Censored and truncated data: living with it and loving it

Gordon A. Fox
University of South Florida

2015

Introduction to censored data

Ad hoc substitutions

Better approaches: a taste

Getting some insight on censored data

Regression and censored data
- Regression with ad hoc substitutions
- Censored regression

Truncation
- Truncated regression
- Truncated event times

Outline

Ecological data

- Our data can be complicated!
- Often we have data that are missing, **censored**, or **truncated**.
- These can severely affect our estimates
- This can seem confusing. We can handle it, though!
- But . . . do we *always* have to use special methods?

Some examples of data censorship

- Is the water safe? Measuring concentrations of fecal coliform bacteria in water; we can only measure concentrations above some level.
- Estimating the range of a species from remotely-sensed data; we can only find them above some density.
- Estimating the survival times of a fish. Some individuals live longer than our study.

Where does the term come from?

- I don't know!
- But think of it as being like a government censoring parts of a newspaper article: you know it's there, you know how long the blank space is, ...

There are poor ways of handling these problems

We should avoid doing this statistically.

Outline

Some poor ways of handling the data ...

- Ignore the problem by deleting data points that have these problems.
- *Ad hoc* substitutions
  - Set values beyond the limits equal to the limits.
  - If it is a lower limit for detectability (LOD), set those values to 0.
  - If it is a lower LOD, set those values to LOD/2.

Why are these bad ideas?

- Deleting the data values . . . no, let's not even talk about it!
- *Ad hoc* substitutions: these introduce structure to the data set that is not in the actual population. They also make estimates **inconsistent** (more on this later).

Outline

R preliminaries

```
options (replace.assign = TRUE)
options (contrasts = c ("contr.treatment", "contr.poly"))

library (car)
library (NADA)
library (survival)
library (MASS)
library (truncreg)
library (eha)
library (truncnorm)
```

A simulation of left-censored data

Lognormal distribution. Assume

- One **LOD** (limit of detection) $= 0.5$.
- *Ad hoc* methods for dealing with left censorship: set values to the LOD, to LOD/2, or to 0.

```
source("SimulateLeftCensoredData.R")
```

Look at the data frame

```
head(example1)
```

```
##         y y_cens_lod y_cens_halflod y_cens_zero cens    TF
## 1 0.98309    0.98309        0.98309     0.98309    1 FALSE
## 2 0.45673    0.50000        0.25000     0.00000    2  TRUE
## 3 3.76831    3.76831        3.76831     3.76831    1 FALSE
## 4 0.79317    0.79317        0.79317     0.79317    1 FALSE
## 5 0.18943    0.50000        0.25000     0.00000    2  TRUE
## 6 7.36636    7.36636        7.36636     7.36636    1 FALSE
```

Data values less than 0.5 are:

- `y_cens_lod -> 0.5`
- `y_cens_halflod -> 0.25`
- `y_cens_zero -> 0`

Descriptive statistics

The descriptive statistics, using the differing approaches, are as follows.

```
means <- sapply (example1 [, 1 : 4], mean, na.rm = TRUE)
medians <- sapply (example1 [, 1 : 4], median, na.rm = TRUE)
sds <- sapply (example1 [, 1 : 4], sd, na.rm = TRUE)
sumStats <- data.frame (cbind (median = medians, mean = means,
  sd = sds))
sumStats
```

```
##                median   mean     sd
## y               1.032 1.6695 2.0735
## y_cens_lod      1.032 1.7115 2.0466
## y_cens_halflod  1.032 1.6580 2.0807
## y_cens_zero     1.032 1.6045 2.1191
```

The estimates are not the same. These differences might matter.

Better approaches: Kaplan-Meier (KM)

- Non-parametric.
- Comes from survival analysis: what is the cumulative distribution of data?
- Modified for this use by astronomers!

Better approaches: robust regression on order statistics (ROS)

- Normal distribution: if the Q-Q plot (probability plot) is $\approx$ linear, fit a regression to it. Intercept estimates the mean; slope estimates the sd of the distribution.
- More general: for any location-scale distribution, if the probability plot is $\approx$ linear, the intercept and slope estimate the location and scale parameters, respectively.
- This seems **immoral, unjust, and improper** to many people! But:
- Based on statistical theory.
- The method provides a direct check of the assumed distribution.
- **Robust ROS**: instead of using the intercept and slope, use the regression model to impute[1] the values of the censored points, and then use that pseudo-data set to to estimate the mean and sd.

[1] see Nakagawa, ch. 4

Better approaches: maximum likelihood (ML)

Censored (tobit) regression. Will discuss below!

Using the NADA library

```
nada.est <- with (example1, censtats (y_cens_lod, TF))

##       n   n.cen pct.cen
##   500.0   107.0    21.4

nada.est

##     median   mean     sd
## K-M 1.0208 1.7118 2.0470
## ROS 1.0320 1.6688 2.0740
## MLE 1.0497 1.6617 2.0391
```

All three methods did pretty good jobs. Not always the case; ML will work poorly for small samples and KM will work poorly if more than half the data are censored.

These methods also work for multiply censored data

Data sets with more than one LOD.

