Duration Models and Matching

Stephen Pettigrew

April 9, 2014

Outline

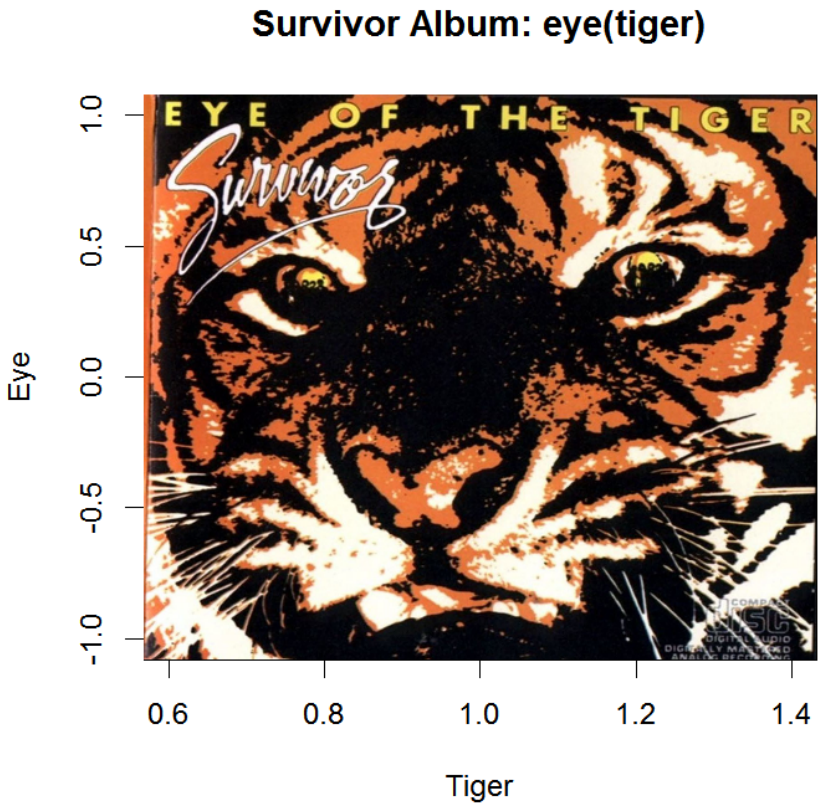
Outline

Logistics

- Pset 6 due tonight. No new pset this week
- Submit a title and abstract for your final paper
  - One per group
  - On Canvas > Modules > Replication Paper > Replication Abstract > Submit Assignment (on right)
- RSVP to the party at Gary's house if you haven't already
  - Saturday, April 19 at 12:30
  - Near the Green Line and the 86 bus that goes through Harvard Square

From last week

I sincerely appreciate feedback on how to make these sections more helpful for you.

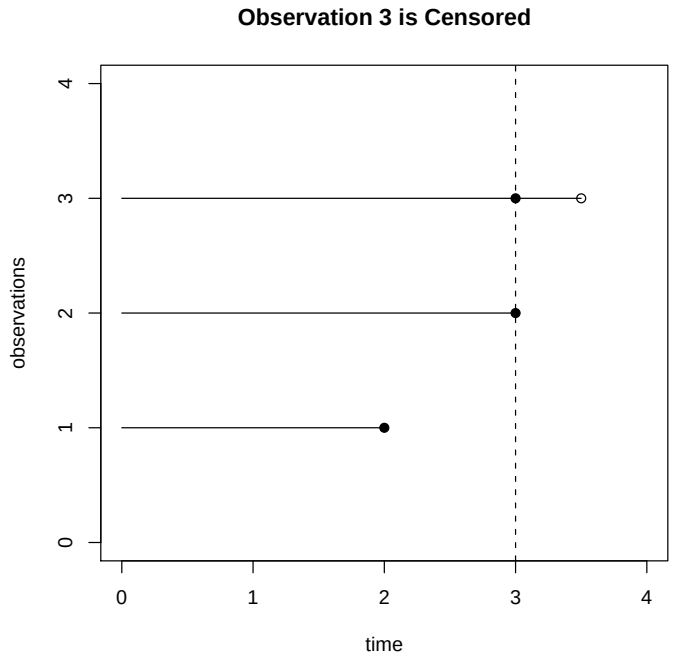


Outline

Review from last week

Three reasons we use duration models:

1. OLS assumes Y is Normal but duration dependent variables are always positive (number of years, number of days. etc.)
2. Duration models can handle time-varying covariates
3. Duration models can handle censoring



Duration Model Notation

T: a continuous, positive random variable representing the duration/survival times ($T = Y$)

f(t): the probability density function of T (the stochastic component)

F(t): the CDF of $f(t)$, $\int_0^t f(u)du = P(T \leq t)$, which is the probability of an event occurring before (or at exactly) time t

Survivor function: the probability of surviving (i.e. no event occurring) until at least time t: $S(t) = 1 - F(t) = P(T > t)$

Hazard function or hazard rate: the probability of an event at time t given survival up to time t: $h(t) = P(t \leq T < t + \tau | T \geq t)$

Duration Model Notation

$$
\begin{aligned}
h(t) &= P(t \leq T < t + \tau | T \geq t) \\
&= P(\text{event at } t | \text{survival up to } t) \\
&= \frac{P(\text{survival up to } t | \text{event at } t) P(\text{event at } t)}{P(\text{survival up to } t)} \\
&= \frac{P(\text{event at } t)}{P(\text{survival up to } t)} \\
h(t) &= \frac{f(t)}{S(t)}
\end{aligned}
$$

Duration Model Notation

Therefore:

$$\underbrace{f(t)}_{\text{density function}} = \underbrace{h(t)}_{\text{hazard function}} \cdot \underbrace{S(t)}_{\text{survival function}}$$

$$\underbrace{f(t)}_{\text{density function}} = \underbrace{\frac{f(t)}{S(t)}}_{\text{hazard function}} \cdot \underbrace{S(t)}_{\text{survival function}}$$

Handling censoring

We know that censored observations they survived at least until some observed time, t^c, and that their true duration, t is greater than or equal to t^c.

For each observation, let's create a censoring indicator variable, c_i, such that

$$c_i = \begin{cases} 1 & \text{if censored} \\ 0 & \text{if not censored} \end{cases}$$

Censoring

We can incorporate the information from the censored observations into the likelihood function.

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [f(t_i)]^{1-c_i} \; [P(T_i \geq t_i^c)]^{c_i} \\
&= \prod_{i=1}^{n} [f(t_i)]^{1-c_i} \; [1 - F(t_i)]^{c_i} \\
&= \prod_{i=1}^{n} [f(t_i)]^{1-c_i} \; [S(t_i)]^{c_i}
\end{aligned}$$

Uncensored observations contribute to the density function and censored observations contribute to the survivor function in the likelihood.

