Stat 547 - Statistics in Ecology: Tutorial 2 - Censored data

Marie Auger-Méthé

1 Goals and set up

The goal of this tutorial is to explore the problems associated with censored data and to look at some of the statistical methods we can use to handle censorship.

1.1 Instructions

Please read the text and reproduce the analyses presented in the code boxes of the tutorial. In addition, in a separate script or as an R markdown document, please answer the questions and do the analyses described in the *Exercise* sections. If you are simply using an R script, answer the questions by using comments in the code and, in addition, please comment your code so I understand each step of your analyses. I'm expecting a comment every 1 to 3 lines. If you are using R markdown, make sure to explain each of the steps of your analyses either in the text or in the code box. Please send me the final document at auger-methe@stat.ubc.ca at the end of the class.

1.2 Setup

First, load the package we will need to complete the analyses. Off course you need to have it installed first.

```
# Load packages
library(survival) # More or less all analyses use this package
```

Then, let's read the dataset we will use. For this tutorial, we will use the dataset on the survival of cougars provided by Moss et al. (2015a) on dryad.

```r
cougars <- read.csv("JAE_00481_Cougar Survival Data.csv") # Read the data
head(cougars) # Take a quick look at the data
```

```
##   Individual.ID EntryMonth ExitMonth TotalMonth Event Sex Mortality.Cause
## 1           F03         60        71         11     0   F
## 2           F04         45        71         26     0   F
## 3           F05         25        63         38     1   F           Human
## 4           F06         41        71         30     0   F
## 5           F07          8        40         32     1   F           Human
## 6           F10         47        71         24     0   F
##   Mortality.Detail  X_13C X_15N X..Diet.Herbivore Housing.Density
## 1                  -21.34  6.80               65%            6.19
## 2                  -21.48  7.27               62%            0.00
## 3       Euthanasia -20.73  8.33                              10.08
## 4                  -21.34  7.81               55%            7.59
## 5         Roadkill -21.17  8.73               45%           23.14
## 6                  -21.46  8.61               48%           40.91
```

The authors estimated survival based on GPS collar information. In particular, the EntryMonth values are either the date the cougar was first captured to place a GPS collar or the date a known individual reached adulthood. The ExitMonth values are either the date when the data was censored (either through lost of communication with the GPS collar or end of the study) or the death of the individual (the cause of the death is explained in the columns Mortality.Cause and Mortality.Detail). TotalMonth is the amount of time the cougar was monitored. The column Event describe whether the data is censored (0) or whether it is associated with a known death (1). The other columns describe the covariates, most of which (X_13C, X_15N, X..Diet.Herbivore) are associated with the diet of the animal, and one associated with sex (Sex) and one with human density (Housing.Density). You can find more information on dryad (Moss et al. 2015a) and in their associated paper (Moss et al. 2015b). For the tutorial, we are generally interested in two things: 1) what's the mean number of months cougars survive, and 2) what covariates affect their survival.

Exercise 1

What type of censorship this dataset fall under?

2 Applying common (not recommended) methods to handle censored data

2.1 Use fix values for censored points

The first common method to handle censored data is to set the value of a censored data point to a fix value. Because it is survival data, we will use the value of the censored point as it is, which would mean simply ignoring the fact that it is censored and using the values from the column TotalMonth as if it was a correct representation of the number of months the cougar was alive. This would be equivalent to set the observation for all cases that are censored to the censored level ($y_i = c_i$). We use c_i because the censored levels change for each individual. While this is different in practice to, let's say, a study where the tool we used cannot measure something smaller than c, where you would have only one c and choose to set observation value (y_i) to a value such as 0 or $c/2$, it's the same theoretically (i.e. you are fixing the value of the censored data.