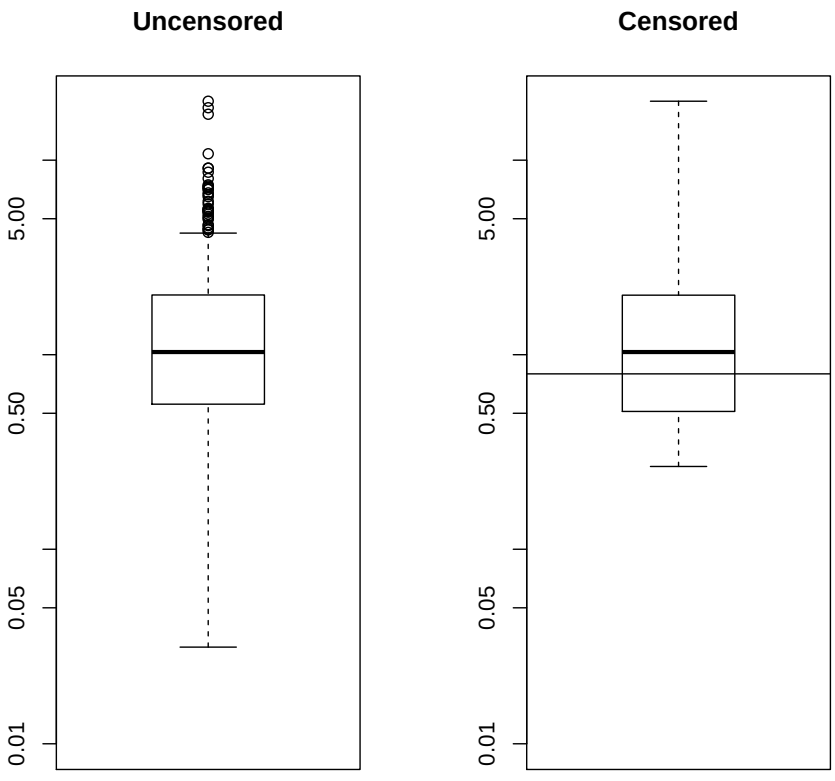
```
source("MultiplyCensoredData.R")
head(example2)
```

```
##         y y_cens_lod y_cens_halflod y_cens_zero cens    TF
## 1 0.98309    0.98309        0.98309     0.98309    1 FALSE
## 2 0.45673    0.56792        0.28396     0.00000    2  TRUE
## 3 3.76831    3.76831        3.76831     3.76831    1 FALSE
## 4 0.79317    0.79317        0.79317     0.79317    1 FALSE
## 5 0.18943    0.48611        0.24305     0.00000    2  TRUE
## 6 7.36636    7.36636        7.36636     7.36636    1 FALSE
```

Censored box plots

Since ROS can estimate a distribution's quantiles, it can also estimate boxplots of distributions including censored values.

```
par (mfrow = c (1,2))
boxplot (example2$y, main="Uncensored", log="y", ylim = c (0.01, 20))
cenboxplot (example2$y_cens_lod, example2$TF, main="Censored", ylim = c (0.01, 20))
```



Outline

Much terminology

- Taxonomy!
- Censored, truncated
  - Left censored, interval censored, right censored
- Much more . . . most of which we will ignore here!
- Don't worry. **Take a deep breath.** There are only a few important concepts to remember; the others you can always look up!

Censored data

Data values for which we know only an inequality, e.g.:

- **Left-censored**: typically from limits to measurement. Technique may say 0, but it has a lower resolution of c (say, 17.2), so $0 \leq x_i < 17.2$. Censored data are on the left end of the distribution.
- **Right-censored**: typically from event-time data. Some events occur after end of study c (say, day 72), so $x_i > 72$. Censored data are on right end of the distribution.
- **Interval censored**: mostly occurs in time-to-event data. We know the event occurred between two times c_1 and c_2, e.g., $38 < x_i > 93$. Censored data are in an interval in the middle.

Characteristics of data values

- A data set can include some data that are right-censored and some that are interval-censored, for example. The data *points* are censored, not the data set.
- Censorship can come from
  - Research design
    - technique for measurement
    - stop measuring after time y
    - stop measuring after n events
  - Random factors (individuals lost to follow-up).
- There can be one or multiple values at which some data are censored.

Let's consider some examples

What can you say about each of these examples? What's going on?

- Measuring concentrations of a pollutant in water. Our machine only measures concentrations > 5 ppm.
- Estimating seed dispersal distances. We cannot follow individuals farther than 500m.
- Estimating survival. Some of our study individuals die because we step on them.
- Estimating flowering time. We mark individuals and check them every week, but have a month in which we cannot get to the field site.

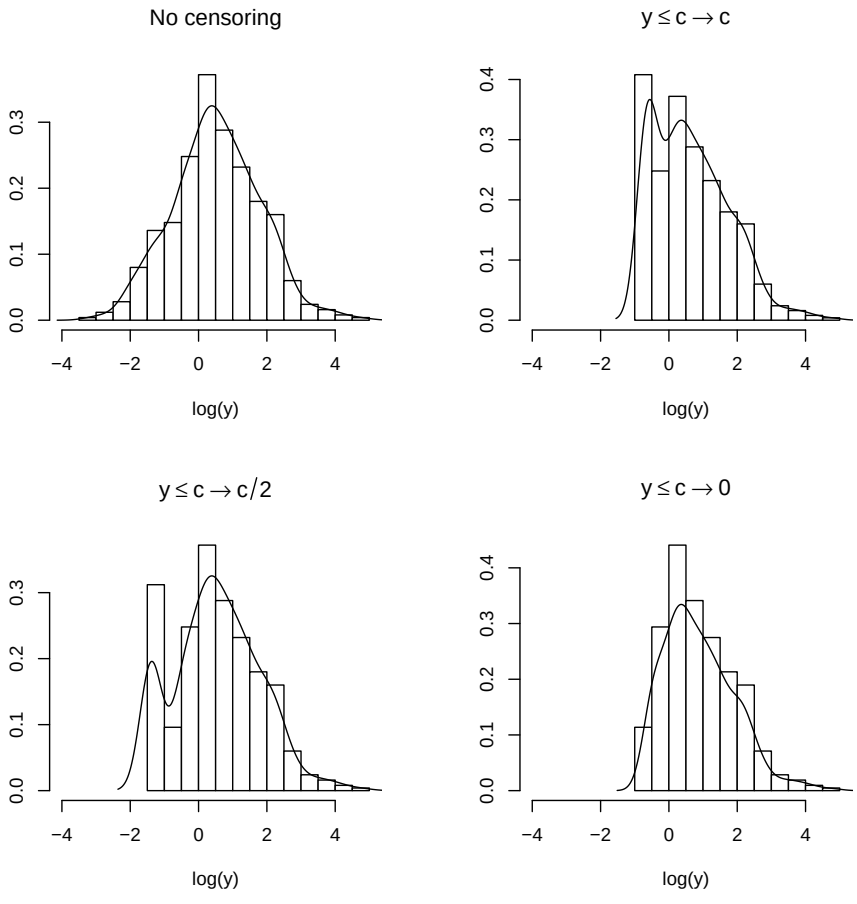
Outline

Simulate some lognormal data for regression

Lognormal, and with the *ad hoc* substitutions in the data frame as before.

```
source("SimulateLogNormal.R")
```