Outline

The Poisson Process

- Popular example of stochastic process
- Principles of Poisson process:
  - **Independent increments**: number of events occurring in two disjoint intervals is independent
  - **Stationary increments**: probability distribution of number of occurrences depends only on the time length of interval (because of common rate)
- Events occur at rate λ (expected occurrences per unit of time)
- N_τ = number of arrivals in time period of length τ
  - $N_\tau \sim \text{Poisson}(\lambda\tau)$

The Poisson Process

- Exponential distribution measures the times between events in a Poisson process
- T = time to wait until next event in a Poisson process with rate λ
- $T \sim \text{Expo}(\lambda)$
- **Memorylessness property**: how much you have waited already is irrelevant

$$P(T > t + k | T > t) = P(t > k)$$

$$P(T > 3 + 5 | T > 3) = P(t > 5)$$

Two Possible Parameterizations of the Exponential Model

- $\lambda_i > 0$ is the **rate** parameter

$$T_i \sim \text{Exponential}(\lambda_i)$$

$$f(t_i) = \lambda_i e^{-\lambda_i t_i}$$

$$E(T_i) = \frac{1}{\lambda_i}$$

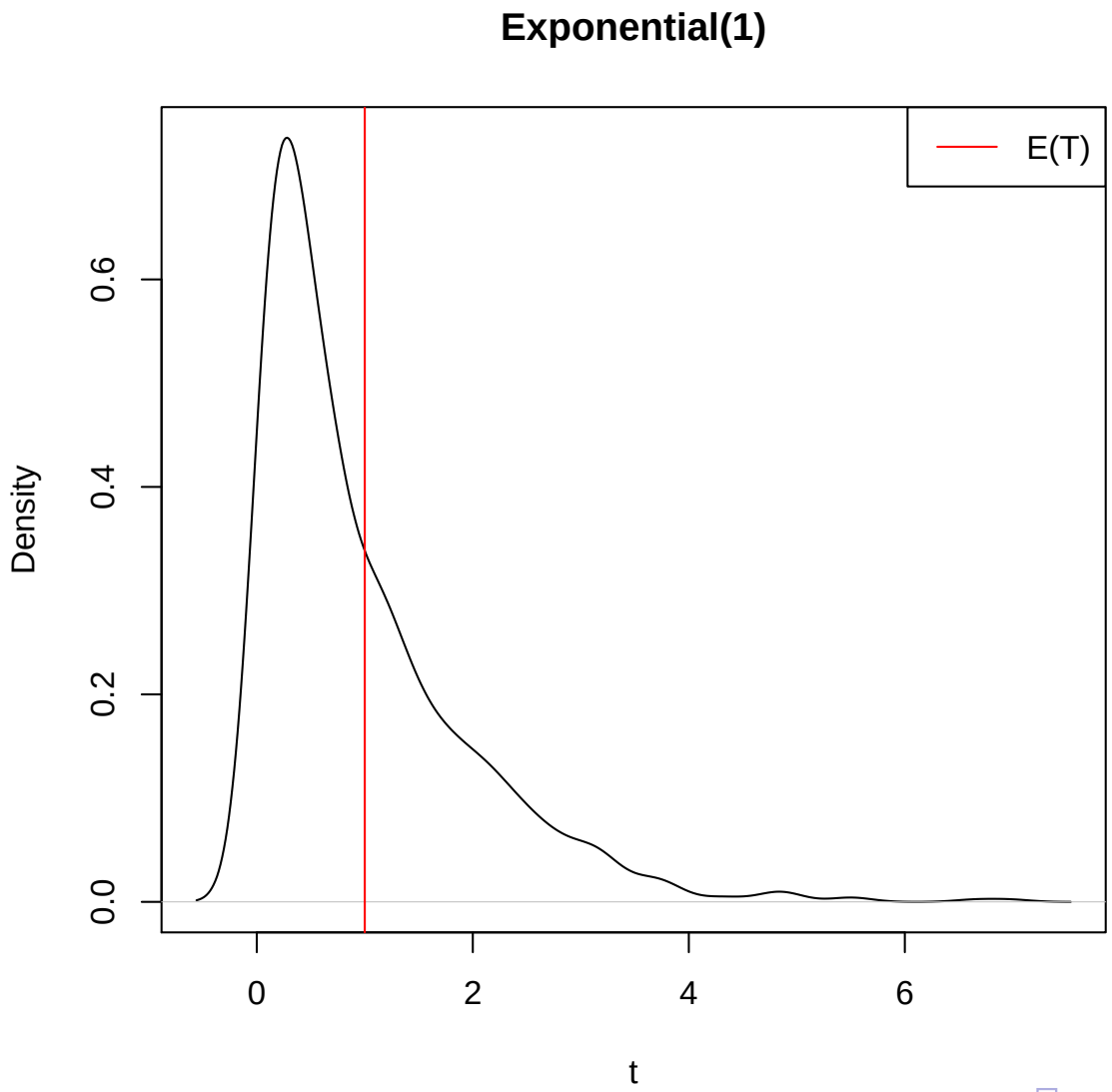
- $\theta_i > 0$ is **scale** parameter ($\theta_i = \frac{1}{\lambda_i}$)

$$T_i \sim \text{Exponential}(\theta_i)$$

$$f(t_i) = \frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}$$

$$E(T_i) = \theta_i$$

The Exponential Model



Link Functions

- If you use a rate parameterization with λ_i:

$$E(T_i) = \frac{1}{\lambda_i} = \frac{1}{\exp(x_i\beta)}$$

  Positive β implies that expected duration time decreases as x increases.

- If you use a scale parameterization with θ_i

$$E(T_i) = \theta_i = \exp(x_i\beta)$$

  Positive β implies that expected duration time increases as x increases.

Hazard Function for Rate Parametrization

For $T_i \sim \text{Exponential}(\lambda_i)$:

$$
\begin{aligned}
f(t) &= \lambda_i e^{-\lambda_i t} \\
S(t) &= 1 - F(t) \\
&= 1 - (1 - e^{-\lambda t}) \\
&= e^{-\lambda_i t} \\
h(t) &= \frac{f(t)}{S(t)} \\
&= \frac{\lambda_i e^{-\lambda_i t}}{e^{-\lambda_i t}} \\
&= \lambda_i
\end{aligned}
$$

Hazard Function for Scale Parametrization

For $T_i \sim \text{Exponential}(\theta_i)$:

$$
\begin{aligned}
f(t) &= \frac{1}{\theta_i} \exp[-\frac{t}{\theta_i}] \\
S(t) &= 1 - F(t) \\
&= 1 - (1 - \exp[-\frac{t}{\theta_i}]) \\
&= \exp[-\frac{t}{\theta_i}] \\
h(t) &= \frac{f(t)}{S(t)} \\
&= \frac{\frac{1}{\theta_i} \exp[-\frac{t}{\theta_i}]}{\exp[-\frac{t}{\theta_i}]} = \frac{1}{\theta_i}
\end{aligned}
$$

Let's work with the scale parametrization

- Note that $h(t) = \frac{1}{\theta_i}$, which does not depend on t!
  - The exponential model thus assume a **flat hazard**: Every unit / individual has their own hazard rate, but it does not change over time
  - Connected to **memorylessness property** of the exponential distribution

Modeling $h(t)$ with covariates:

$$h(t) = \frac{1}{\theta_i} = \exp[-x_i\beta]$$

Positive β implies that hazard decreases and average survival time increases as x increases.

Estimation via ML:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [f(t_i)]^{1-c_i} [1 - F(t_i)]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}} \right]^{1-c_i} \left[e^{-\frac{t_i}{\theta_i}} \right]^{c_i} \\
\ell &= \sum_{i=1}^{n} (1 - c_i)(\ln \frac{1}{\theta_i} - \frac{t_i}{\theta_i}) + c_i(-\frac{t_i}{\theta_i}) \\
&= \sum_{i=1}^{n} (1 - c_i)(\ln e^{-\mathbf{x}_i\beta} - e^{-\mathbf{x}_i\beta} t_i) + c_i(-e^{-\mathbf{x}_i\beta} t_i) \\
&= \sum_{i=1}^{n} (1 - c_i)(-\mathbf{x}_i\beta - e^{-\mathbf{x}_i\beta} t_i) - c_i(e^{-\mathbf{x}_i\beta} t_i) \\
&= \sum_{i=1}^{n} (1 - c_i)(-\mathbf{x}_i\beta) - e^{-\mathbf{x}_i\beta} t_i
\end{aligned}
$$

