Week 10: Duration Models, Censoring and Truncation, Selection Models

Applied Statistical Analysis II

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Spring 2023

Roadmap through Stats Land

Where we've been:

- Over-arching goal: We're learning how to make inferences about a population from a sample
- Last time: We learned how to conduct a linear regression when our outcome is a count (Poisson)

Today we have a lot to cover! ☺

- Estimate & interpret a duration/survival model
- Intuition behind selection models
  - Censored data, Tobit
- Heckman
  - 2SLS

Introduction to Duration/survival models

- Survival distributions and hazard functions
- Kaplan-Meier curves and Cox regression

Unfortunately, we won't have time for

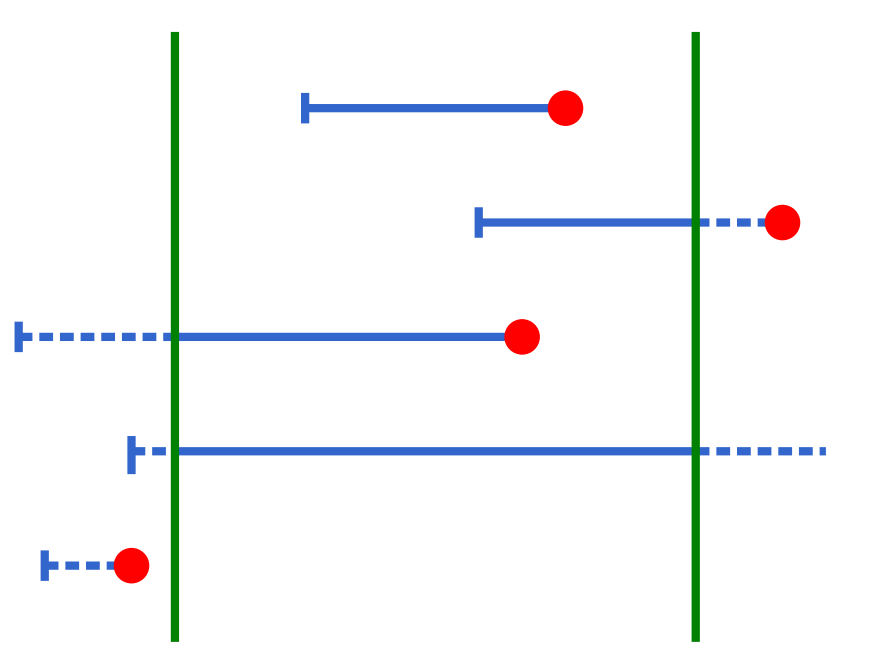
- Time-varying covariates and time-dependent effects
- Weibull, Log-normal, etc.

Introduction to Duration/survival models

- We're interested in time-to-event or survival data with well-defined start and end points
- Sometimes observation stops before event occurs, and waiting time is right-censored, so all we know is that $T > t$
- We can also have delayed entry: observation starts when process is ongoing and we treat waiting time as left-truncated, working with $T|T > t_0$
- We could also have both
- R uses function `Surv( )` for these models

Survival data

Often we have a window of observation and can use a cohort or period sample



Note which episodes are included in each sampling frame, and which are right-censored, left-truncated, or both

Survival data: Some definitions

- Survival time is the period from a start time to when an event of interest occurs
- Three elements must be defined
  - Time origin
  - Scale that defines time periods
  - An observable event
- Outcome of interest in survival models is time to event
- A key problem in some settings is difficulty of observing exact time of event

Tools: Kaplan-Meier Curve

- Basic idea: count # of "exits" at some point of interest and divide by # still unterminated
- Graph looks like stairs
- KM is most useful when comparing two conditions, i.e. treatment and control
  - Ex: Hawthorne et al.(1992) conducted a randomized clinical trial of 67 ulcerative colitis (inflammatory bowel disease that affects the large intestine and rectum) patients with 2 months of remission while taking azathiprine
  - Randomization: continue with azathioprine or placebo
  - Following figure shows that at every time-point treatment group has a higher proportion in remission

Ex: Kaplan-Meier Curve

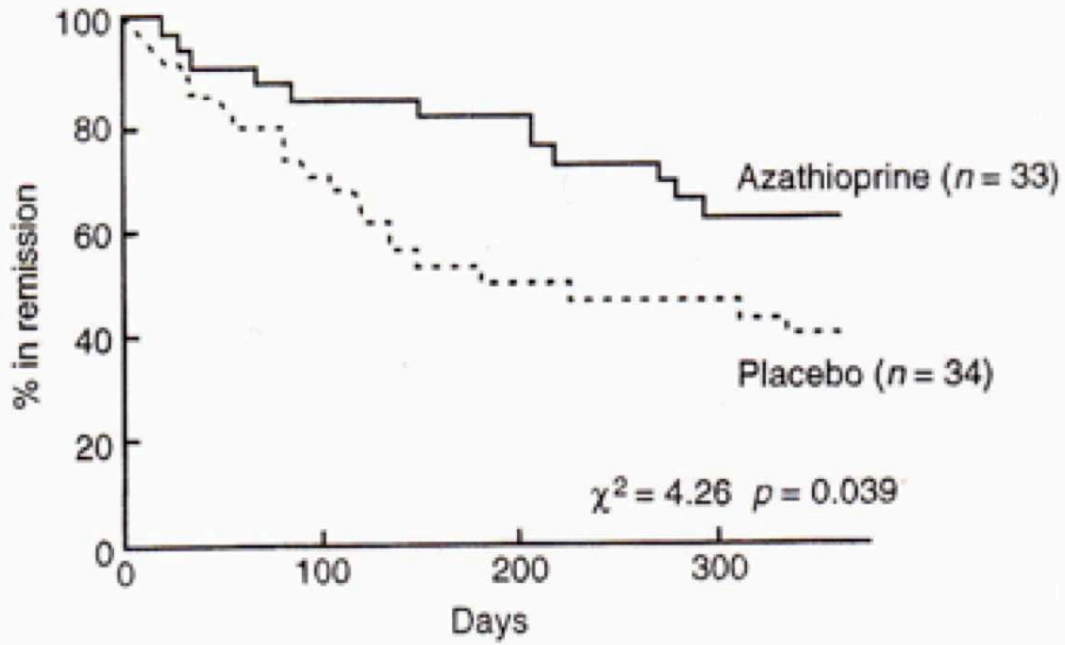


Figure 10.3 Kaplan–Meier survival curves for time from randomisation to recurrence of ulcerative colitis in 67 patients who had achieved remission by initially taking azathioprine. From Hawthorne et al (1992). Randomised controlled trial of azathioprine withdrawal in ulcerative colitis. *British Medical Journal*, **305**, 20–22: reproduced by permission of the BMJ Publishing Group