The data distributions



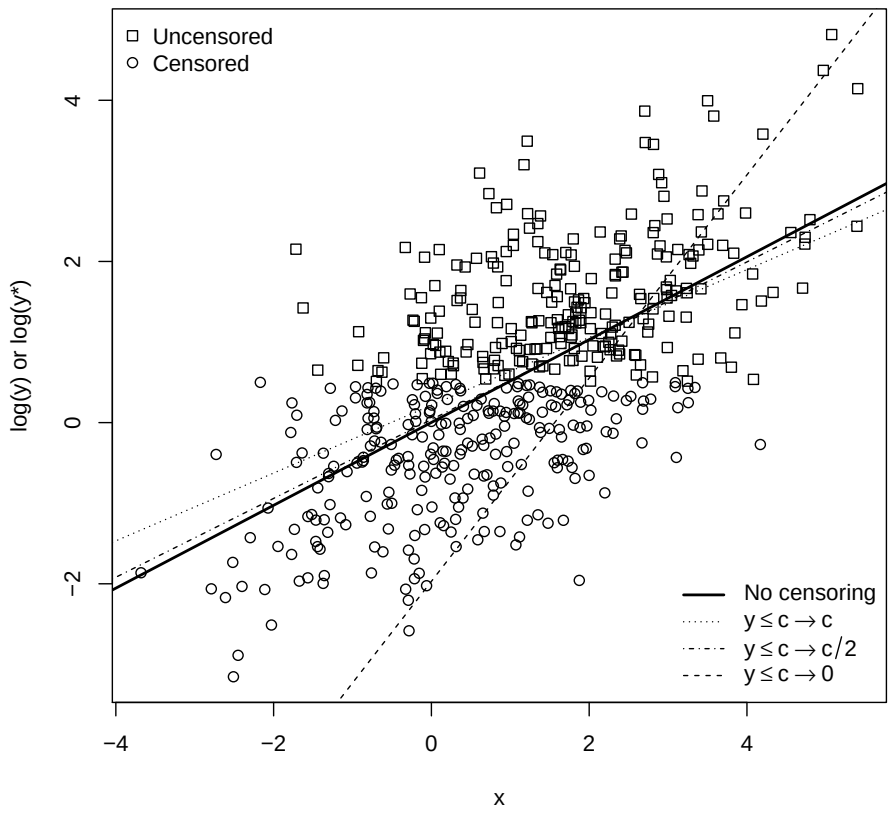
How well does OLS do with these?

Estimate the relationship between x and y using `lm` for these different ys.
Estimated parameter values:

```
estimates_y_error

##              Uncensored y_cens_lod   y.half y_cens_zero
## Intercept     0.0019132    0.21558 0.037733    -1.96973
## Slope         0.5140465    0.42075 0.489130     1.26100
## Adjusted R2   0.3964416    0.36038 0.377203     0.27298
```

The differing estimates



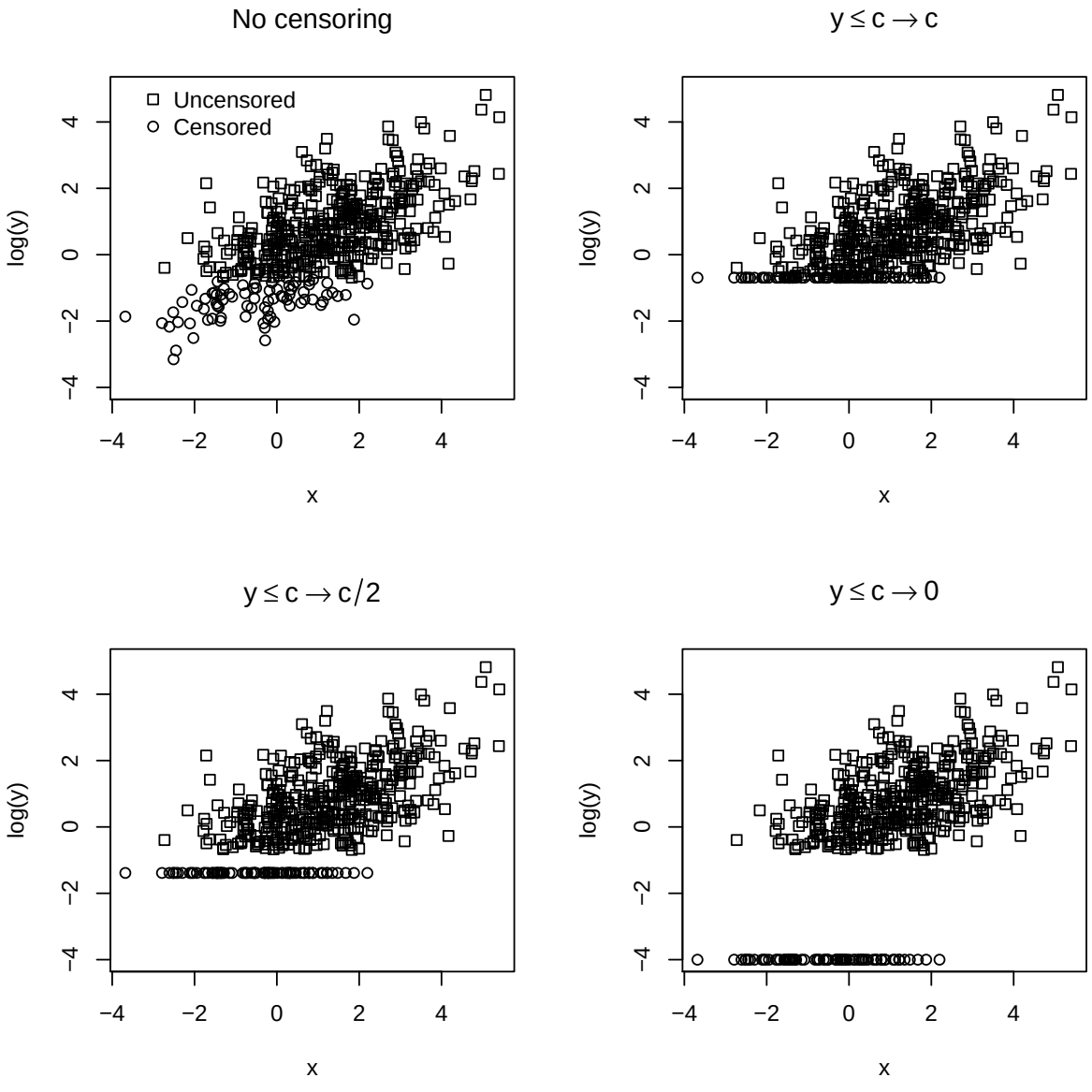
The intended relationship was $\log y = x/2$. The uncensored estimation ("true" values) recovers this relationship. Some of the substitutions are better than others.