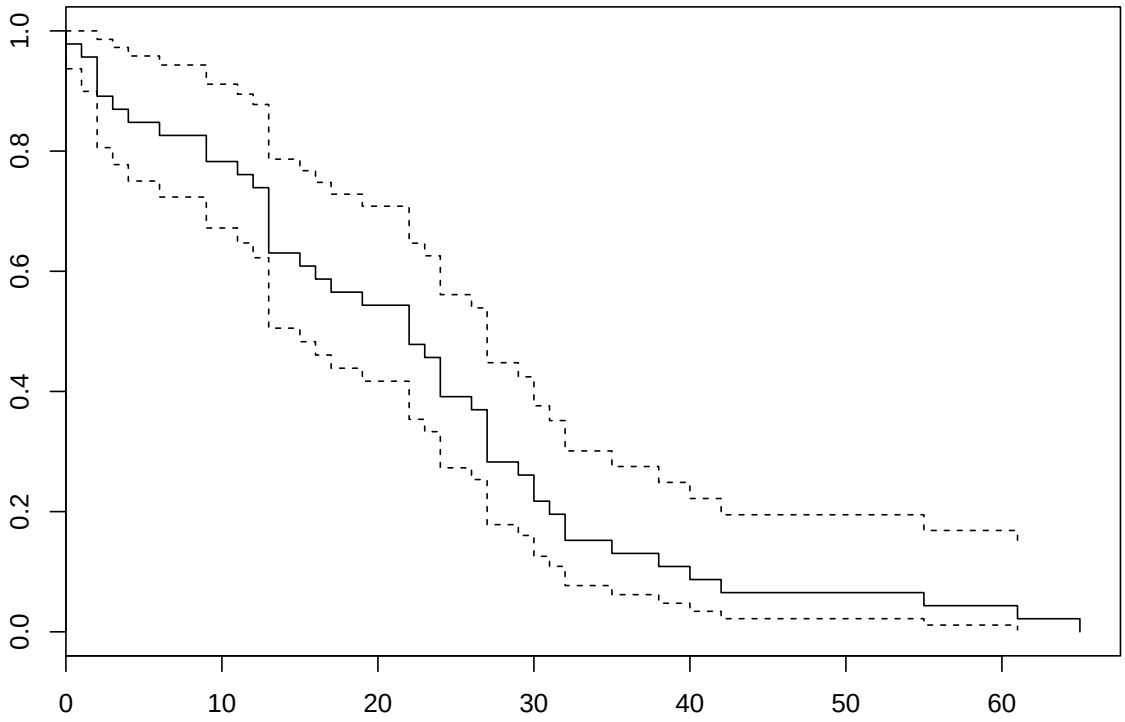
So first let's estimate the mean length of survival using the sample mean.

```
# Sample mean survival in months
mean(cougars$TotalMonth)

## [1] 21.91304

# Standard deviation of the survival
sd(cougars$TotalMonth)

## [1] 15.1611
```

We can also estimate the survival function. For this we will use the `survfit` function from the package `survival` and just indicate that the end of all points are considered as death (i.e. ignore the fact that it's censored).

```
# First create a survival object.
# For the argument event we create a vector that has 1 (death) for each row.
survV <- Surv(time=cougars$TotalMonth, event=rep(1, nrow(cougars)))
# Now estimate a survival function.
# The formula indicates that the survival is not dependent on covariates
```