Kaplan-Meier Overview

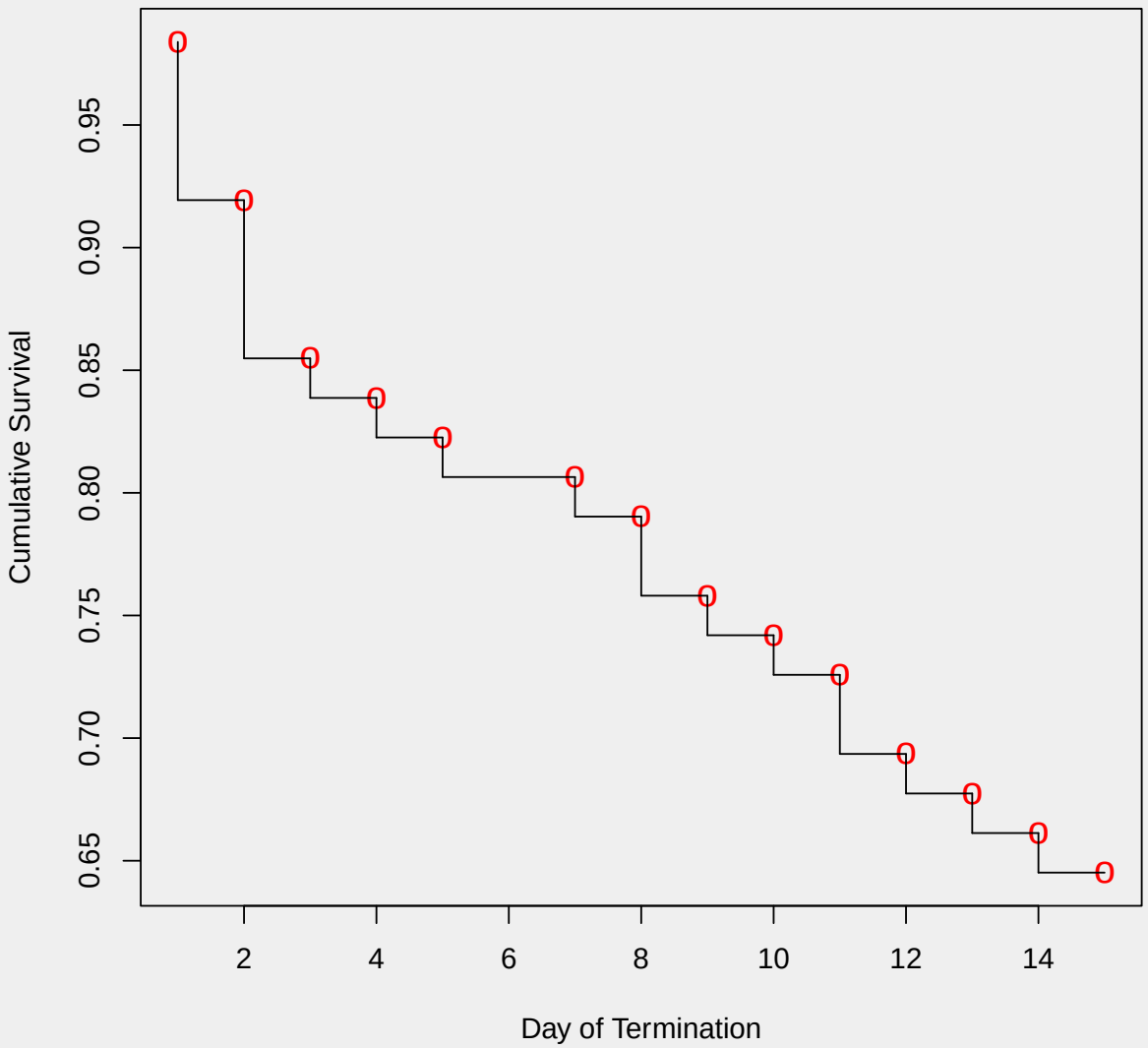
- If there are no “left-censored” cases then the KM curve for n subjects starts at time 0 with value 1 (100% alive)
- It then continues horizontally until first event, dropping by k/n, for k deaths at that time
- This continues until “curve” hits 0 or study ends
- Value k is recorded when measured time is less granular, say days instead of minutes, for time of death
- If there are censored values in the middle of recorded period (these are not included in k), then these only affect denominator n, giving uneven stairstep values

Ex: Government Duration

Create data and plot in R

```
# create "time periods"
tj <- seq(1,15,length=15)[-6]
# NO EVENT AT TIME PERIOD 6
# incidents
d <- c(1,4,4,1,1,1,1,2,1,1,2,1,1,1)
# total
n <- c(62,61,57,53,52,51,50,49,47,46,45,43,42,41)
# calculate risk
risk <- 1 - d/n
# calculate cumulative prod
survivor <- cumprod(risk)
# open plot
pdf("../graphics/govt_KM.pdf")
plot(tj,survivor,pch="o",col="red",cex=1.3,xlab="Day of Termination",ylab="Cumulative
     Survival")
# for each t, connect dots
for (i in 2:length(tj)){
  segments(tj[i-1],survivor[i-1],tj[i-1],survivor[i])
  segments(tj[i-1],survivor[i],tj[i],survivor[i])
  }
dev.off()
```

Ex: Government Duration



Limitations of Kaplan-Meier Models

- Descriptive rather than inferential
- Does not include effects of explanatory variables
- Requires categorical outcomes
- Omits time-dependent effects

Essential concept: Survival function

Survival function is probability that event has not occurred by time t

$$S(t) = Pr(T > t)$$

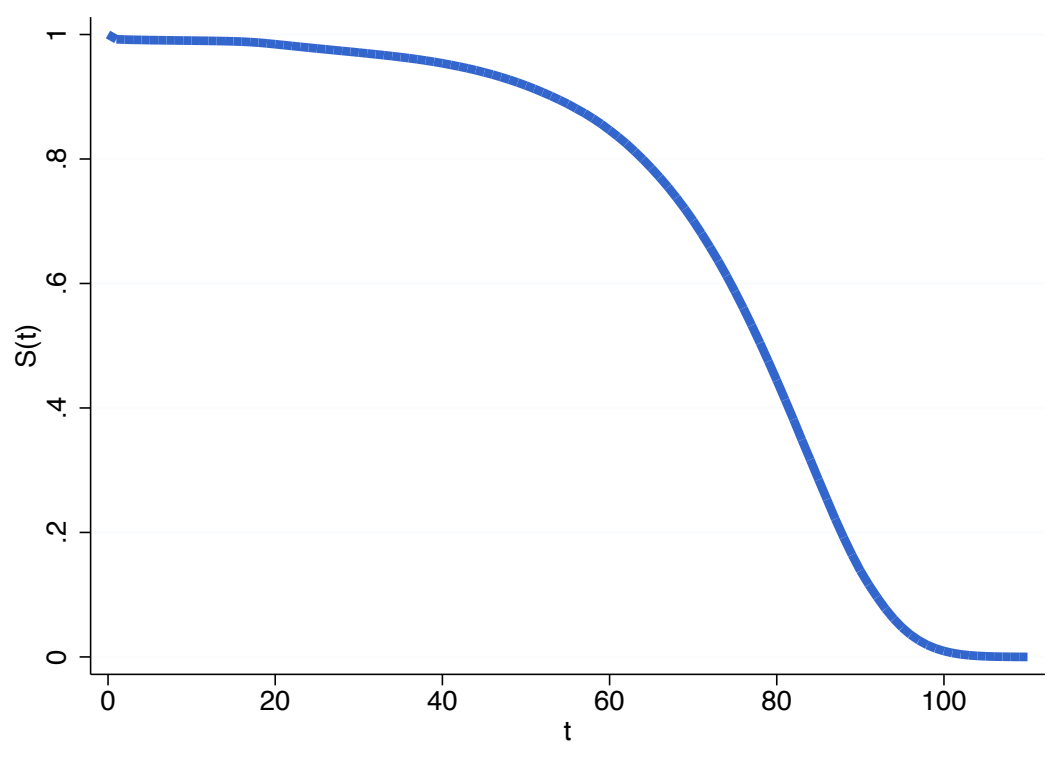


Figure: Survival function for U.S. males.