Residuals – new patterns of variation!

- Substitutions imposed new patterns of variation on the data.
- Technical problem: cannot plot the case of y set to 0 on a log scale.
- Crude fix: set logs of those values to a small value. For these data, -4 works well.

```
yln0 <- log (logn_example$y_cens_zero)
yln0 [which (is.infinite (yln0))] <- -4
```

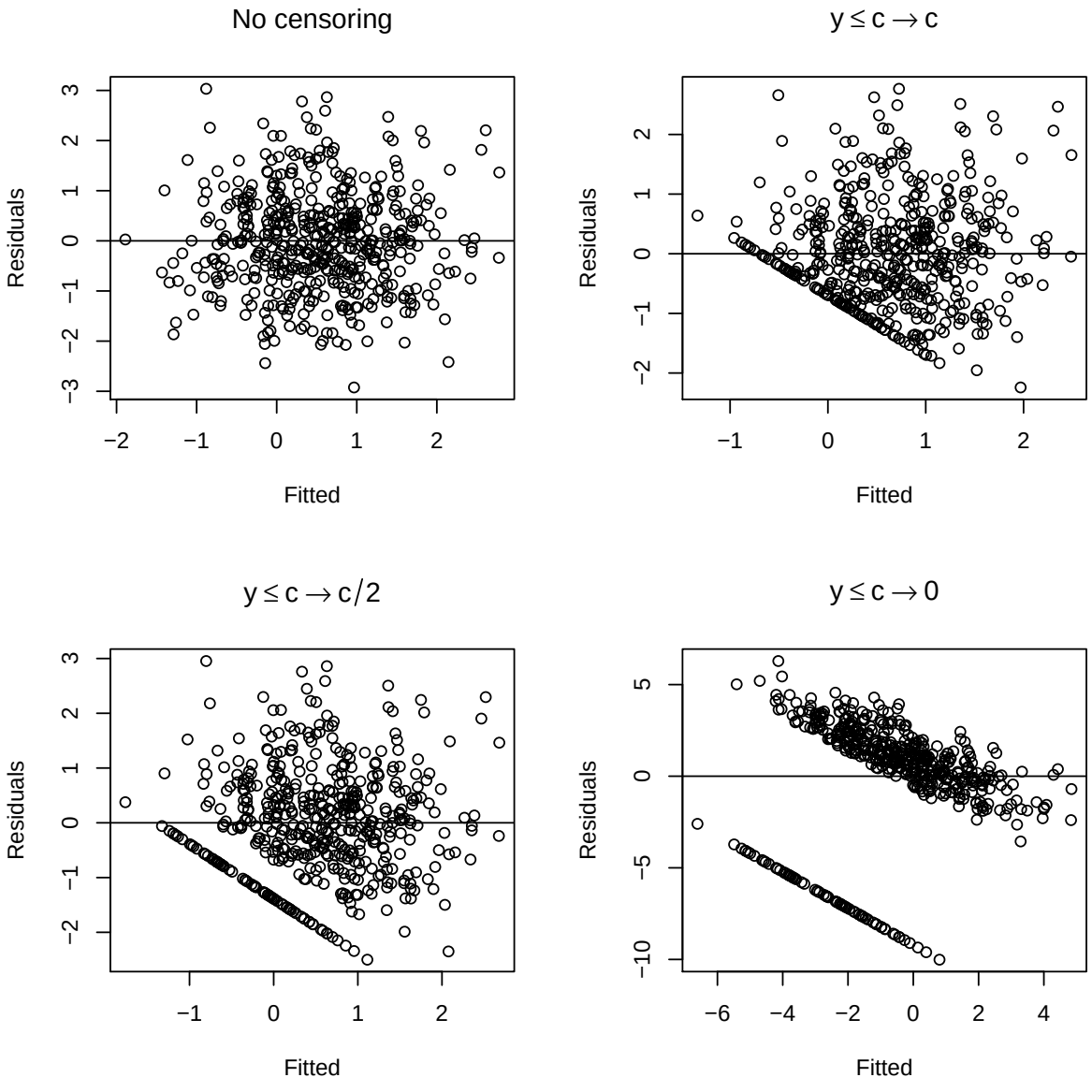
Scatter plots



Even worse

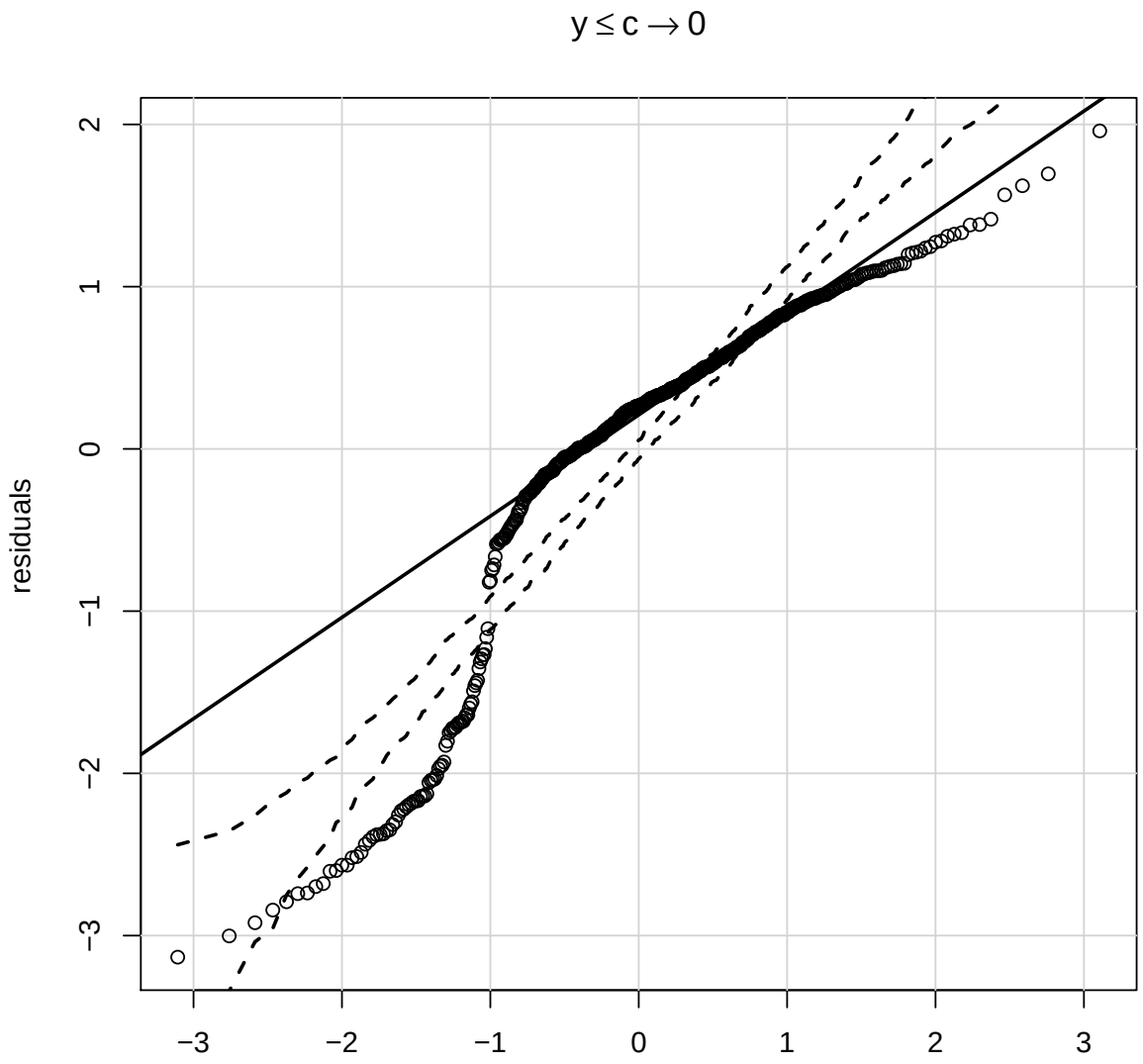
Not shown here, but in Appendix 5a: when we set observations of 0 to a small value, that value has very strong effects on the OLS estimate, with smaller values having stronger effects.

Residual plots



We can also use probability plots

```
qqPlot (lm_y_cens_zero, line = "quartiles", ylab = "Studentized
  residuals", main = cens_labels [4], col.lines = "black")
```



Bias and consistency

- **Bias**. An unbiased estimator, on average, estimates the true value. A property of small samples.