Quantities of interest

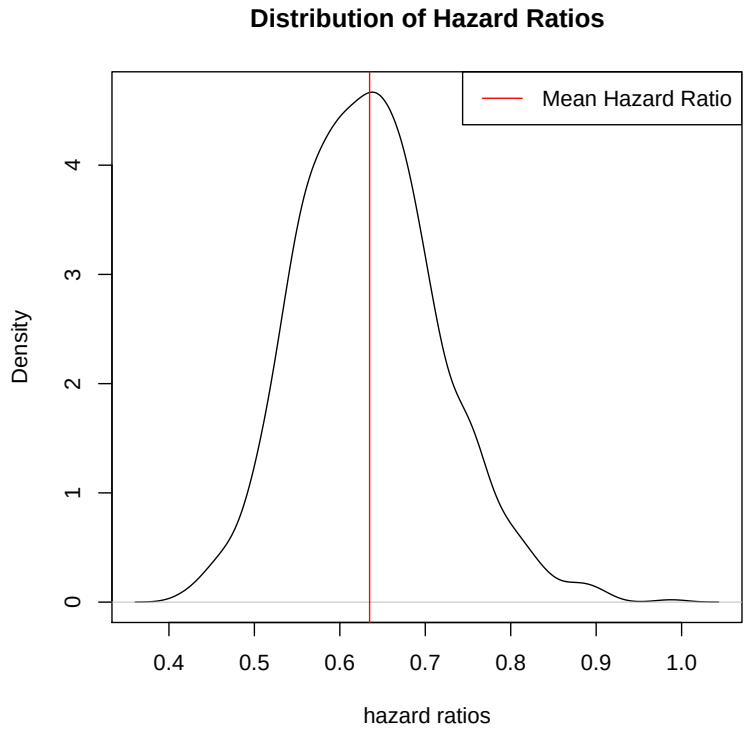
If our outcome variable is how long a parliamentary government lasts, and we're interested in the effect of majority versus minority governments. We could calculate:

- Find the hazard ratio of majority to minority governments
- Expected survival time for majority and minority governments
- Predicted survival times for majority and minority governments
- First differences in expected survival times between majority and minority governments

Hazard Ratios

$$
\begin{aligned}
\text{HR} &= \frac{h(t|\mathbf{x}_{\text{maj}})}{h(t|\mathbf{x}_{\text{min}})} \\
&= \frac{e^{-\mathbf{x}_{\text{maj}}\beta}}{e^{-\mathbf{x}_{\text{min}}\beta}} \\
&= \frac{e^{-\beta_0}e^{-x_1\beta_1}e^{-x_2\beta_2}e^{-x_3\beta_3}e^{-x_{\text{maj}}\beta_4}e^{-x_5\beta_5}}{e^{-\beta_0}e^{-x_1\beta_1}e^{-x_2\beta_2}e^{-x_3\beta_3}e^{-x_{\text{min}}\beta_4}e^{-x_5\beta_5}} \\
&= \frac{e^{-x_{\text{maj}}\beta_4}}{e^{-x_{\text{min}}\beta_4}} \\
&= e^{-\beta_4}
\end{aligned}
$$

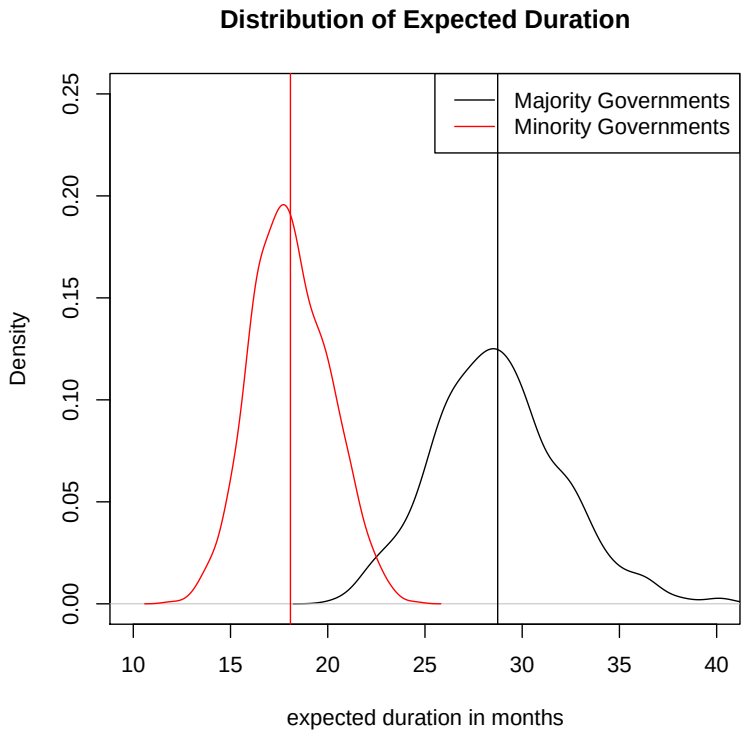
Hazard ratio greater than 1 implies that majority governments fall faster (shorter survival time) than minority governments.



Majority governments survive longer than minority governments.

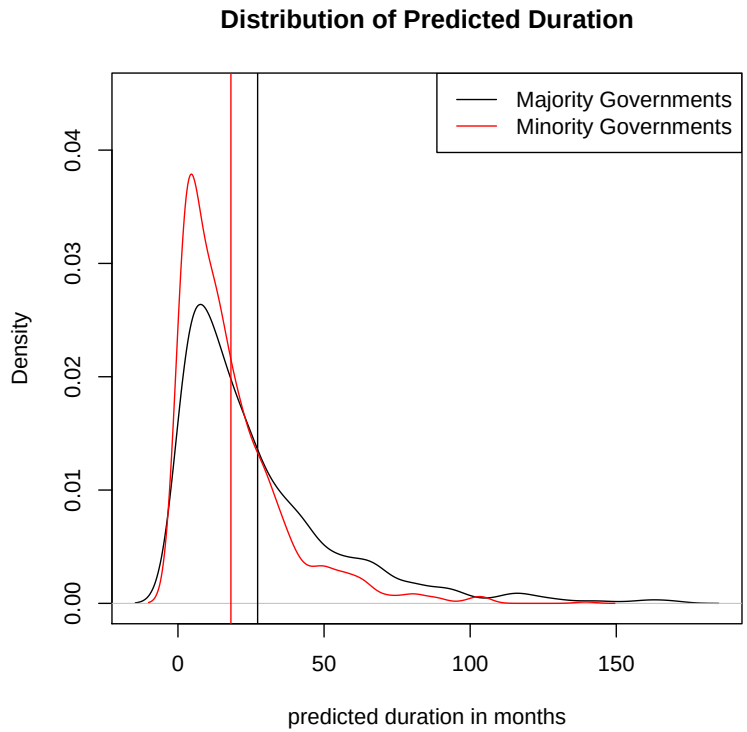
Expected (average) Survival Time

$$\begin{aligned} E(T|\mathbf{x}_i) &= \theta_i \\ &= \exp[\mathbf{x}_i\beta] \end{aligned}$$



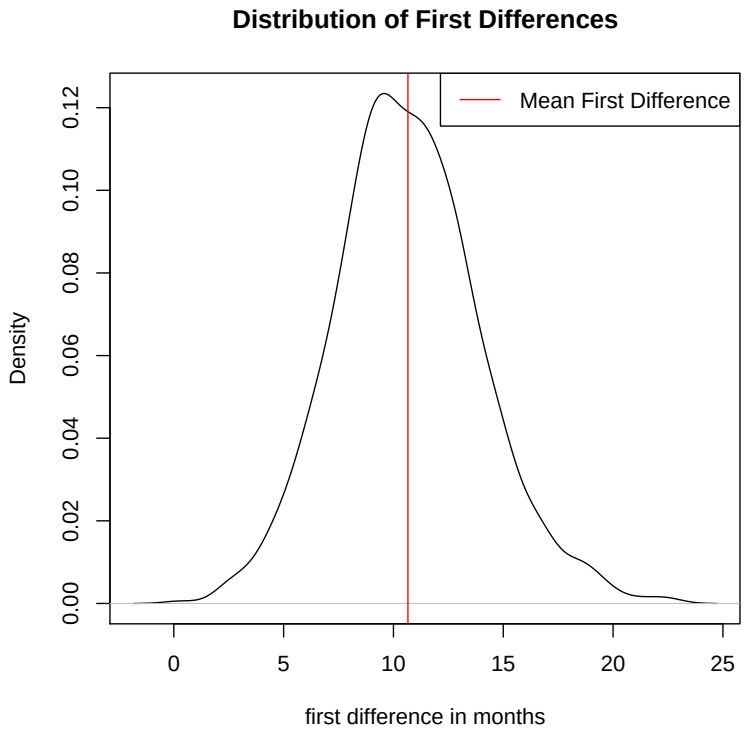
Predicted Survival Time

Draw predicted values from the exponential distribution.