Essential concept: Density function

Density, or unconditional frequency of events by time, tells us how quickly survival drops over time

$$f(t) = lim_{d_t \downarrow 0} Pr(T \in (t, t + d_t))/d_t = -S'(t)$$

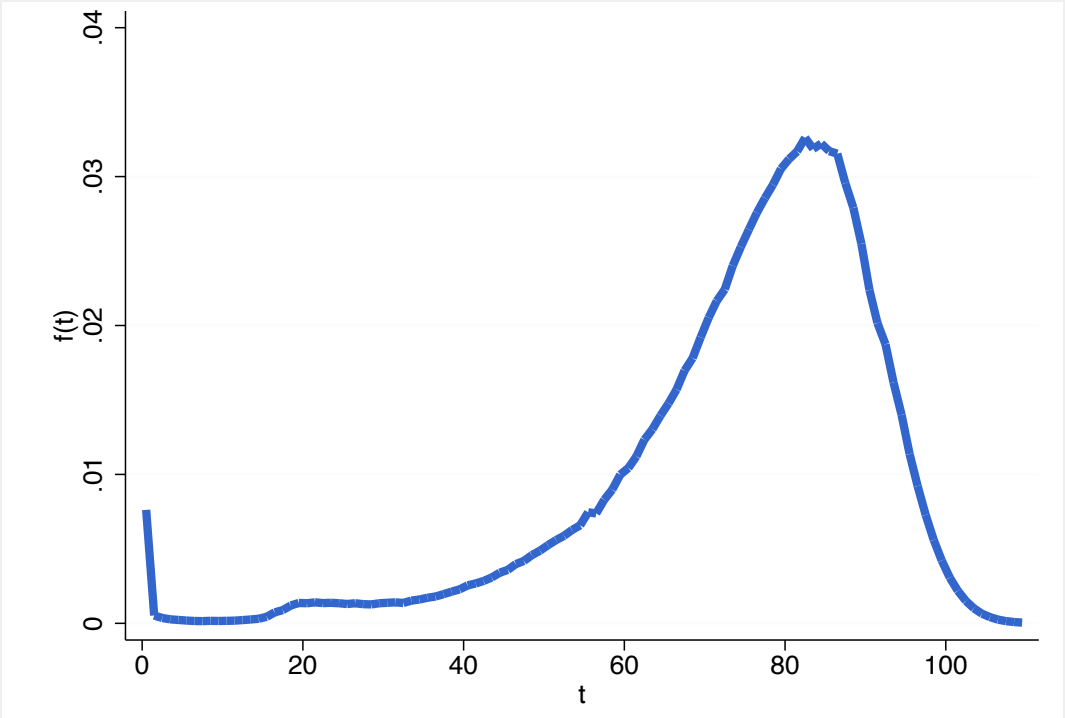


Figure: Density of U.S. male deaths by age.

Essential concept: Hazard function

Hazard is conditional event rate among people at risk

$$\lambda(t) = lim_{d_t \downarrow 0} Pr(T \in (t, t + d_t) | T > t)/d_t = f(t)S(t)$$

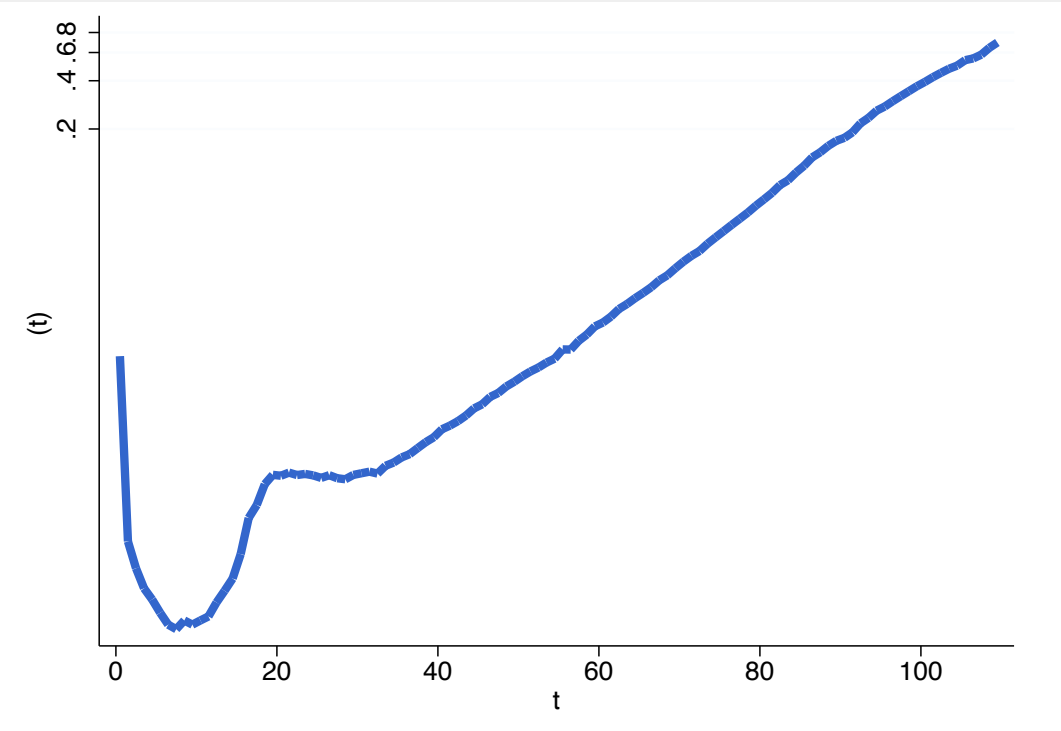


Figure: U.S. death rates by age, plotted in log scale.

From Risk to Survival

- Survival time can be characterized by any of these functions
- For example we can go from hazard to survival

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{d_t}\log S(t)$$

- Then integrate both sides using $S(0) = 1$ as a boundary condition to obtain

$$S(t) = exp[-\int_0^t \lambda(u)du]$$

- Integral is called *cumulative hazard* and is denoted $\Lambda(t)$
- Example: If hazard is constant, $\lambda(t) = \lambda$ then cumulative hazard is $\Lambda(t) = \lambda t$ and survival function is $S(t) = exp[-\lambda t]$, an exponential distribution

Survival distributions and models

- Many to select from: Weibull, Gompertz, Gamma, Log-normal, Log-logistic
- There are four ways to introduce covariates in parametric survival models
  1. Parametric families, where parameters of a distribution, such as λ and p in a Weibull, depend on covariates
  2. Accelerated life, where log of survival time follows a linear model
  3. Proportional hazards, where log of hazard function follows a linear model
  4. Proportional odds, where logit of survival function follows a linear model

Hazard Ratio For Comparing Two Groups

- Observe O_A and O_B
- H_O: no difference between groups, H_A: groups are different
- Calculate E_A and E_B, as before: $E_A = \sum_T e_{A_i}$, $E_B = \sum_T d_i - E_A$
- O_A/E_A is relative death rate in group A, and O_B/E_B is relative death rate in group B
- Hazard Ratio is: $H_R = \frac{O_A/E_A}{O_B/E_B}$, which is near 1 under null hypothesis of no difference

Proportional Hazards

- Also called “relative risk”
- Most widely used general survival regression specification
- It starts with a baseline or underlying hazard function that all units share, $h(t)$
- Then predictors in form of covariates and associated coefficients act on each units hazard through $exp(X\beta)$, which is typically called relative hazard function
- General form of regression specification is:

$$h(t|X) = h(t)exp(X\beta)$$

- An appropriately bounded parametric hazard function can be used for baseline hazard function
- This week we parametrically define $h(t)$ but can leave it unspecified

Proportional Hazards Interpretation

- Consider how to interpret *j*th coefficient from proportional hazards model
- Recall elegant interpretation of linear model regression coefficients?
- Regression coefficient for $X_j(\beta_j)$ is increase in log hazard rate at some fixed point in time t if X_{ji}s increased by one unit and all other covariates are held constant at observed means
- This can be re-expressed as:

$$exp(\beta_j) = \left[\frac{t|X_1, X_2, ..., [X_{j+1}], X_{j+1}, ...X_k)}{t|X_1, X2_, ..., [X_j], X_{j+1}, ..., X_k)}\right]$$