- **Consistency**. As the sample size increases towards infinity, the estimates approach the true value of the population quantity. A property of large samples.
- Use of *ad hoc* substitutions leads to **inconsistent** estimates.

Ad hoc substitutions . . .

. . . are a *really*, *really* bad idea. **Please** don't use them!

First, some more data

Same as earlier, but a larger data set to make patterns clear.

```
source("LargeDataSet.R")
```

Key idea behind tobit regression

Assume (for now) linear model, normal distribution, one censorship level c, univariate. Then observed data are:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i & \text{if } y_i^* > c \\ c & \text{if } y_i^* \leq c \end{cases}$$

where y_i^* is the true value. Then the expectation is

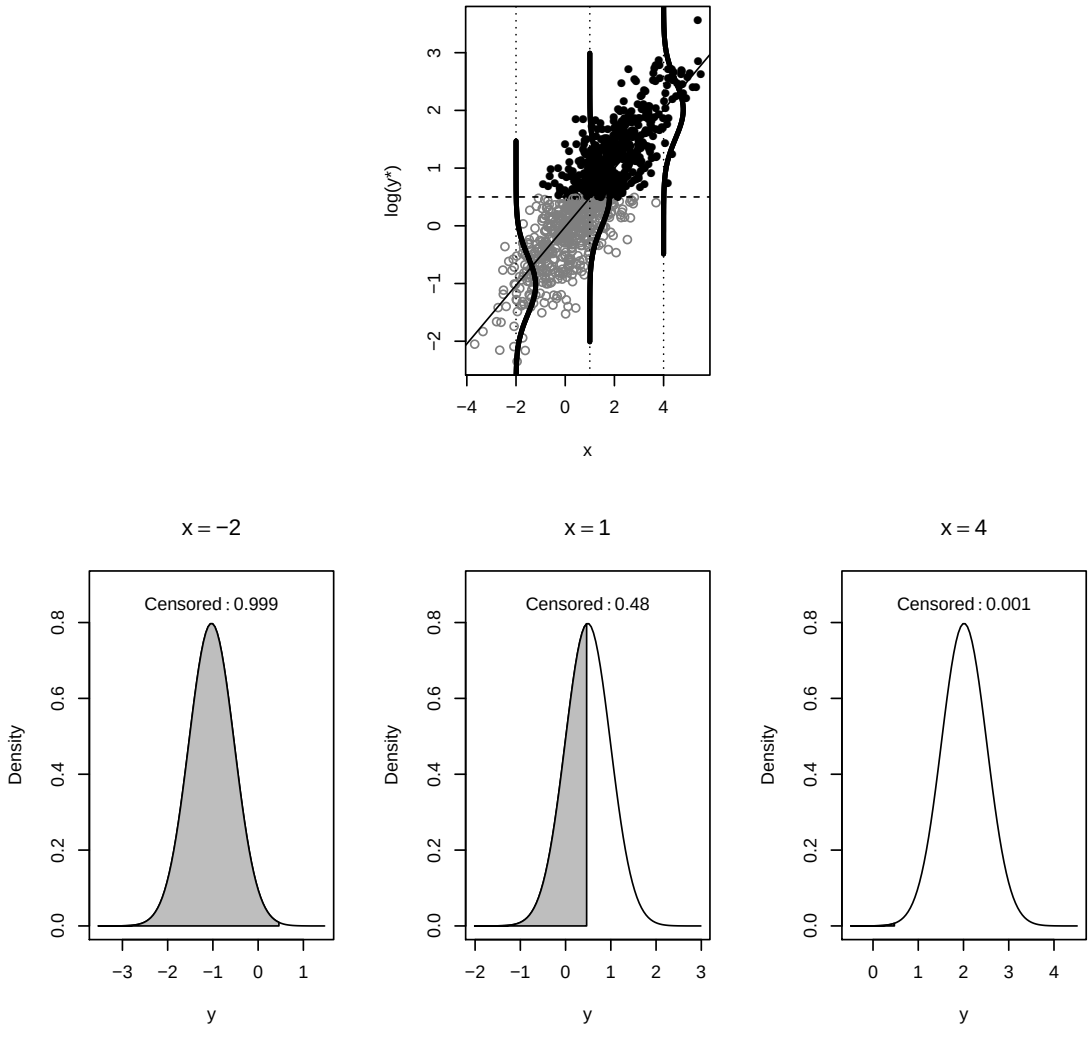
$$E[y_i|x_i] = P\left(\text{uncensored}|x_i\right) \times \mathbf{E}\left[y_i|y_i > c, x_i\right] + P\left(\text{censored}|x_i\right) \times c$$