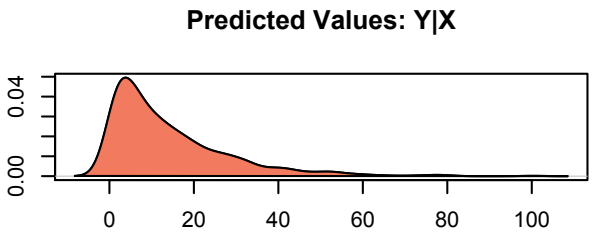
First Differences

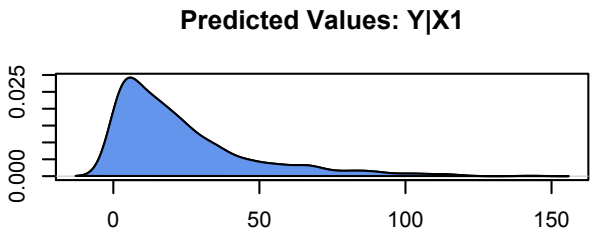
$$E(T|\mathbf{x}_{\text{maj}}) - E(T|\mathbf{x}_{\text{min}})$$

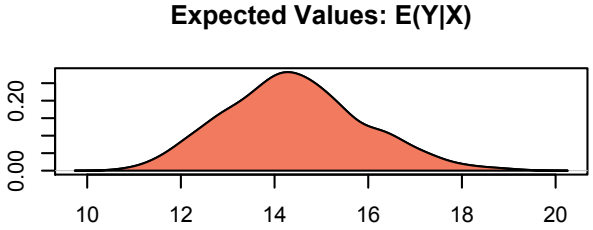


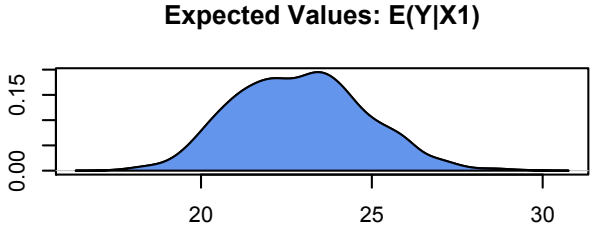
Quantities of Interest in Zelig

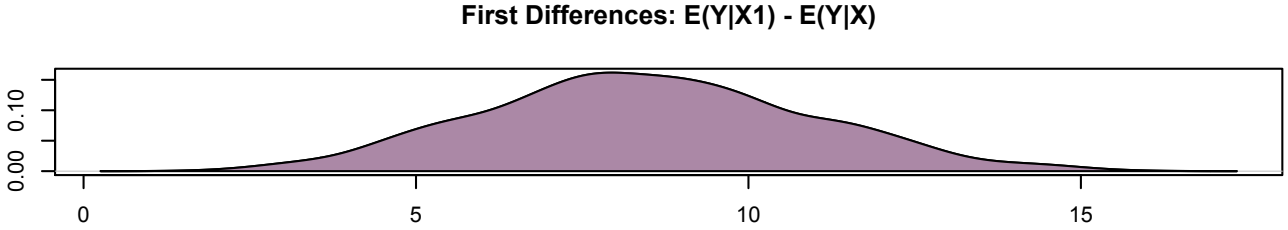
```
x.min <- setx(z.out,numst2=0)
x.maj <- setx(z.out,numst2=1)
s.out <- sim(z.out, x=x.min,x1=x.maj)
summary(s.out)
plot(s.out)
```

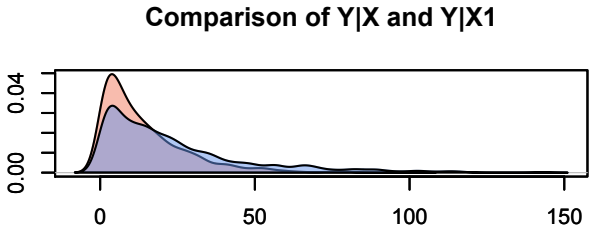

Predicted Values: Y|X
0.00
0.04
0
20
40
60
80
100

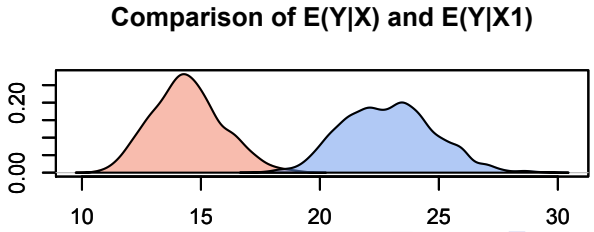

Predicted Values: Y|X1
0.000
0.025
0
50
100
150


Expected Values: E(Y|X)
0.00
0.20
10
12
14
16
18
20


Expected Values: E(Y|X1)
0.00
0.15
20
25
30


First Differences: E(Y|X1) - E(Y|X)
0.00
0.10
0
5
10
15


Comparison of Y|X and Y|X1
0.00
0.04
0
50
100
150


Comparison of E(Y|X) and E(Y|X1)
0.00
0.20
10
15
20
25
30

The exponential model is nice and simple, but the assumption of a flat hazard may be too restrictive.

What if we want to loosen that restriction by assuming a monotonic hazard?

We can use the Weibull model.

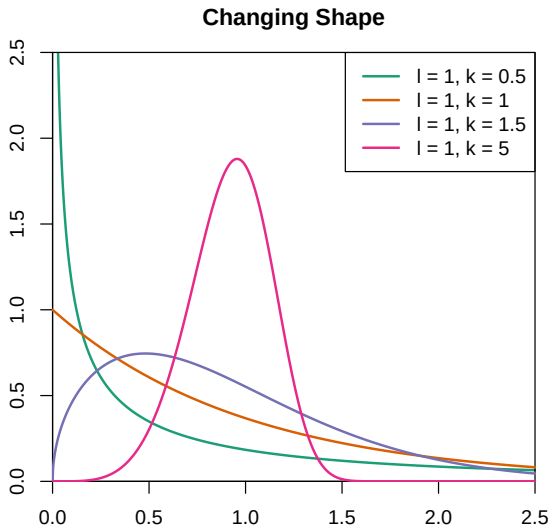
Outline

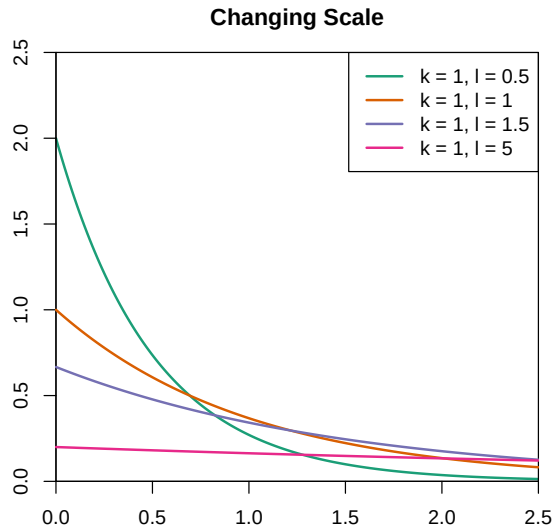
The Weibull Model

Similar to how we generalized the Poisson into a Negative Binomial by adding a parameter, we can do the same with the Exponential by turning it into a Weibull:

$$\begin{aligned} T_i &\sim \text{Weibull}(\lambda_i, \alpha) \\ E(T_i) &= \lambda_i \Gamma\left(1 + \frac{1}{\alpha}\right) \end{aligned}$$

$\lambda_i > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter.