Proportional Hazards Interpretation

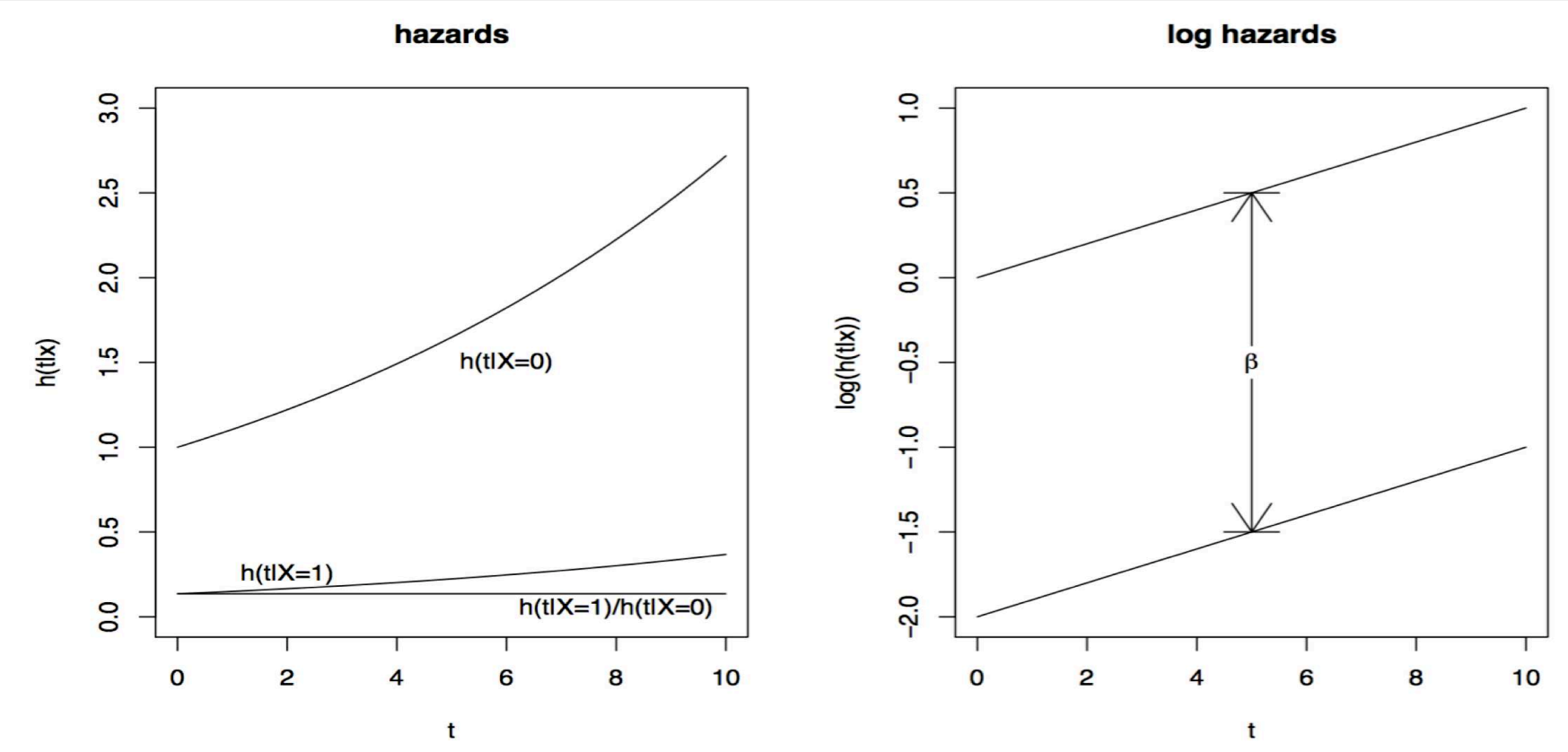
- Suppose X_1 is treatment indicator: zero for control ($X_1 = 0$), one for treatment ($X_1 = 1$) with no other explanatory variables
- In this case, proportional hazards model with no intercept is given for two cases by:

$$h(t|X_1 = 0) = h(t)$$

$$h(t|X_1 = 1) = h(t)exp(\beta_1)$$

- We can also express same general effect holding other explanatory variables constant

Proportional Hazards Interpretation



Surv Function In Detail

- Surv creates a survival object, which is response variable in an R model formula
- Argument matching is important for this function since it gives structure to the outcome variable
- Syntax: `Surv(time, time2, event, type=c('right', 'left', 'interval', 'counting','interval2', 'mstate'), origin=0)`
- `time`: For continuous data this is starting time for interval

Surv Function In Detail

- `event`: 0/1 or 1/2, not-occured/occured
  - For interval censored data, status indicator is 0=right censored, 1=event at time, 2=left censored, 3=interval censored
  - Where 1/2 coding is used and all subjects are censored model will be wrong (no 2's to use)
  - If you want to be extra safe about coding use something like `Surv(time, status==3)` where 3 indicates event
- `time2`: Ending time of the interval for interval censored or counting process data only
- `origin`: for counting process data, hazard function origin, not commonly used

Surv Function In Detail

- `type`: a character string specifying type of censoring: “right”, “left”, “counting”, “interval”, “interval2”, or “mstate”
- For “mstate” status variable will be treated as a factor where first indicates censoring and remaining values are transitions to given state
- When `type` argument is assumed that
  - If there are two unnamed arguments, they match `time` and `event` in that order
  - If there are three unnamed arguments, they `matchtime`, `time2`, and `event` in that order
  - If `event` variable is a factor then type `mstate` is assumed, otherwise type`right` if there is no `time2` argument, and type `counting` if it is present
- Due to these rules, `type` argument is often not used

Surv Function In Detail

- Example from the R help file: `Surv(heart$start, heart$stop, heart$event)`

```
[1] (  0.0,  50.0] (  0.0,   6.0] (  0.0,   1.0+] (  1.0,  16.0] (  0.0,  36.0+]
```

- Notice use of brackets in conventional mathematical sense
- First of pair of #s is start time in the Stanford Heart Transplant study
- Second of pair of #s is either exit time (death), or right-censoring time denoted by the "+"

Surv Function In Detail

- So consider case #5 in the data

```
heart[5,]
start stop event       age      year surgery transplant id
5     0   36     0 -7.737166 0.4900753       0          0  4
```

- In Surv output:

```
Surv(heart$start, heart$stop, heart$event)[5]
[1] (0,36+]
```

- They started at zero, exited at 36 but had no event, so they must be censored

Ex: Male Mortality in 1800s

- Males born in years 1800-1820 and surviving at least 40 years in parish Skellefte in northern Sweden are followed from their fortieth birthday until death or sixtieth birthday, which ever comes first
- 2058 observations with 6 variables
- `id`: personal identification number
- `enter`: start of duration in years since the fortieth birthday
- `exit`: end of duration in years since the fortieth birthday
- `event`: logical vector indicating death at end of interval
- `birthdate`: birthdate in decimal form
- `ses`: socio-economic status, a factor with levels lower (565), upper (643)

Ex: Male Mortality in 1800s

```
data(mort)
dim(mort)
```

```
[1] 1208    6
```

```
mort[1:5,]
```

```
id  enter   exit event birthdate   ses
1   1  0.000 20.000     0  1800.010 upper
2   2  3.478 17.562     1  1800.015 lower
3   3  0.000 13.463     0  1800.031 upper
4   3 13.463 20.000     0  1800.031 lower
5   4  0.000 20.000     0  1800.064 lower
```

Cox P-H Regression in R

- Note that `coxreg` is a "wrapper" for `coxph` in survival package, but gives different printing of results
- `survfit` gives summaries from a fit `codeph` object, including estimates of $S(t)$ at mean values of the covariates and plot method for objects estimated survival function, along with 95 % confidence bands
- A requirement with these functions is that we stipulate a survival object on LHS of the model specification with function Surv
- First, let's run a model with only two explanatory variables...