Log likelihood function

- Uses CDF to estimate probability a point is censored.
- Uses PDF to estimate probability of y, given that it is not censored.

PDFs and CDFs

PDFs for $y|x$ (top). CDFs for $y|x$ for three values; shaded areas give probability of censorship (bottom).



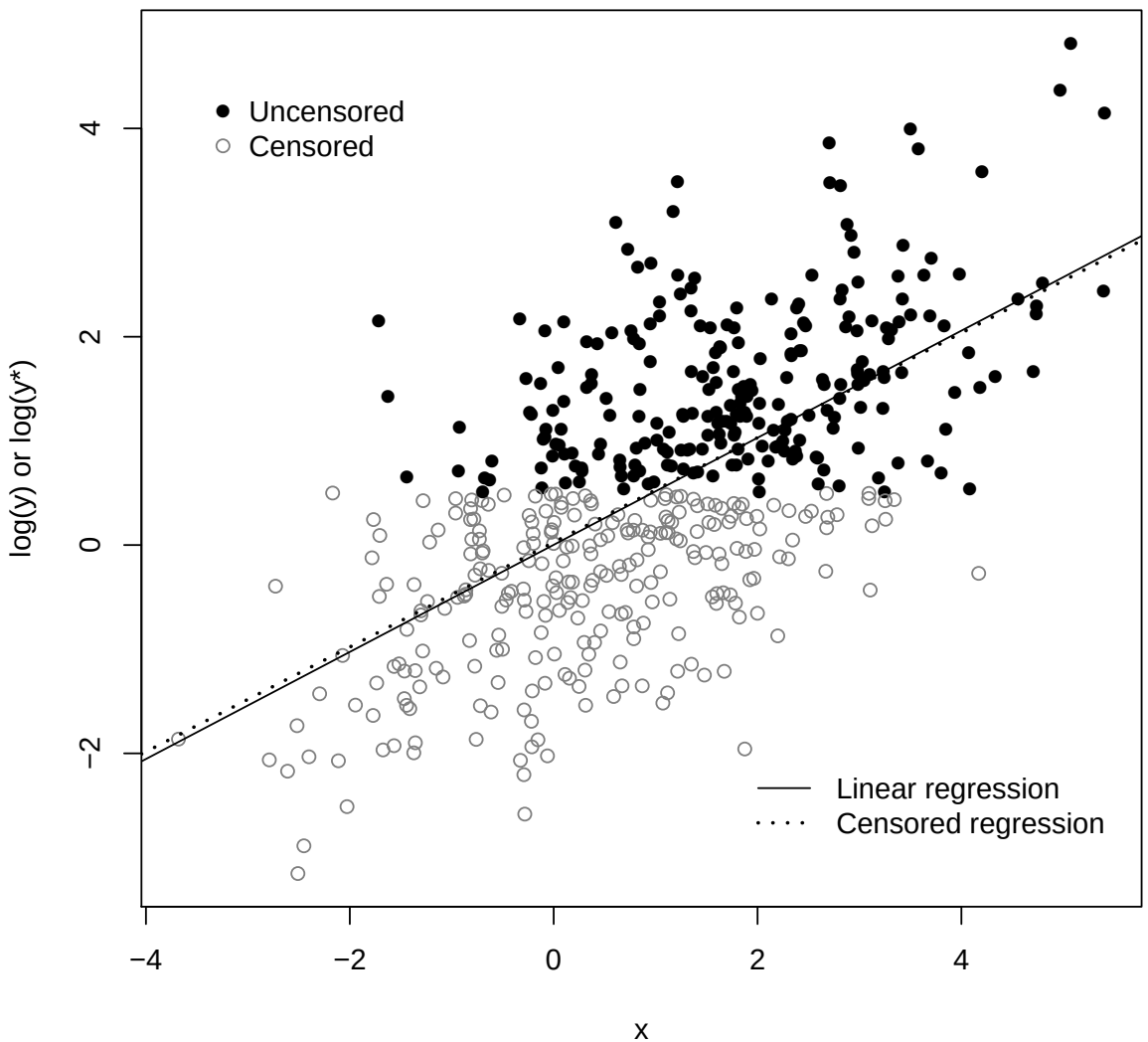
Fitting a censored regression model

```
sfit <- survreg (Surv (y_cens_lod, cens == 1, type = "left") ~ x,
  dist = "loggaussian", data = logn_example)
summary (sfit)
```

```
##
## Call:
## survreg(formula = Surv(y_cens_lod, cens == 1, type = "left") ~
##     x, data = logn_example, dist = "loggaussian")
##                Value Std. Error      z        p
## (Intercept)  0.0257     0.0569  0.453 6.51e-01
## x            0.5025     0.0300 16.731 7.83e-63
## Log(scale)  -0.0087     0.0353 -0.246 8.05e-01
##
## Scale= 0.991
##
## Log Normal distribution
## Loglik(model)= -1039   Loglik(intercept only)= -1157.6
##  Chisq= 237.16 on 1 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 4
## n= 500
```

An excellent fit

True intercept and slope are 0 and 0.5.