The Weibull Model

$$f(t_i) = \left(\frac{\alpha}{\lambda_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\lambda_i}\right)^{\alpha}\right]$$

Model λ_i with covariates in the systematic component:

$$\lambda_i = \exp(x_i\beta)$$

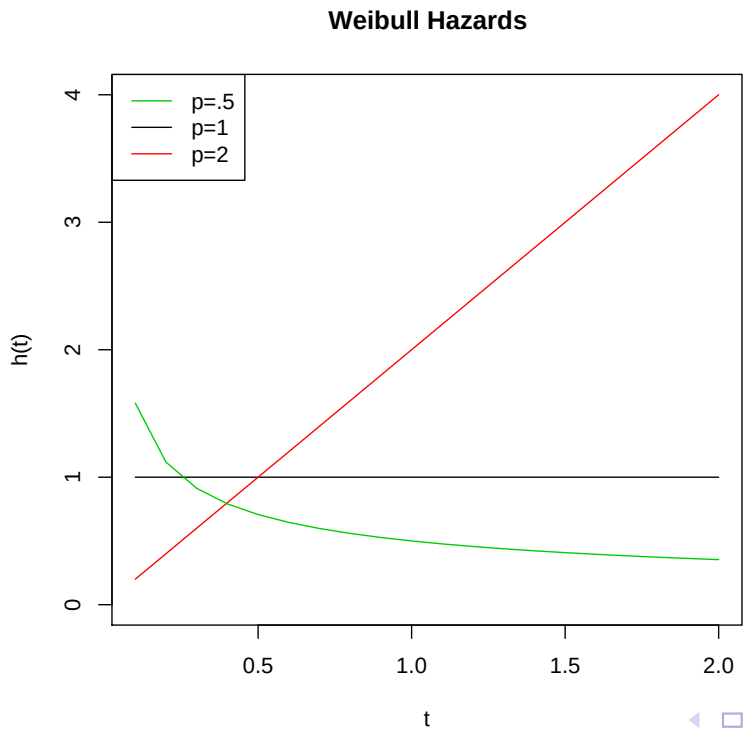
Positive β implies that expected duration time increases as x increases.

$$
\begin{aligned}
f(t_i) &= \left(\frac{\alpha}{\lambda_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\lambda_i}\right)^{\alpha}\right] \\
S(t_i) &= 1 - F(t_i) \\
&= 1 - (1 - e^{-(t_i/\lambda_i)^{\alpha}}) \\
&= e^{-(t_i)/\lambda_i)^{\alpha}} \\
h(t_i) &= \frac{f(t_i)}{S(t_i)} \\
&= \frac{\left(\frac{\alpha}{\lambda_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\lambda_i}\right)^{\alpha}\right]}{e^{-(t_i/\lambda_i)^{\alpha}}} \\
&= \left(\frac{\alpha}{\lambda_i}\right)\left(\frac{t_i}{\lambda_i}\right)^{\alpha-1} \\
&= \left(\frac{\alpha}{\lambda_i^{\alpha}}\right) t_i^{\alpha-1}
\end{aligned}
$$

Hazard monotonicity assumption

$h(t_i)$ is modeled with both λ_i and α and is a function of t_i. Thus, the Weibull model assumes a **monotonic hazard**.

- If $\alpha = 1$, $h(t_i)$ is flat and the model is the exponential model.
- If $\alpha > 1$, $h(t_i)$ is monotonically increasing.
- If $\alpha < 1$, $h(t_i)$ is monotonically decreasing.



The shape parameter α for the Weibull distribution is the reciprocal of the scale parameter given by `survreg()`.

The scale parameter given by `survreg()` is NOT the same as the scale parameter in the Weibull distribution, which should be $\lambda_i = e^{\mathbf{x}_i\beta}$.

Hazard Ratios

One quantity of interest is the hazard ratio:

$$HR = \frac{h(t|x=1)}{h(t|x=0)}$$

With the Weibull model we make a **proportional hazards** assumption: hazard ratio does not depend t.

Other Parametric Models

- Gompertz model: monotonic hazard
- Log-logistic or log-normal model: nonmonotonic hazard
- Generalized gamma model: nests the exponential, Weibull, log-normal, and gamma models with an extra parameter

But what if we don't want to make an assumption about the shape of the hazard?

The Cox Proportional Hazards Model

Often described as a semi-parametric model.
Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

- Only quantities of interest are hazard ratios.
- Can be subject to overfitting
- Shape of hazard is unknown (although there are semi-parametric ways to derive the hazard and survivor functions)

More resources about survival modeling

Box-Steffensmeier, Janet M. and Bradford S. Jones. 2004. *Event History Modeling.* Cambridge University Press.

King, Gary, James E. Alt, Nancy E. Burns, and Michael Laver. 1990. "A Unified Model of Cabinet Dissolution in Parliamentary Democracies." *American Journal of Political Science* 34(3): 846-971

Long, S. J. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: SAGE Publications, Inc.

McCullagh, Peter; Nelder, John (1989). *Generalized Linear Models*, Second Edition. Boca Raton: Chapman and Hall/CRC

Lam, Patrick. Survival Model Notes. http://www.patricklam.org/teaching.html

Outline

Setup

- Let's denote treatment as $T \in 0, 1$. $T = 1$ is treated group, $T = 0$ is control group.
- We have an outcome Y
- SUTVA: stable unit treatment value assumption
  - No interference between units - i.e. units don't talk to each other about the experiment
  - No hidden levels of treatment
- Under SUTVA, we have 2 potential outcomes per unit:

$$Y_i(1) \text{ and } Y_i(0)$$

- Before we think about how treatment was assigned, we can express this information in a potential outcomes table

Potential Outcomes Table

i	Name	$Y_i(1)$	$Y_i(0)$
1	George Washington	$Y_1(1)$	$Y_1(0)$
2	Toucan Sam	$Y_2(1)$	$Y_2(0)$
3	Anne Boleyn	$Y_3(1)$	$Y_3(0)$
4	Lisa Nowak	$Y_4(1)$	$Y_4(0)$
5	Dr. Phil	$Y_5(1)$	$Y_5(0)$
6	Herschel Walker	$Y_6(1)$	$Y_6(0)$

Individual causal effect for George Washington: $Y_1(1)$-$Y_1(0)$

All potential outcomes are fixed pre-treatment characteristics of individuals

Defining the Estimand

We should define an estimand before we think about estimation! What is a causal quantity we are interested in?

For example, an average treatment effect (ATE):

$$\mathrm{E}[Y_i(1) - Y_i(0)]$$

Fundamental Problem of Causal Inference

i	Name	$Y_i(1)$	$Y_i(0)$	T_i
1	George Washington	$Y_1(1)$	?	1
2	Toucan Sam	$Y_2(1)$	?	1
3	Anne Boleyn	?	$Y_3(0)$	0
4	Lisa Nowak	$Y_4(1)$	?	1
5	Dr. Phil	?	$Y_5(0)$	0
6	Herschel Walker	?	$Y_6(0)$	0

The **fundamental problem of causal inference** is that we only observe one potential outcome per unit!

Fundamental Problem of Causal Inference

How do we estimate the average treatment effect (ATE) from observed data?

$$\mathrm{E}[Y_i(1) - Y_i(0)]$$

Causal inference is a missing data problem!

We'll have to impute missing potential outcomes.