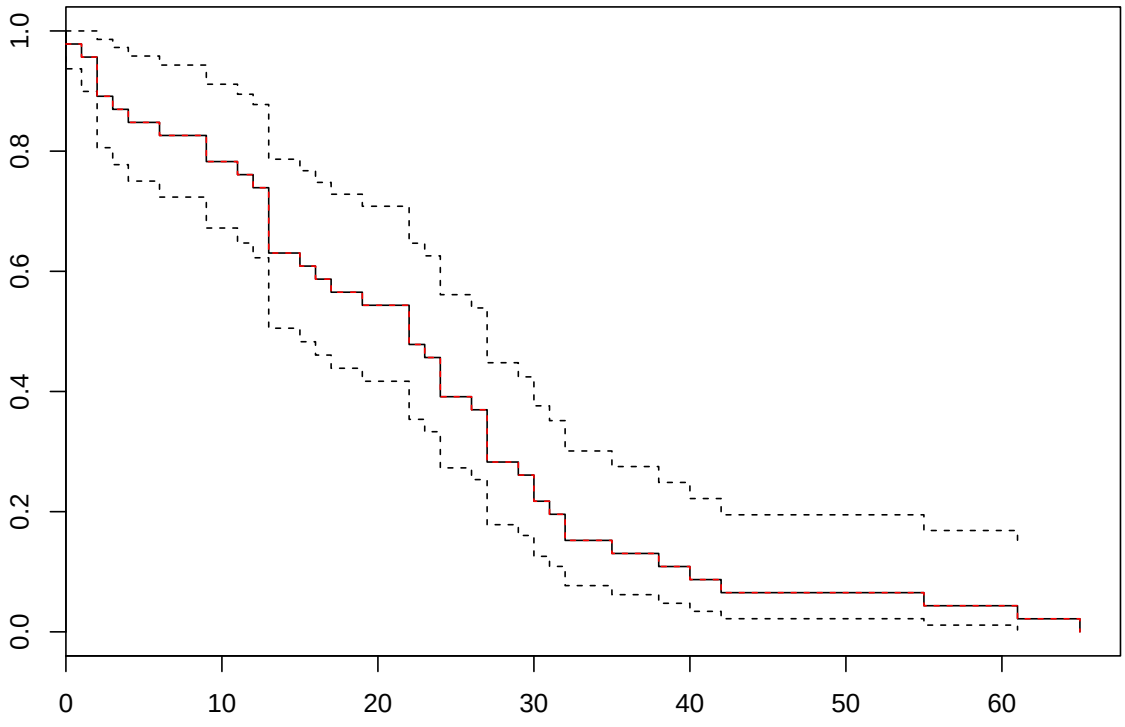
```r
fixV <- survfit(survV~1)
# Plot the survival curve
plot(fixV)
```



Here the function **survfit** uses the Kaplan-Meier estimator to estimate the survival curve, but because there is no censored data, it is equivalent to the empirical survival function, which we can estimate by hand as follow.

```r
# Create an empty matrix for the empirical survival function,
# with a column for the each unique value of y and s(y).
esf <- data.frame(matrix(ncol=2, nrow=length(unique(cougars$TotalMonth))))
colnames(esf) <- c("y", "s")
```

```r
# Fill in the values of y (ordered)
esf$y <- sort(unique(cougars$TotalMonth))

# Create a loop that calculates the proportion of
# the y values that are > than each y_i
for(i in 1:nrow(esf)){
  esf$s[i] <- sum(cougars$TotalMonth>esf$y[i])/nrow(cougars)
}

# Plot the empirical survival function on top
# of the Kaplan-Meier estimated function
points(esf, type="s", col="red", lty=2)
```



You can see that they match perfectly.

You can use the area under the curve (AUC) to estimate the mean survival using the Kaplan-Meier or the empirical survival function.

```
print(fixV, print.rmean=TRUE)
```

```
## Call: survfit(formula = survV ~ 1)
##
##          n     events     *rmean *se(rmean)     median    0.95LCL
##      46.00      46.00      21.91       2.21      22.00      15.00
##    0.95UCL
##      27.00
##     * restricted mean with upper limit =  65
```