Outline

Truncated data

- Our design means that there are data values over some range that we aren't even sampling, and (for our analysis) don't even know about.
- Examples: (1) Estimating spread of an invasive plant. We have remotely-sensed data with a minimum resolution of 30m. (2)Estimating species diversity in phytoplankton. Our net has pores 20 μm in diameter.
- Left- and right-truncated data.
- This is quite different from censorship, although it sometimes sounds similar! Ignoring the problem leads to inconsistent estimates.

Simulated data

Eliminate the censored values to get a truncated data set.

```
lgn <- subset (logn_example, y > lod)
```

Take a random subset of the original (untruncated) data set, so we are comparing samples of the same size.

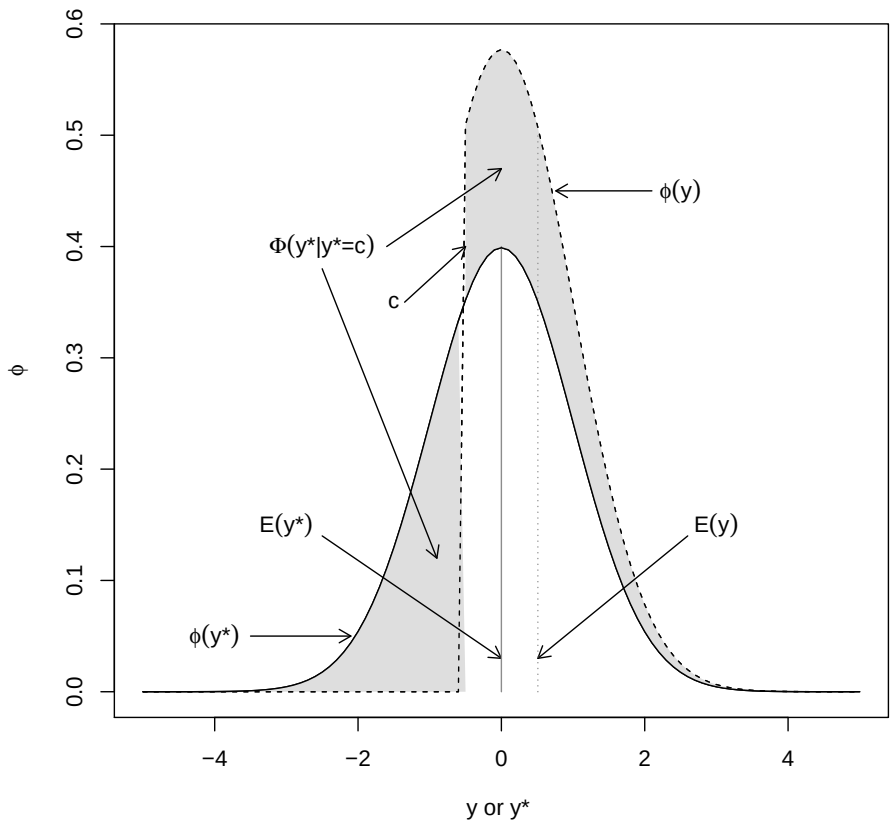
```
sample_size <- nrow (lgn)
subs_logn <- logn_example [sample (nrow (logn_example), sample_size), ]
```

Similarities and differences

- Many ideas are similar to those for censored data.
- We don't know about the truncated points, so we can't use methods like ROS or KM.
- Truncated regression uses a logic similar to censored regression. Key is thinking about how the PDF changes from truncation.

Change in the PDF from trunctation

Untruncated (solid); truncated (dashed). Shaded region to the left of c is truncated (and $= \Phi(c)$). A region of the same area augments the PDF for the truncated distribution. The mean of the truncated distribution is larger, and the variance is smaller.



The likelihood function in truncreg makes the appropriate adjustments.

Regression on truncated data

Truncated regression

```
treg <- truncreg (log (y) ~ x, lgn, point = log (lod), direction = "left")
```

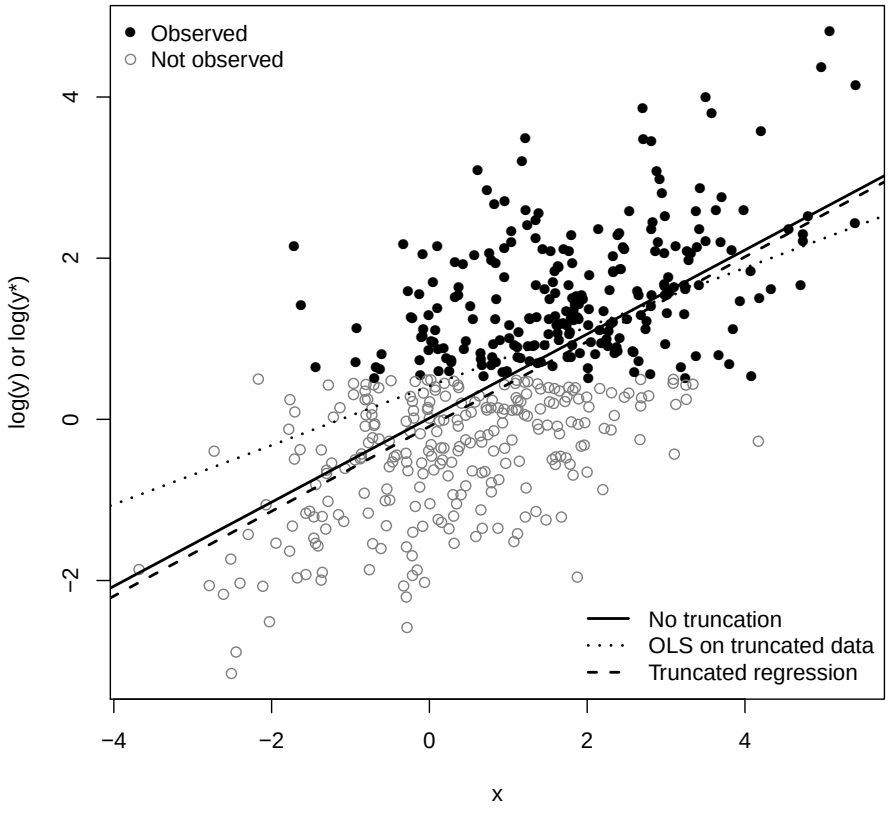
"Naive" model – OLS on truncated data

```
lm_naive <- lm (log (y) ~ x, data = lgn)
```

OLS for the untruncated data

```
lm_untruncated <- lm (log (y) ~ x, data = subs_logn)
```

How well do the models perform?