Unconfoundedness

To do this, we generally make another assumption: **unconfoundedness** of treatment,

$$P(T|Y(0), Y(1), X) = P(T)$$

Unconfoundedness translates to the following:

$$\mathrm{E}[Y(1)] = \mathrm{E}[Y(1)|T = 1] = \mathrm{E}[Y(1)|T = 0]$$
$$\mathrm{E}[Y(0)] = \mathrm{E}[Y(0)|T = 1] = \mathrm{E}[Y(0)|T = 0]$$

Now:

$$\begin{aligned} \mathrm{E}[Y(1) - Y(0)] &= \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] \\ &= E[Y(1)|T = 1] - E[Y(0)|T = 0] \\ &= E[Y^{obs}|T = 1] - E[Y^{obs}|T = 0] \end{aligned}$$

Causal Effects with Unconfoundedness

	T=1 (Treatment)	T=0 (Control)
$E[Y\|T=t]$	6.6	2.4

If this is our data, how would we estimate the average causal effect of T on Y? Assuming unconfoundedness:

$$\begin{aligned} \mathrm{E}[Y(1)-Y(0)] &= E[Y(1)|T=1]-E[Y(0)|T=0] \\ &= E[Y^{obs}|T=1]-E[Y^{obs}|T=0] \\ &= 6.6-2.4 \\ &= 4.2 \end{aligned}$$

Unconfoundedness

Unconfoundedness implies:

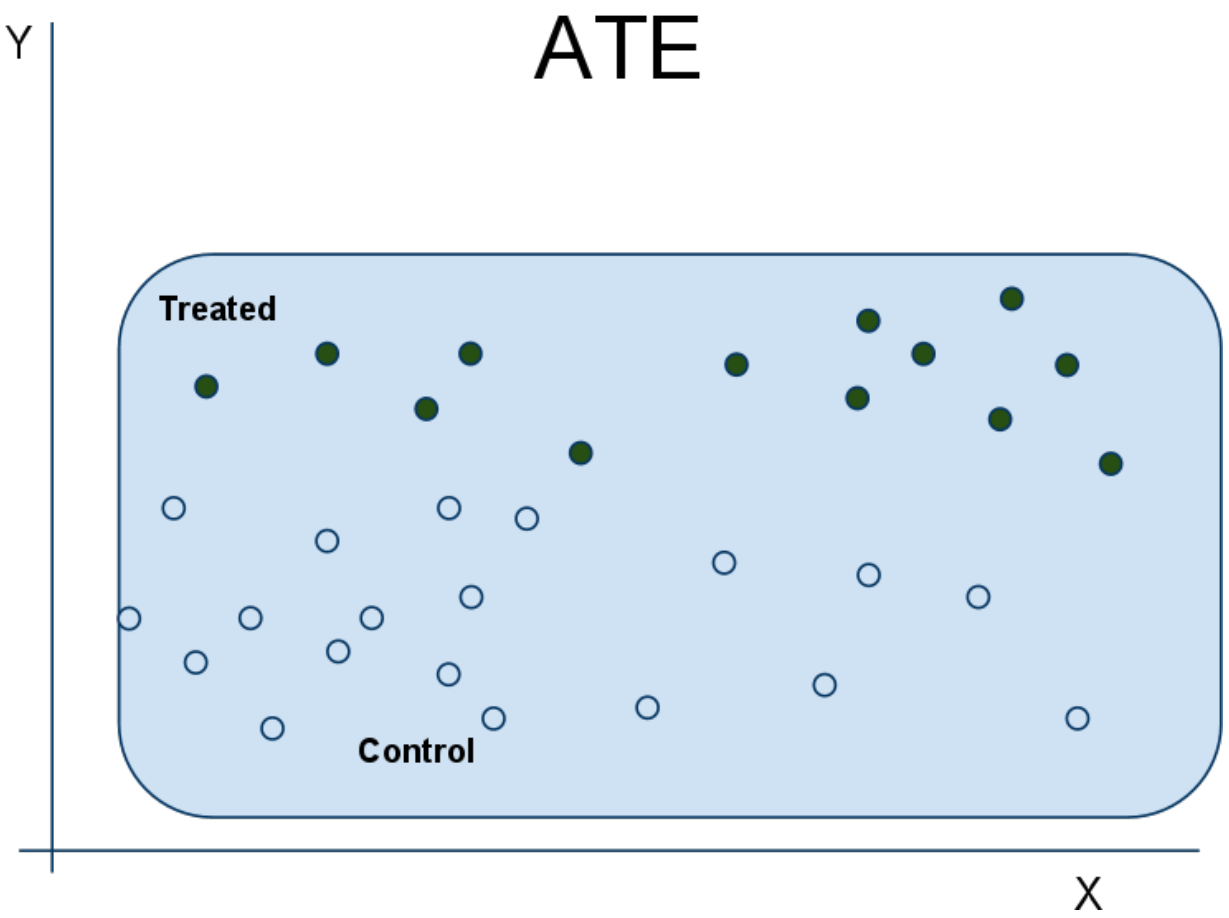
1. $\mathrm{E}[Y(1)] = \mathrm{E}[Y(1)|T = 1] = \mathrm{E}[Y(1)|T = 0]$
2. $\mathrm{E}[Y(0)] = \mathrm{E}[Y(0)|T = 1] = \mathrm{E}[Y(0)|T = 0]$

If we only assume #2 then we can still calculate the average treatment effect on the treated (ATT):

$$\begin{aligned} E[Y(1) - Y(0)|T = 1] &= E[Y(1)|T = 1] - E[Y(0)|T = 1] \\ &= E[Y(1)|T = 1] - E[Y(0)|T = 0] \\ &= E[Y^{obs}|T = 1] - E[Y^{obs}|T = 0] \end{aligned}$$