Cox P-H Regression in R

```
# estimate Cox PH model
add_surv <- coxph(Surv(enter, exit, event) ~ ses +
    birthdate, data=mort)
summary(add_surv)
```

```
Call:
coxph(formula = Surv(enter, exit, event) ~ ses + birthdate, data = mort)

n= 1208, number of events= 276

coef exp(coef) se(coef)      z Pr(>|z|)
sesupper  -0.48414   0.61623  0.12074 -4.010 6.08e-05 ***
birthdate  0.02876   1.02918  0.01080  2.663  0.00773 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95
sesupper     0.6162     1.6228    0.4864    0.7808
birthdate    1.0292     0.9716    1.0076    1.0512

Concordance= 0.585  (se = 0.017 )
Likelihood ratio test= 23.08  on 2 df,   p=1e-05
Wald test            = 22.96  on 2 df,   p=1e-05
Score (logrank) test = 23.3  on 2 df,   p=9e-06
```

Cox P-H Regression in R

```
interact_surv <- coxph(Surv(enter, exit, event) ~ ses *
    birthdate, data=mort)
summary(interact_surv)
```

```
Call:
coxph(formula = Surv(enter, exit, event) ~ ses * birthdate, data = mort)

n= 1208, number of events= 276

coef  exp(coef)   se(coef)      z Pr(>|z|)
sesupper           -5.351e+01  5.742e-24  3.977e+01 -1.346    0.178
birthdate           1.624e-02  1.016e+00  1.417e-02  1.146    0.252
sesupper:birthdate  2.926e-02  1.030e+00  2.194e-02  1.334    0.182

exp(coef) exp(-coef) lower .95 upper .95
sesupper           5.742e-24  1.741e+23 8.117e-58 4.062e+10
birthdate          1.016e+00  9.839e-01 9.885e-01 1.045e+00
sesupper:birthdate 1.030e+00  9.712e-01 9.864e-01 1.075e+00

Concordance= 0.585  (se = 0.017 )
Likelihood ratio test= 24.87  on 3 df,   p=2e-05
Wald test            = 22.82  on 3 df,   p=4e-05
Score (logrank) test = 23.72  on 3 df,   p=3e-05
```

Assessing model quality

With one statement we can LRT both explanatory variables

```
drop1(add_surv, test = "Chisq")
```

```
Surv(enter, exit, event) ~ ses + birthdate
Df    AIC     LRT  Pr(>Chi)
<none>       3687.3
ses         1 3701.4 16.1095 5.978e-05 ***
birthdate   1 3692.6  7.2752  0.006991 **
```

Equivalently

```
2*(coxph(Surv(enter, exit, event) ~ ses + birthdate, data=
    mort)$loglik[2] - coxph(Surv(enter, exit, event) ~
    birthdate, data=mort)$loglik[2])
```

```
[1] 16.10947
```

Advice On Model Selection

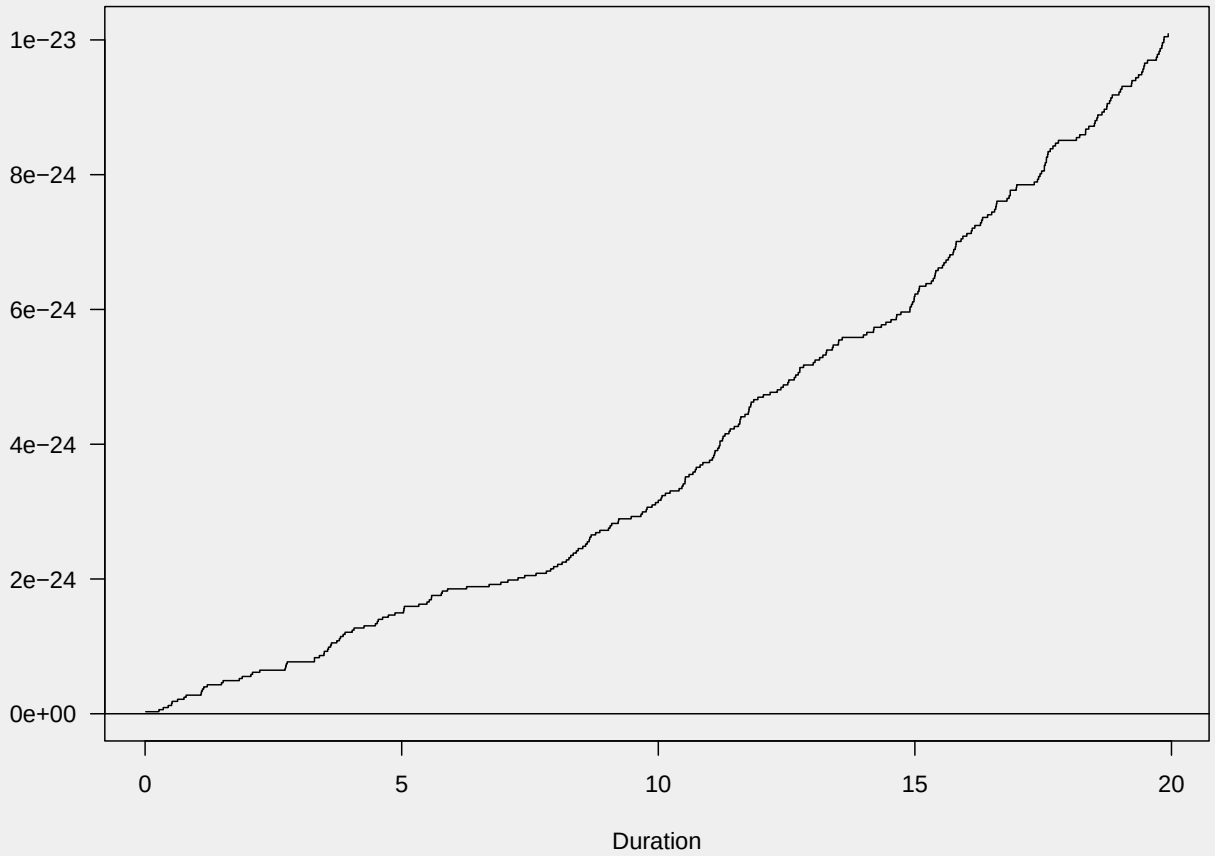
- It's okay to keep explanatory variables in the model if they aren't reliable predictors of the outcome variable if they help in some other way
- All models with non-identity link functions imply interactions (input does not yield same output, some transformation)
- Deviance comparisons are generally the best way to compare nested models
- R^2 does not mean much

Plotting cumulative hazard function

You must use coxreg not coxph to do this

```
plot_CoxPH <- coxreg(Surv(enter, exit, event) ~ ses +
    birthdate, data=mort)
```

```
plot(plot_CoxPH)
```



What we didn't cover

- Time-varying covariates and time-dependent effects
- Connection to Poisson (offset)
- Using other distributions; Weibull, log-normal
- Checking assumptions; parallel proportionality

Intuition of Selection Models: Truncation

- Imagine that we have a random variable, y, with a normal distribution
- However, we don't observe distribution of y below some value a, because distribution is truncated
- We need to add an additional term to our regression model to correctly estimate our βs
- If we don't include term that adjusts for probability that we see y to begin with, we'll get an inconsistent estimate
  - Similar to what would happen if we omitted a relevant variable (more on this later)

Intuition of Selection Models: Censored Data

- Suppose that instead of data being simply missing, we observe value of $y_i = a$, whenever $y_i < a$
- We can imagine a latent variable, y_i^*, but we observe a whenever $y_i^* \leq a$, and y_i otherwise
  - Remember from last week, this is a case of "left-censoring"
- If censoring cut point $\neq$ 0, and if we also have an upper-censoring cut point, then conditional expectation ofy_i will be more complicated

Intuition of Selection Models: Tobit Models

- Tobin wanted to explain relationship between household income and household luxury expenditures
  - Noticed large concentration of households who spend exactly $0 on luxury goods
- **Problem:** Need to account for fact that explanatory variable might influence probability of whether a household spent $0 on luxury items and how much they actually spent, given that they spent something