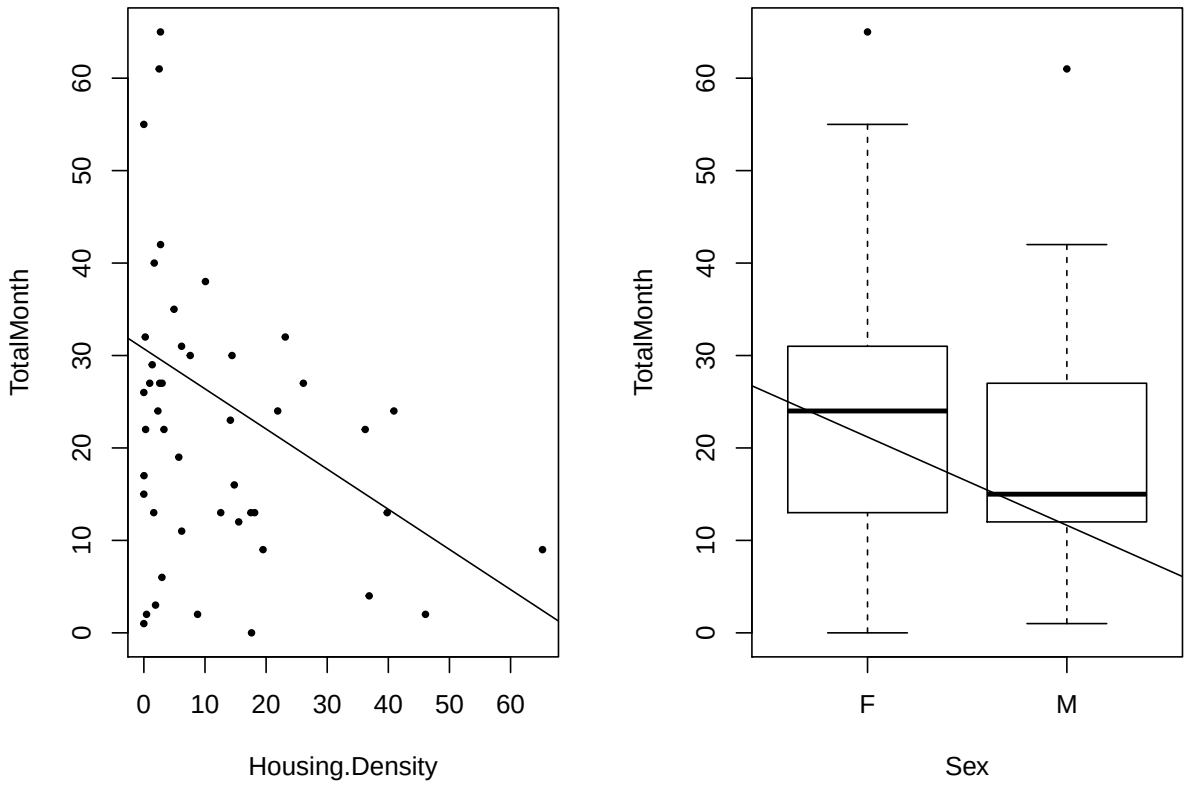
First, note that this returns the same values as the sample mean found above. See below for a discussion of the upper limit, but note that since we are ignoring the censorship here, changing the upper limit will not affect the results.

Next, we might be interested in understanding the relationship between covariates and survival. So for example, we will look at whether human density (column Housing.Density) and sex affect survival. The most naive way we could do this is using a linear regression.

```
fixlm <- lm(TotalMonth~Housing.Density+Sex, data=cougars)
summary(fixlm)
```

```
##
## Call:
## lm(formula = TotalMonth ~ Housing.Density + Sex, data = cougars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.450  -8.697  -0.005   6.892  40.907
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      30.7358     3.6400   8.444 1.12e-10 ***
## Housing.Density  -0.4342     0.1530  -2.838  0.00689 **
```

```
## SexM             -9.5528     4.6658  -2.047  0.04676 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.09 on 43 degrees of freedom
## Multiple R-squared:  0.175,Adjusted R-squared:  0.1366
## F-statistic: 4.559 on 2 and 43 DF,  p-value: 0.016


# Relationship with human density
plot(TotalMonth~Housing.Density, data=cougars, pch=19, cex=0.5)
abline(fixlm$coef[1:2])
# Relationship with sex
plot(TotalMonth~Sex, data=cougars, pch=19, cex=0.5)
abline(fixlm$coef[c(1,3)])
```



It looks like increasing human density decreases the survival of cougars and that females live longer (which is pretty common in animals).

2.2 Deleting cases with censored data

Another common method to handle censored data is to delete all cases that are censored. So in our example, we delete all of the rows where the column Event is 0.

```
# Let's look at the % of the data that would be thrown away
sum(cougars$Event == 0)/nrow(cougars)*100

## [1] 43.47826
```

```r
# Create a new data set with these rows removed
cougarsD <- cougars[cougars$Event == 1, ]
```

Exercise 2

a. Calculate the mean survival for this new dataset and compare the results to those found when we fixed the values for the censored data.

b. Estimate the survival function and compare it graphically to the one estimated when we fixed the values for the censored data.