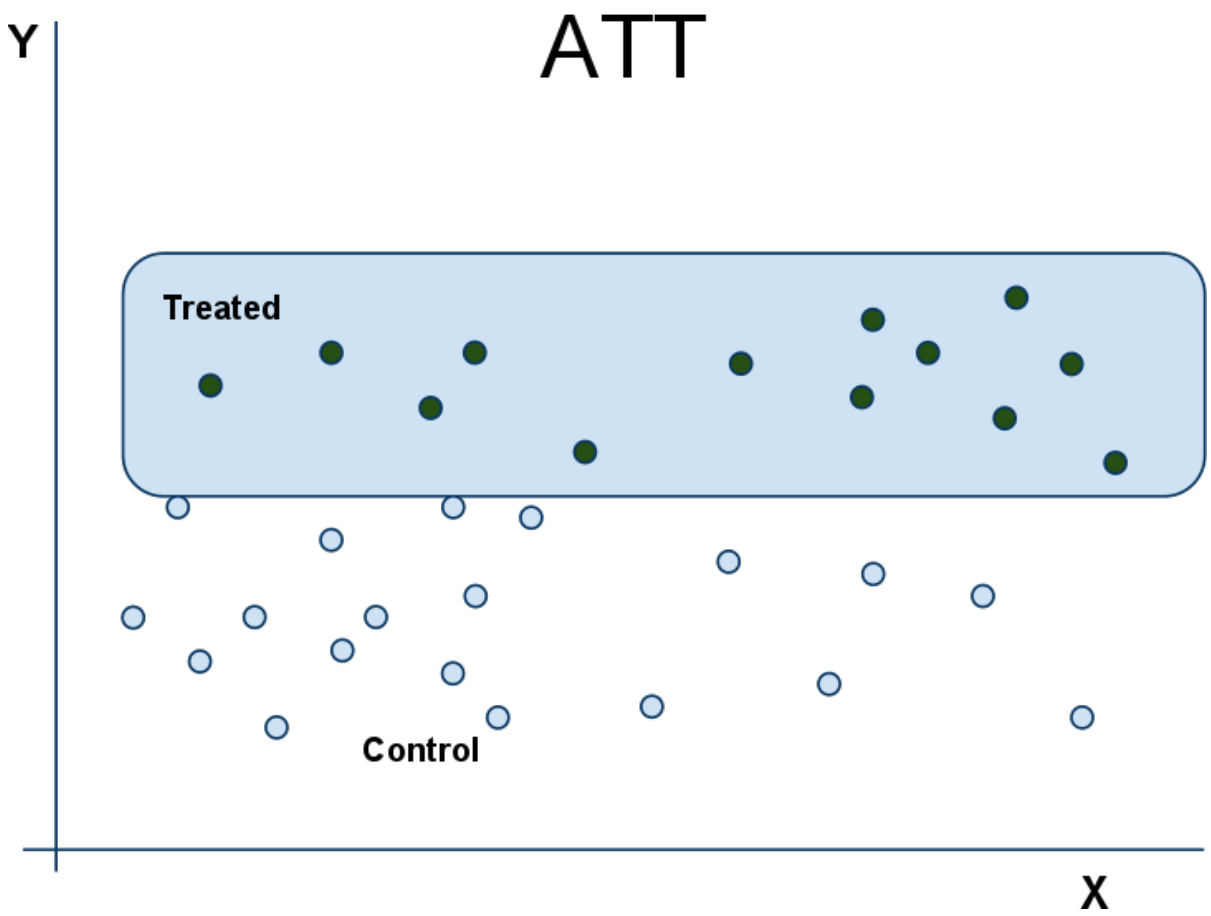
ATT vs. ATE vs. ATC vs. FSATT

Average Treatment Effect: need counterfactuals for all units



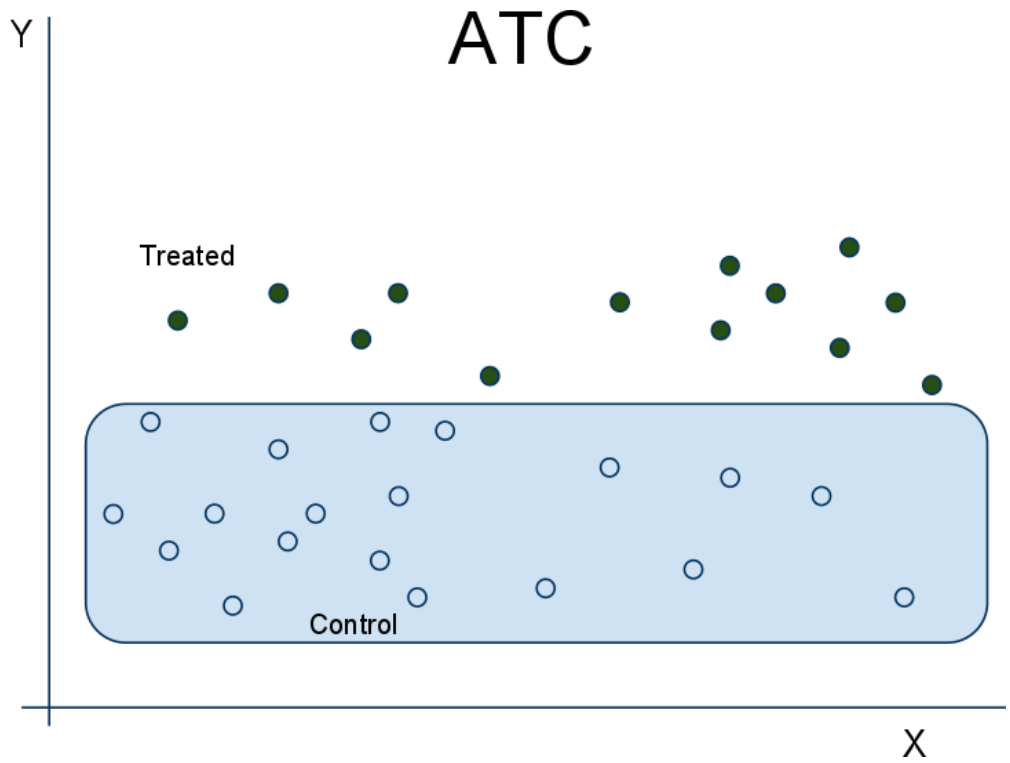
ATT vs. ATE vs. ATC vs. FSATT

Average Treatment Effect on the Treated: need counterfactuals for all treated units



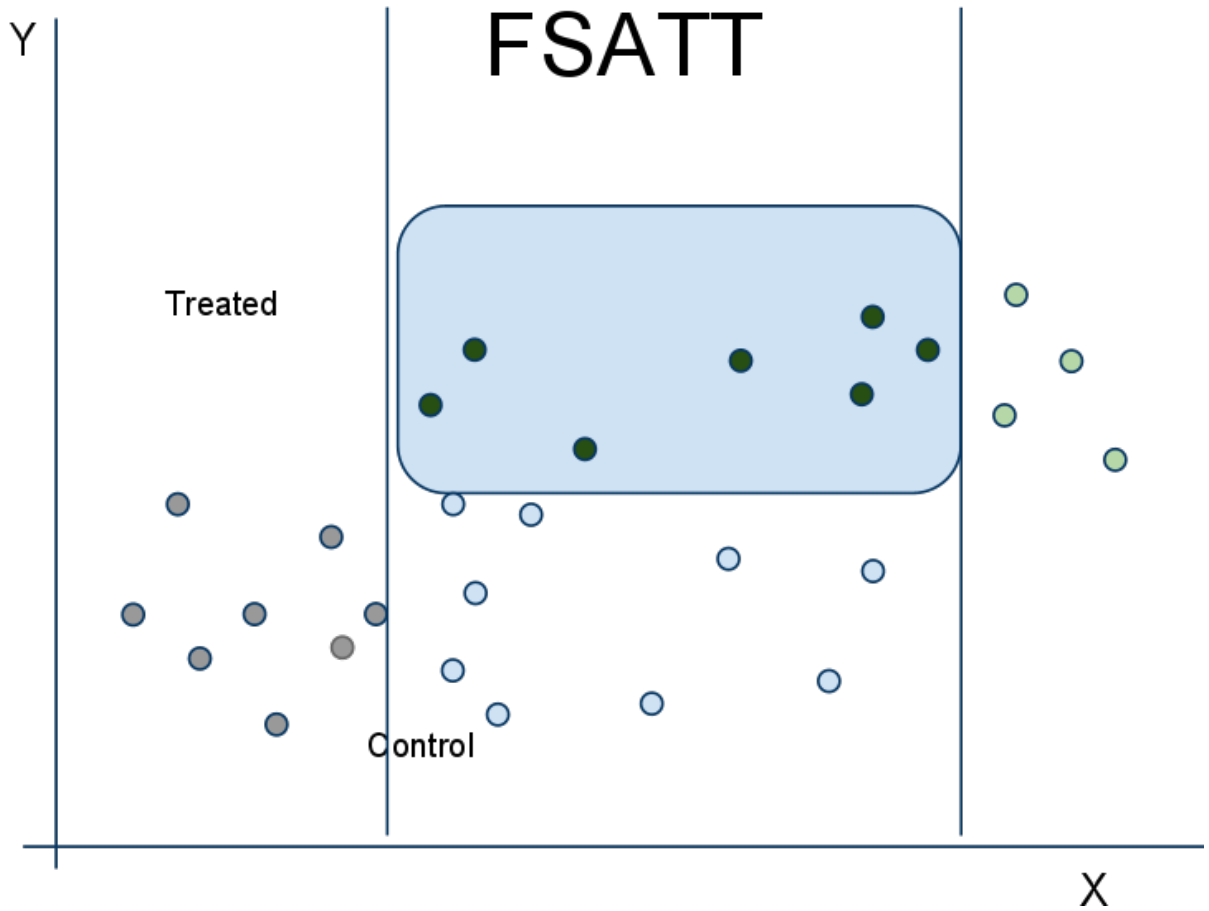
ATT vs. ATE vs. ATC vs. FSATT

Average Treatment Effect on the Controls: need counterfactuals for all control units Think of this as the group that we need to find counterfactuals for.