Setup: Tobit Regression models

Censored from of y_L when latent variable $y_j^* \leq y_L$

Let Φ be standard normal cumulative distribution function and φ be standard normal probability density function

$$\sum_{y_j > y_L} \log\left(\frac{1}{\sigma}\varphi\left(\frac{y_j - X_j\beta}{\sigma}\right)\right) + \sum_{y_j = y_L} \log\left(\Phi\left(\frac{y_L - X_j\beta}{\sigma}\right)\right)$$

- We know Y_is are distributed normally with a mean of βX and a variance of σ^2
- We know density function of a normally distributed variable
- This equation is what results when you substitute that density function into joint probability density function of Y's and take log of equation

Interpreting Tobit coefficients

- $X_i'\beta$ = expected value of underlying latent variable (Y^*)
- β = effect of a change in given x on expected value of latent variable, holding all other x constant
- Expected value of observed y is equal to relevant coefficient weighted by probability that an observation will be uncensored
- Can also calculate expected value of y conditional on y exceeding censoring threshold

Ex: School Aptitude Tests

- Consider academic aptitude (scaled 200-800) which we want to model using
  - Reading and math test scores
  - Type of program students are enrolled in (academic, general, or vocational)
- **Problem**: Students who answer all questions on academic aptitude test correctly receive a score of 800, even though it's likely that these students are not "truly" equal in aptitude
  - Same is true of students who answer all of the questions incorrectly
  - All such students would have a score of 200, although they may not all be of equal aptitude

Ex: Tobit in R

```r
# load data
aptitude <- read.csv("https://stats.idre.ucla.edu/stat/data/tobit.csv")
aptitude$prog <- factor(aptitude$prog, labels = c('acad', 'general', 'vocational'))
```

```r
# run using AER
tobit_aer <- AER::tobit(apt ~ read + math + prog,
  data = aptitude, left = -Inf, right = 800
)
# using VGAM
tobit_vglm <- vglm(apt ~ read + math + prog, tobit(Upper = 800), data = aptitude)
# using surv
tobit_surv <- survreg(Surv(apt, apt < 800, type = 'right') ~ read + math + prog,
                      data = aptitude, dist = 'gaussian')
```

Ex: Tobit in R

```r
# tobit log-likelihood func
tobit_ll <- function(par, X, y, ul = -Inf, ll = Inf) {
  # this function only takes a lower OR upper limit
  # parameters
  sigma = exp(par[length(par)])
  beta  = par[-length(par)]

  # create indicator depending on chosen limit
  if (!is.infinite(ll)) {
    limit = ll
    indicator = y > ll
  } else {
    limit = ul
    indicator = y < ul
  }

  # linear predictor
  lp = X %*% beta

  # log likelihood
  ll = sum(indicator * log((1/sigma)*dnorm((y-lp)/sigma)) ) +
    sum((1-indicator) * log(pnorm((lp-limit)/sigma, lower = is.infinite(ll))))
  # return -log like func
  -ll
}
```

Ex: Interpret Tobit Results

(Intercept)	209.57^{***}
	(32.77)
read	2.70^{***}
	(0.62)
math	5.91^{***}
	(0.71)
$\text{prog}_{general}$	-12.71
	(12.41)
$\text{prog}_{vocational}$	-46.14^{***}
	(13.72)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

- For a 1 unit $\uparrow$ in read, there is ≈ 2.7 $\uparrow$ in predicted aptitude
- For a 1 unit $\uparrow$ in math, there is ≈ 5.91 $\uparrow$ in predicted aptitude
- Predicted aptitude is -46.14 $\downarrow$ for students in a vocational program than for students in an academic program

Tobit Model Limitations

- Same set of variables, same coefficients determine both Pr(censored) and outcome
- Lack of theory as to why observations are censored
- Selection models allow us to:
  - Different variables and coefficients in censoring (selection equation) and DV (outcome equation)
  - Theory of censoring, observations are censored by some variable Z
  - Take account of censoring process because selection and outcome are not independent

Heckman Selection Models

- Imagine you want to estimate impact of participating in a community group on one's level of satisfaction with democratic institutions
  - Perhaps taking part in some community organization $\uparrow$ one's trust in government and $\uparrow$ satisfaction with government performance
  - Taking part in a community organization is going to be our "treatment effect"
- We want to see how this treatment could affect outcomes for "average" person

Heckman Selection Models

- **Problem:** People participating in community groups are unlikely to be same as "average person"
- May be something that leads them to take part in community initiatives and, if this unobserved characteristic also makes them likely to trust government, we'll bias our estimates of treatment effect
- Analogous to one of truncation
  - Here, truncation is "incidental"
  - It does not occur at given, definite values of y_i, but instead cuts out people for whom underlying attractiveness of community service did not reach some critical level

Interpreting coefficients: Heckman

- If covariate appears ONLY in outcome equation, coefficient can be interpreted as marginal effect of 1 unit change in that covariate on Y
- However, if covariate appears in both selection and outcome equations, coefficient in outcome equation is affected by its presence in selection equation (we'll see in a minute)

Heckman Ex: Women's Pay (Mroz 1987)

- 1975 Panel Study of Income Dynamics (PSID) on married women's pay and labor force participation
  - Age, # of children and years of education
- Want to estimate a wage equation for married women in 1975
- However, wages are only observed for women that participate in the labor-force (possible selection bias)
  - Estimated effects in wage equation may suffer from a large bias since substantial proportion of married women in 1975 did not participate in labor-force

Women's Pay: In R

First, let's just estimate a base OLS with log(wage) as outcome, and education, experience, and city/rural as our predictors

```
# load data from sampleSelection package
data("Mroz87")
# simple OLS, first using sample of
    labor-force participants, only for
    those with observed wage
# lfp = 1 (labor force paticipant)
base_heck <- lm(log(wage) ~ educ + exper
     + city + kids5 + huswage, data=
    subset(Mroz87, lfp==1))
```

(Intercept)	-0.44^{*}
	(0.19)
educ	0.10^{***}
	(0.02)
exper	0.02^{***}
	(0.00)
city	0.02
	(0.07)
kids5	-0.05
	(0.08)
huswage	0.02
	(0.01)
R^2	0.16
Adj. R^2	0.15
N	428

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Women's Pay: In R

Now, let's use 2-step estimation with labor force selection

- Probit equation to estimate selection process (who's in labor force)
- Result from that equation are used to construct a variable that captures selection effect in wage equation
- Selection equation should include predictors variables likely affect whether married women would be in labor force

```
full_heck <- heckit(lfp ~ educ + exper + city + kids5 + huswage,
       log(wage) ~ educ + exper + city + kids5 + huswage, data=
    Mroz87 )
```

Women's Pay: In R

(Intercept)	−0.44*	−1.71***
	(0.19)	(0.29)
educ	0.10***	0.14***
	(0.02)	(0.02)
exper	0.02***	0.05***
	(0.00)	(0.01)
city	0.02	−0.06
	(0.07)	(0.11)
kids5	−0.05	−0.50***
	(0.08)	(0.10)
huswage	0.02	−0.03*
	(0.01)	(0.01)
O: (Intercept)		0.57
		(1.03)
O: educ		0.06
		(0.05)
O: exper		0.00
		(0.02)
O: city		0.04
		(0.08)
O: kids5		0.13
		(0.20)
O: huswage		0.03
		(0.01)
invMillsRatio		−0.61
		(0.60)
sigma		0.79
rho		−0.77
R^2	0.16	0.16
Adj. R^2	0.15	0.15
N	428	753
Censored		325
Observed		428

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

- Notice how all of our covariates are not statistically different than zero in outcome equation
- λ (inverse MIills ratio) is not reliable, selection bias is not a significant issue
- Probably because there's no variation in outcome equation to explain

Tangential Issue: Measurement Error

- Regression model has no trouble with measurement error in outcome; measurement error can be conceived as part of stochastic term
- Measurement error in covariates introduces bias in coefficient estimates:

$$Y = \beta_1 T + \beta_2 X_2 + \epsilon \text{ Population regression}$$

$$X_1 = T + \delta, E(\delta) = 0$$

- Thus, sample regression model is:

$$Y = \beta_1 T + \beta_2 X_2 + (\epsilon - \beta_1 \delta)$$

- β_1 will be attenuated
- β_2 will also be biased, though we don't know direction and magnitude