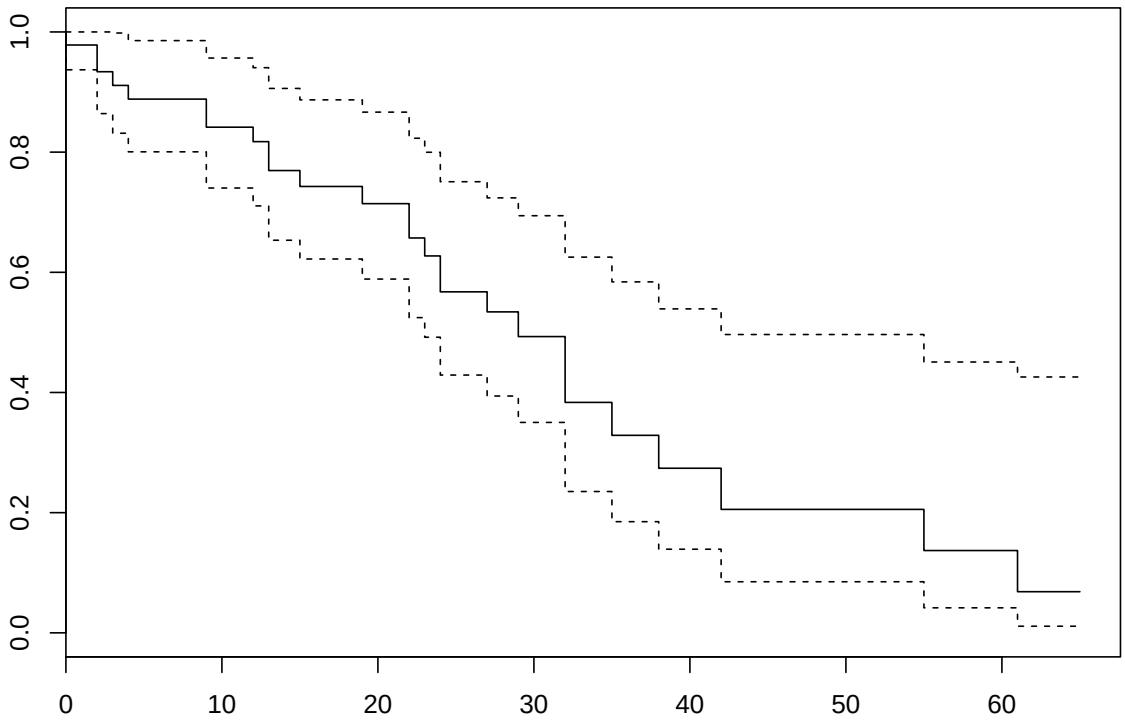
c. Fit a linear regression to the data to understand the relationship between the survival and the human density and sex. How do the results differ from the previous analysis where we fixed the censored values?

3 Estimating the mean properly

3.1 Kaplan-Meier estimate for the mean

Here we will use the `Surv` function again to estimate the survival curve with the Kaplan-Meier estimator. However, this time we will use the information on the censorship.

```r
# First create a survival object.
survKM <- Surv(time=cougars$TotalMonth, event=cougars$Event)
# Now estimate a survival function.
kmV <- survfit(survKM~1)
# Plot the survival curve
plot(kmV)
```



Exercise 3

Demonstrate that when we use censored data the survival function estimated with the Kaplan-Meier estimator differs from the empirical survival function. Explain how it differs and why it makes sense.

Next, we will use the survival function estimated with the Kaplan-Meier estimator to estimate the mean.

```
print(kmV, print.rmean=TRUE)

## Call: survfit(formula = survKM ~ 1)
##
```

```
##          n     events      *rmean *se(rmean)     median    0.95LCL
##      46.00      26.00      30.77       3.41      29.00      23.00
##    0.95UCL
##      42.00
##     * restricted mean with upper limit =  65
```

Here our estimate of the mean survival time is much longer than the one estimated when we fixed or deleted the censored values. Note also that in this case the values we set from the upper survival limit will make a difference. For example, let's say that we know cougars can survive 13 years, so would have a adult life of 132 months, we could set the maximum value of the survival to 132.

```
print(kmV, print.rmean=TRUE, rmean=132)

## Call: survfit(formula = survKM ~ 1)
##
##          n     events      *rmean *se(rmean)     median    0.95LCL
##      46.00      26.00      35.36       6.67      29.00      23.00
##    0.95UCL
##      42.00
##     * restricted mean with upper limit =  132
```

This will increase the mean survival even more. This might be appropriate if you have additional information on the survival of the animals.