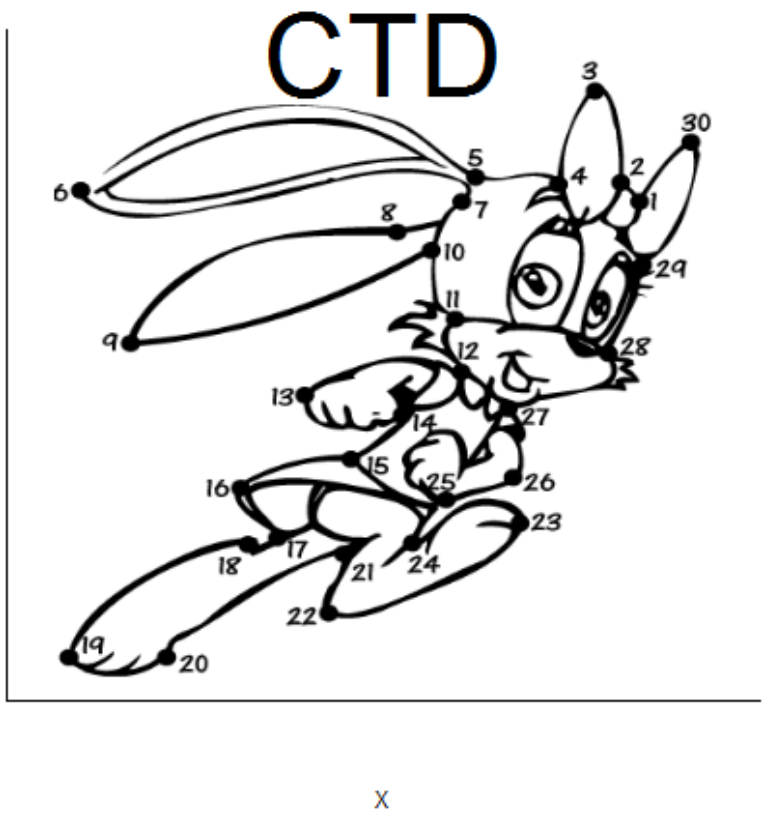
ATT vs. ATE vs. ATC vs. FSATT

Feasible Sample Average Treatment Effect on the Treated: need counterfactuals for all units which we can feasibly match to other units



CTD

Connect the dots



Yes, I did an online connect the dots puzzle while prepping these slides

Yes, I did screw up as I was doing the puzzle which apparently upset this cartoon bear:

Confounding (with *measured* covariates)

What if we do not have unconfoundedness?

$$P(T|Y(0), Y(1), X) = P(T|X)$$

- **Consequence**: imbalance between treated and control units in X
- **Possible solution**: matching

An example with confounding

Suppose that those who take the treatment are systematically different than those who don't take the treatment.

	T=1 (Treatment)	T=0 (Control)
$E[Y\|T=t,X=0]$	-6	2
% Data	.15	.40
$E[Y\|T=t,X=1]$	12	4
% Data	.35	.10

In this case, $X = 0$ strata responds negatively to treatment and $X = 1$ strata responds positively to treatment. Moreover, it appears that those who will be negatively affected by treatment are opting for control.

General Matching Strategy

- 1. Condition on observed, pretreatment variables such that treatment assignment is uncorrelated with potential outcomes conditional on those covariates
- 2. Match data according to the strata defined by the values of these variables
- 3. Assess our matching procedure (check balance)

General Matching Strategy

- 4a. Recombine these strata-specific causal effects into an overall treatment effect by appropriately weighting
- 4bi. Proceed with parametric analysis (regression, t-test, etc.)
- 5. Sensitivity testing for either the confoundedness assumption or the parametric model.

Matching and Causal Inference

If you're interested in the causal effect of A on B, will the results of your parametric analysis be interpretable as a causal effect if you match?

NO $_t$

unless the covariates that you use to match perfectly characterize the data generation process

The purpose of matching for causal inference is to make your treatment seem as if it were truly randomly assigned

Get ready to be disappointed...

If you have unmeasured confounders: matching $\neq$ causal

If you have an omitted variable: matching $\neq$ causal

If you have the wrong functional relationship between confounders and treatment: matching $\neq$ causal

So when can matching help us to make causal statements?

Now the good news...

There's still instances where matching helps you make causal statements:

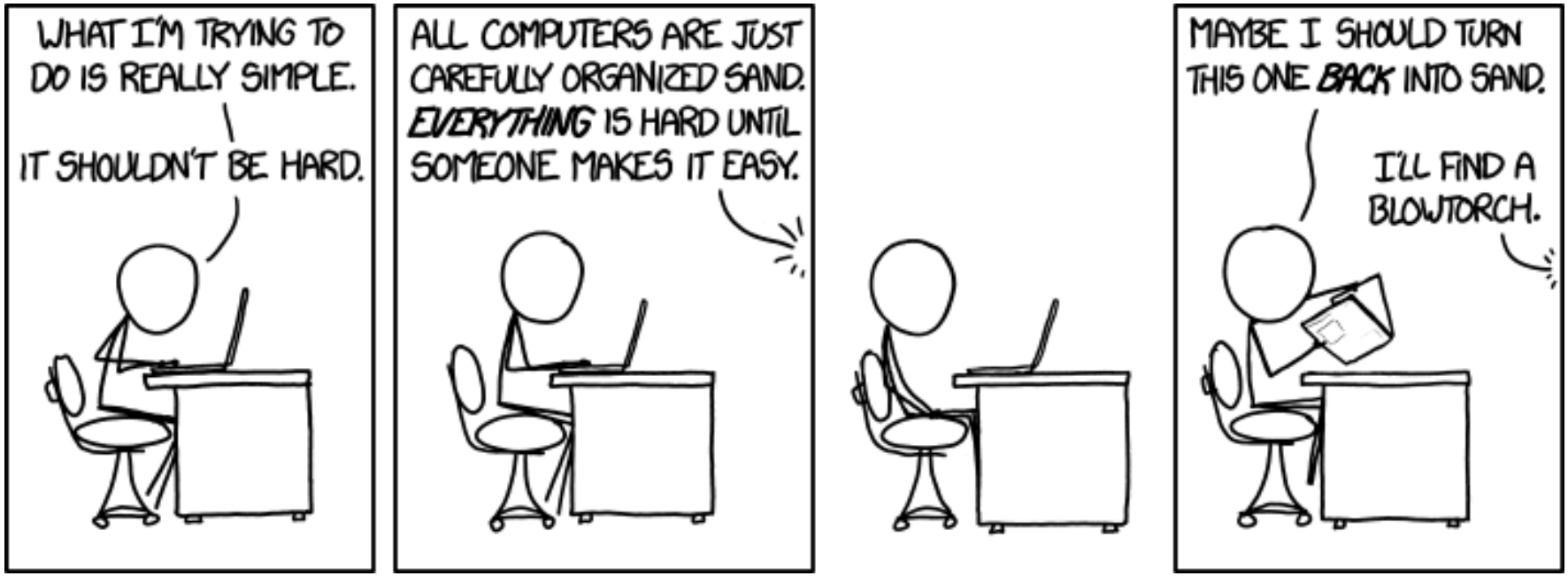
- You ran an experiment, and your randomization didn't work perfectly but you know where it went wrong
- You know all covariates which predict treatment, and you have them measured

How matching is still useful even is you're not making causal statements:

- Matching can help alleviate model dependence
- Matching can help deal with outliers
- Matching can help you understand the convex hull of your data and help you avoid extrapolating outside of it

Next week: Matching in R, balance checking, multiple equation models

Questions?


WHAT I'M TRYING TO DO IS REALLY SIMPLE.
IT SHOULDN'T BE HARD.
ALL COMPUTERS ARE JUST CAREFULLY ORGANIZED SAND.
EVERYTHING IS HARD UNTIL SOMEONE MAKES IT EASY.
MAYBE I SHOULD TURN THIS ONE BACK INTO SAND.
I'LL FIND A BLOWTORCH.