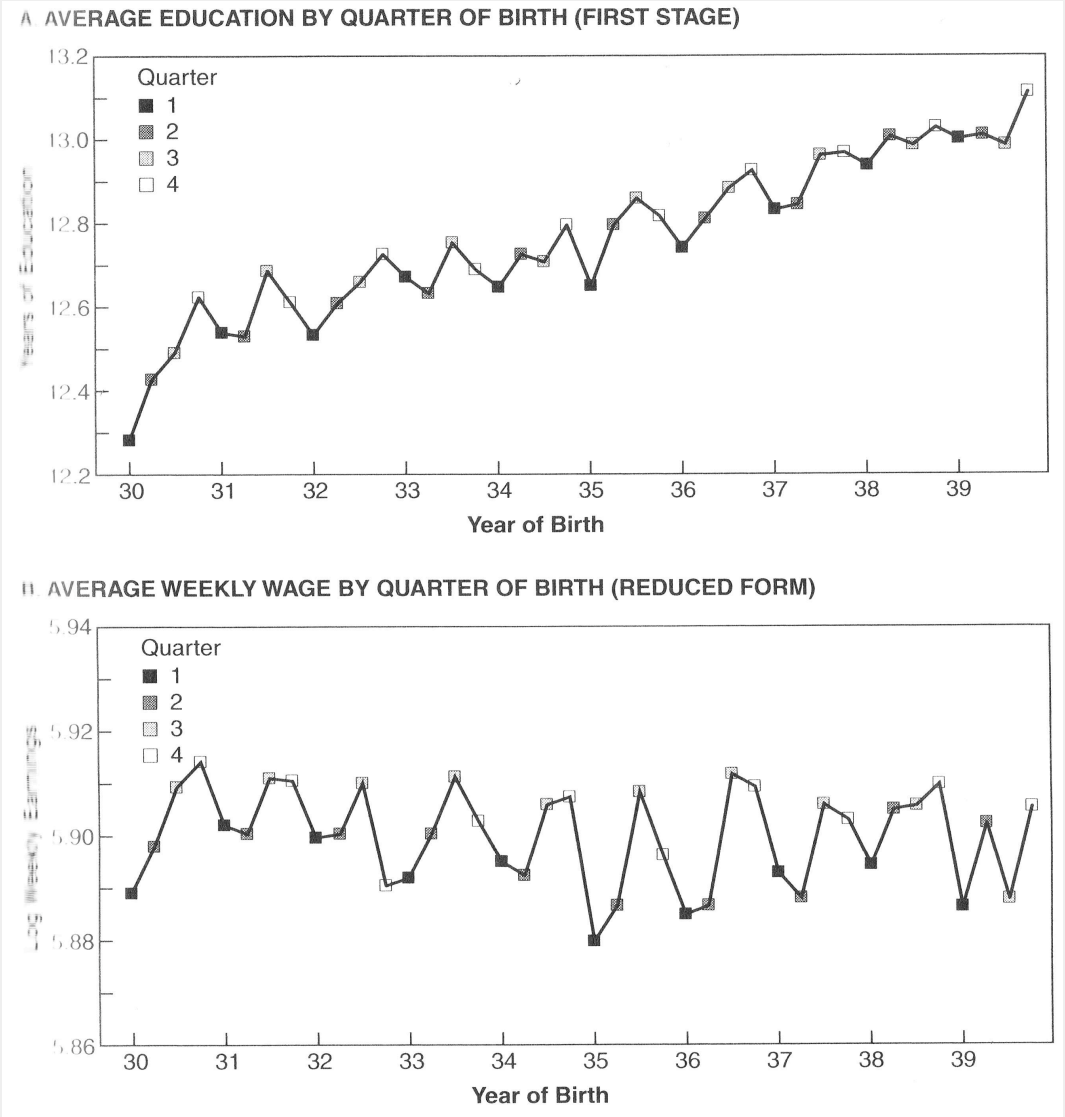
Possible solution: Instrumental variables

- IV are used in simultaneous equations models and to correct bias from measurement error
- IV are also used to solve omitted variable bias (selection on unobservables)
- To motivate the model, assume following causal model

$$Y_i = \alpha + \rho T_i + X_i'\beta + \underbrace{A_i'\gamma + \epsilon}_{\eta_i} \text{ (structural equation)}$$

  - Notice A is unobserved
- To provide a causal interpretation for ρ, we need an instrument Z for T
- Instrument Z must be correlated with T but uncorrelated with A and ϵ (exclusion restriction)

Ex: School attendance → future earnings[1]



[1] Angrist, J. D., & Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics, 106(4), 979-1014.

Instrumental variables: Setup

- Causal model can be represented as:

$$T_i = X'\pi_{10} + Z_i\pi_{11} + \zeta_{1i} \text{ (first stage regression equation)}$$

$$Y_i = X'\pi_{20} + Z_i\pi_{21} + \zeta_{2i}\ )$$

1. Y and T are endogenous variables
2. X are exogenous covariates
3. Z are instruments

- IV estimator of causal effect is

$$\hat{\rho}_{IV} = \frac{\hat{\pi_{21}}}{\hat{\pi_{11}}} = \frac{cov(Y, \widetilde{Z})}{cov(T, \widetilde{Z})}$$

where $\widetilde{Z}$ is residual of $Z \sim X$

Two-Stage Least Squares

- To derive 2SLS estimator of ρ, substitute first-stage equation into structural equation

$$\begin{aligned} Y_i &= \alpha + \rho \underbrace{\left(X'\pi_{10} + Z_i\pi_{11} + \zeta_{1i}\right)}_{T_i} + X_i'\beta + \eta_i \\ &= \alpha + \rho \underbrace{\left(X'\pi_{10} + Z_i\pi_{11}\right)}_{\hat{T}_i} + X_i'\beta + \rho\zeta_{1i} + \eta_i \\ &= \alpha + \rho\hat{T}_i + X_i'\beta + \zeta_{2i} \end{aligned}$$

- 2SLS estimator of ρ is simply OLS estimator of ρ in last equation; T are fitted values of first-stage regression

Two-Stage Least Squares

- Intuition: 2SLS only retains variation in T that is generated "as if at random"
- 2SLS is biased but consistent
- Bias is towards "naive" OLS estimate of ρ in model $\alpha + \rho T_i + X_i'\beta + \zeta_{2i}$
- Bias is worst with "weak instruments"; if "weak instruments" exist, bias worsens with "over identification"
- Careful with estimate of residual variance (use ρT, not $\rho \hat{T}$ in defining η_i)

Ex: Causal impact of FDI on inequality

- Typical design regresses $Gini_t$ on FDI_{t1}, but:
  1. Omitted variables: We do not have fully-specified models of inequality
  2. Selection bias: MNCs locate near pools of highly-trained labor in cities
  3. Reverse causation: MNCs might avoid high-inequality places
- IV design uses "distance to border" as instruments for FDI

$$Y_t = \beta_0 + \beta_1 FDI_{t-1} + \beta_2 Education$$

$$FDI_{t-1} = \gamma_0 + \gamma_1 Education + \gamma_2 distance$$

Ex: Causal impact of FDI on inequality

Let's try simple OLS first

```r
# create data
set.seed(5)
makeBivariateNormalVars <- function(N, mu_x, mu_y,
                                    sigma_x, sigma_y,
                                    rho){
  # Begin with two independent N(0,1) variables
  Z1 <- rnorm(N, 0, 1)
  Z2 <- rnorm(N,0,1)
  # create error terms
  u <- sigma_x*Z1 + mu_x
  v <- sigma_y *(rho*Z1 + sqrt(1-rho^2)*Z2) + mu_y
  # package up errors as list
  return(list(u,v))
}
# get 10k observations
N <- 10000
errorTerms <- makeBivariateNormalVars(10000, 0, 0, 1,
      1, .8)
u <- errorTerms[[1]]
v <- errorTerms[[2]]
z <- rnorm(N, 2,1)
x1 <- 0 + z + v
x2 <- rnorm(N, 0, 1)
y <- 0 + 0.5*x1 + 2*x2 + u
# run "bad"/naive OLS
baseReg <- lm(y ~ x1 + x2)
```