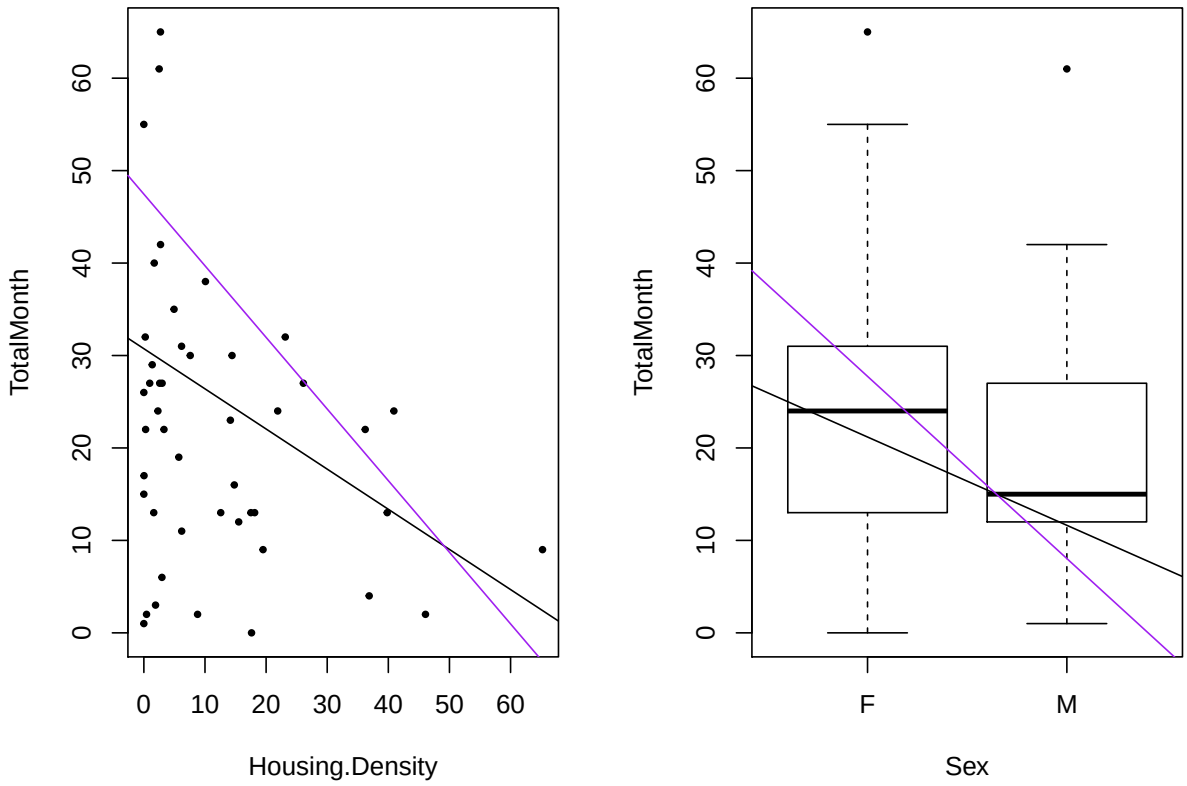
4 Understanding the relationship with covariates