Diagnostics etc.

- Be careful!
- Interpret residuals with care – there will be an excess of residuals in the truncated region! This is true even for the truncated regression model.
- This is not linear.
- Normality is questionable. I recommend bootstrapping the CIs; some code for this is in Appendix 5A.

Common in survival studies

Truncation is common in time-to-event data! Example: study of lifetime reproductive success, in which individuals younger than some age are ignored.

The `survival` package doesn't process truncated data, but the `eha` (event history analysis) package does.

I am using a large n here because we are truncating it quite severely – there will be just under 1000 events in the data set analyzed.

Generating the data, 1

Large sample because we are truncating severely.

```
n <- 5000
set.seed (59234)
```

Two covariates

```
x1 <- rnorm (n, 8, 2)
x2 <- rnorm (n, 12, 0.5)
```

Set the Weibull parameters

```
baseline <- 0.03
beta1 <- 2
beta2 <- -1
```

Model the failure times using the parameters set above.

```
T = rweibull (n, shape = 1.2, scale = baseline * exp (- (beta1 * x1 + beta2 * x2)))
wei_dat <- cbind (x1, x2, T)
```

Now truncate

Set a LOD

```
wlod <- 0.01
trunc_wei <- subset (wei_dat, wei_dat [, 3] > wlod)
```

Create a data set that's the same length, but randomly selected from the full set

```
wei_sample <- length (trunc_wei [, 1])
subs_wei <- wei_dat [sample (length (wei_dat[, 1]), wei_sample), ]
```

Make these into data frames

```
trunc_wei <- as.data.frame (trunc_wei)
subs_wei <- as.data.frame (subs_wei)
```

Fit survival models

Fit accelerated failure time models to the untruncated and truncated data. Also fit a "naive" model to the truncated data.

```
source("WeibullModels.R")
```

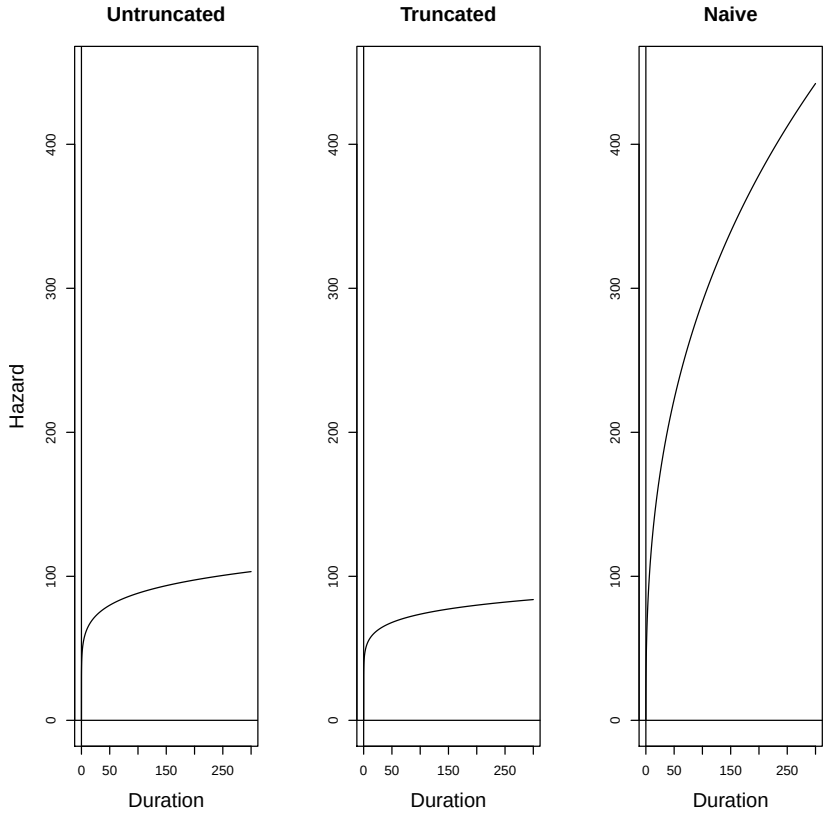
Here's a view of these parameters:

```
true <- c (beta1, beta2, baseline, 1.2)
wei_params <- cbind (true, c (coef (wei_untrunc) [1 : 2], exp (coef (wei_untrunc) [3 : 4])),
  c (coef (wei_naive) [1 : 2], exp (coef (wei_naive) [3 : 4])), c (coef (wei_trunc) [1 : 2],
  exp (coef (wei_trunc) [3 : 4])))
rownames (wei_params) <- c ("x1", "x2", "scale", "shape")
colnames (wei_params) <- c ("True", "Untruncated", "Naive", "Truncated")
wei_params
```

```
##         True Untruncated     Naive Truncated
## x1      2.00    2.011676  1.854940  2.049618
## x2     -1.00   -0.985367 -0.878853 -0.996535
## scale  0.03    0.039691  0.074626  0.038186
## shape  1.20    1.143004  1.382305  1.117383
```

Very different model fits!

Ignoring truncation yields poor parameter estimates. The two covariates are both underestimated in importance, and the Weibull scale parameter is estimated at more than twice its true value. This translates into a big difference in the hazard functions (the rate at which events occur).



Censorship and truncation

- Common in ecological data
- Can severely affect estimates
- A fair amount is known about how to handle these data
- Do we **always** have to use these methods?