(Intercept)	0.185^{***}
	(0.019)
FDI	0.406^{***}
	(0.008)
Education	0.004
	(0.011)
R^2	0.216
Adj. R^2	0.216
N	10000

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Ex: Causal impact of FDI on inequality

Now, use an IV approach

```
# by hand 1st step
stage1 <- round(lm(y ~ z + x2)$
    coefficient[2], 4)
stage2 <- round(lm(x1 ~ z + x2)$
    coefficient[2], 4)
round(stage1/stage2, 4)
```

```
z
0.0095
```

```
# check using package "ivreg"
ivReg <- ivreg(y ~ x1 + x2| x2 + z)
```

(Intercept)	0.185^{***}	0.975^{***}
	(0.019)	(0.028)
FDI	0.406^{***}	0.010
	(0.008)	(0.012)
Education	0.004	0.001
	(0.011)	(0.012)
R^2	0.216	0.010
Adj. R^2	0.216	0.010
N	10000	10000

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Which instruments are good instruments?

- Condition $T|X \perp \epsilon|X$ does not hold
- We need to find Z that correlates with T but not with ϵ
  1. Relevance of instrument (avoid weak instruments)
  2. Monotonicity
  3. Exclusion restriction

Which instruments are good instruments?

```r
# check assumptions
summary(ivReg, diagnostics =
    TRUE)
```

```
Diagnostic tests:
                 df1  df2 statistic p-value
Weak instruments    1 9997      9852  <2e-16 ***
Wu-Hausman          1 9996      3487  <2e-16 ***
Sargan              0   NA        NA      NA
```

1. Weak instruments: F-test on instruments in the first stage
   - H_0: We have weak instruments, so a rejection means our instrument(s) are not weak, which is good
2. Wu-Hausman: Consistency of OLS estimates under assumption that IV is consistent
   - H_0: OLS is consistent, suggesting endogeneity is not present
   - If we accept null, OLS and IV estimates are similar, and endogeneity may not be big problem

Which instruments are good instruments?

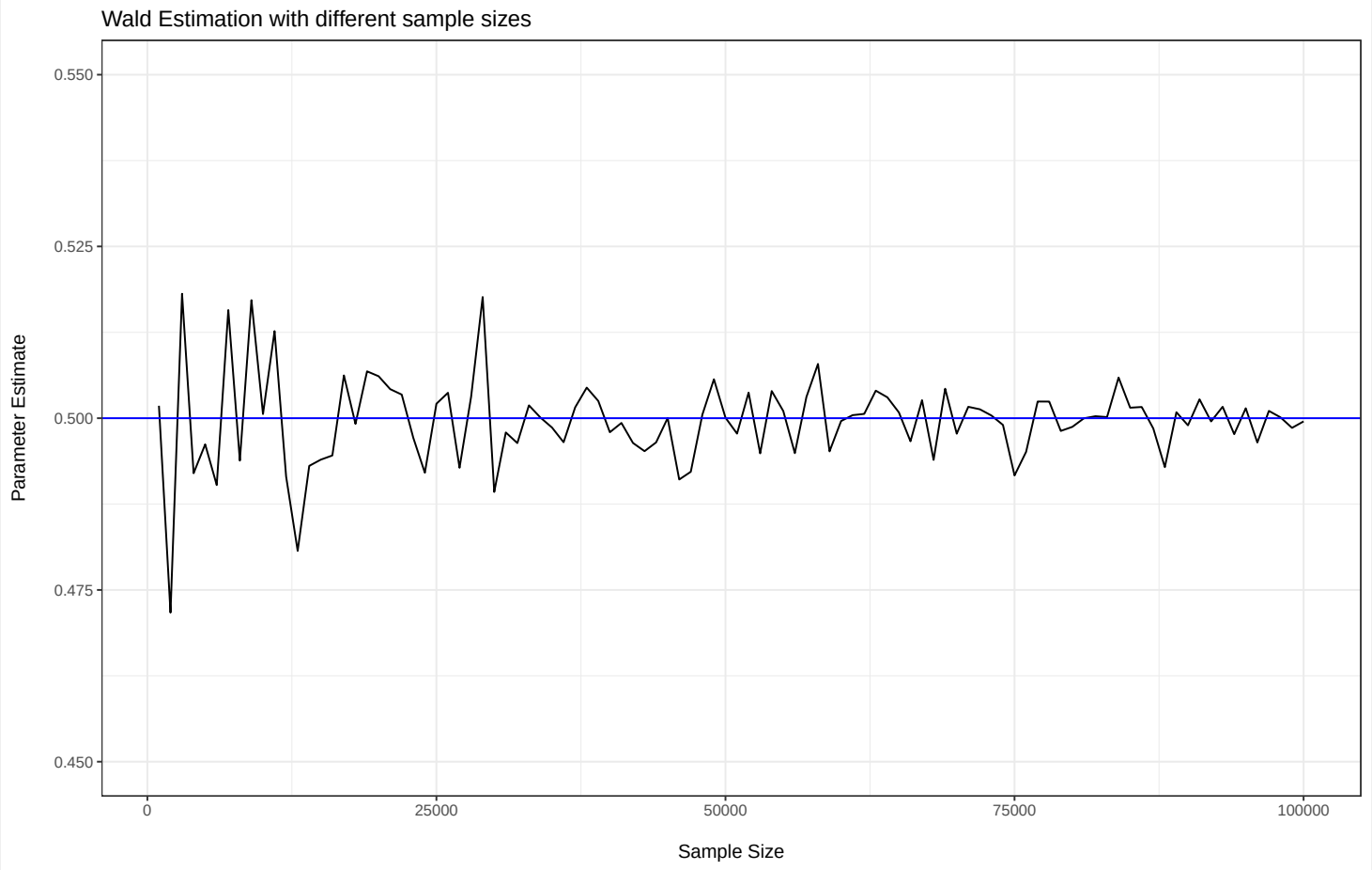
```r
# check assumptions
summary(ivReg, diagnostics =
    TRUE)
```

```
Diagnostic tests:
                 df1  df2 statistic p-value
Weak instruments   1 9997      9852  <2e-16 ***
Wu-Hausman         1 9996      3487  <2e-16 ***
Sargan             0   NA        NA      NA
```

3. Sargan: Exogeneity using over-identifying restrictions, called *J*-statistic
    - Can only be used if you have more instruments than endogenous regressors, which we don't
    - If null is rejected, ≥ 1 of our instruments is invalid (possibly all of them)

Reminder about consistency of IVs

```
# create sequence to loop over
obs <- seq(1000, 100000, by = 1000)
iv_consistency <- purrr::modify(obs, simulate)
```



Summary: Adjustment for observational data

We can interpret a coefficient as causal estimate only if we are persuaded that treatment assignment is ignorable:

- Conditioning on "observables" to approximate ignorability:

$$Y_i = \alpha + \beta T_i + \gamma X_i \beta + \epsilon_i, T|X \perp \epsilon|X$$

- Non-ignorable assignment (cannot condition on "unobservables"): Isolate exogenous variation in T

Wrap Up

In this lesson, we went over how to...

- Estimate & interpret a duration model for "count" data!
  - Survival distributions and hazard functions
  - Kaplan-Meier curves and Cox regression
- Selection models
  - Censored data, Tobit
  - Heckman
  - 2SLS

Next time, y'all will present your replications

- Upload your replication pdf and R file to GitHub before 23:59 Sun

Time for Evaluations!