4.1 Regression for censored data

To understand the relationship between the survival of cougars and human density and sex, we will use the function `survreg` to apply a regression for censored data (also know as tobit model).

```
# We apply survreg to the survival curve we estimated above using KM
couTobit <- survreg(survKM~Housing.Density+Sex,data=cougars, dist="gaussian")
summary(couTobit)
```

```
##
## Call:
## survreg(formula = survKM ~ Housing.Density + Sex, data = cougars,
##     dist = "gaussian")
##                   Value Std. Error     z        p
## (Intercept)      47.457      5.151  9.21 3.16e-20
## Housing.Density  -0.775      0.182 -4.25 2.09e-05
## SexM            -19.712      5.805 -3.40 6.85e-04
## Log(scale)        2.670      0.136 19.63 8.65e-86
##
## Scale= 14.4
##
## Gaussian distribution
## Loglik(model)= -113.8   Loglik(intercept only)= -123.1
##  Chisq= 18.51 on 2 degrees of freedom, p= 9.6e-05
## Number of Newton-Raphson Iterations: 4
## n= 46


# Relationship with human density (and compare to fixing the censored data)
plot(TotalMonth~Housing.Density, data=cougars, pch=19, cex=0.5)
abline(fixlm$coef[1:2]) # fix
abline(couTobit$coef[1:2], col="purple") # tobit
# Relationship with sex
plot(TotalMonth~Sex, data=cougars, pch=19, cex=0.5)
abline(fixlm$coef[c(1,3)])
abline(couTobit$coef[c(1,3)], col="purple") # tobit
```



Here using the censorship information changes the magnitude of the relationship with the covariates.

Exercise 4

In the previous example, we only took into account the fact that the death data was censored. However, the EntryMonth is also censored as it represents both the real start of adulthood and when the individual was first tagged. The package `survival` allows you to use interval censorship, which can be done by specifying the arguments time2 and type of the function `Surv`, see `?Surv` for detail.

a. Estimate the mean survival time when accounting for the censorship in the entry date. How do the results change from the previous analyses?

b. Explore the relationship between survival and the covariates sex and timber harvesting. How does using interval censorship affect the results?

Exercise 5

Here you will explore the same techniques using the data that Gilbert et al. (2014a) published on dryad.

```
# Read the data
deers <- read.csv("Fawn_survival_70day.csv")
# Look at the data
head(deers)
```

```
##    i  j  k fate julcapt sex mass_cap timber new_hoof_mm VIT year
## 1 10 23 24    1     154   1      1.9      1           0   0 2011
## 2 10 69 70    0     154   1      2.5      1           0   0 2011
## 3 14 14 15    1     158   1      2.4      1           0   0 2011
## 4 13 16 17    1     157   1      2.1      0           0   0 2011
## 5 16 69 70    0     160   1      2.1      1           0   0 2011
## 6 18 47 48    1     162   1      2.7      1           0   0 2011
```

In their paper Gilbert et al. (2014b) look at the survival of Sitka neonate deer. Here the column i represents the day of entry in the study, j represents the last day the individual was checked prior to existing the study, k represents the day the individual exited the study, or when the values is 70 the end of the study (so 70 means the data is censored), and fate indicates whether the animal died with 1 meaning the fawn died and 0 it was alive the last time it was checked (again a censored case). Most of the rest of the columns are covariates, with the two covariates we will focus on being sex (0 indicate a male) and timber (1 indicates that the watershed has been commercially harvested for timber).

a. Estimate the mean survival length of the fawns.

c. Investigate whether the survival of fawns is affected by sex and timber harvesting.

5 References

Gilbert SL, Lindberg MS, Hundertmark KJ, Person DK (2014a) Data from: Dead before detection: addressing the effects of left truncation on survival estimation and ecological inference for neonates. Dryad Digital Repository. https://doi.org/10.5061/dryad.p1r40

Gilbert SL, Lindberg MS, Hundertmark KJ, Person DK (2014b) Dead before detection: addressing the effects of left truncation on survival estimation and ecological inference for neonates. Methods in Ecology and Evolution 5(10): 992-1001. https://doi.org/10.1111/2041-210X.12234

Moss WE, Alldredge MW, Pauli JN (2015a) Data from: Quantifying risk and resource use for a large carnivore in an expanding urban-wildland interface. Dryad Digital Repository. https://doi.org/10.5061/dryad.23qp6

Moss WE, Alldredge MW, Pauli JN (2015b) Quantifying risk and resource use for a large carnivore in an expanding urban-wildland interface. Journal of Applied Ecology 53(2): 371-378. https://doi.org/10.1111/1365-2664.12563