.178 438.4341 743.9957
pv
         mean       sd      50%    2.5%   97.5%
[1,] 562.8317 550.6102 382.9658 11.5627 2016.61
fd
     mean      sd      50%     2.5%    97.5%
1 222.4824 85.0493 217.0278 61.08082 396.5632
```

Zelig step 4: Use graph to plot simulation results

```
pdf("ZeligFigures.pdf",
    height = 5, width = 7)
plot(sims)
dev.off()
```



Summarizing Zelig

Estimate your model:

```
#install.packages("Zelig")
require(Zelig)
fit <- zelig(Surv(time, event) ~ age + sex,
             model = "exp",
             data = lung)
```

Set your covariates:

```
men <- setx(fit, sex = 1, fn = mean)
women <- setx(fit, sex = 2, fn = mean)
```

Simulate your QOI:

```
sims <- sim(obj = fit, x = men, x1 = women)
```

Plot:

```
plot(sims)
```

After break: expectation maximization, missing data

Cards! Questions?

Photo credit: Riley Brian