Animal Shelter Outcomes for Labradors in Austin Animal Center

Demi Dang
Rosalia Hernandez
Yanjie Qi

11/13/2019

Abstract

Although Labradors are widely viewed as one of the most popular breeds in the United States, they are also a popular breed found in animal shelters https://www.cleartheshelters.com. Taking a look at the outcomes of over 6,000 Labradors in the Austin Animal Center (the largest “no-kill” animal shelter in the U.S.), we build a Cox Proportional Hazards model to see how neuter status affects Labrador adoption rates.

Data Source

The data we used was from the Austin Animal Center, downloaded from kaggle at https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-outcomes-and. The raw data included cats, dogs, birds, livestock and other. Since we were only interested in labradors we created a subset of the data which included 2,128 Labrador and Labrador Mixed breeds for a total of 6,980 observations. We decided to delete one observation which had NULL values for `outcome` and `age upon outcome` since we thought our analysis without this observation would not change. We also excluded 48 data points where the gender and neuter status of the Labrador were not known. We split the column of `sex upon outcome` into two columns for analysis: `sex` with only Male/Female entries and `neuter` with Intact/Fixed entries where "Fixed" means the dog is either neutered or spayed and intact means it was not. Another covariate we changed was `color`. There were a ridiculous amount of "color" descriptions. Our favorite color that we saw was "peach." We decided to change this column to entries of either Mixed fur pattern or Solid fur patterns to denote dogs with more than one color versus dogs that are solid colored. Our final dataset included the following:

- `color`: The descriptive fur pattern of Mixed color fur pattern or solid color fur pattern.
- `event`: The event was 1 if the Labrador had been adopted, and censored was 0 if the Labrador had been transferred, missing (off-site, in foster care or in kennel), possible theft, euthanasia or death.
- `date_diff`: The amount of time in days the Labrador spent in the shelter.
- `sex`: The gender of the dog.
- `neuter`: If the dog is fixed or intact.

Table 1: Final Dataset

color	event	date_diff	sex	neuter
Mixed	1	63 days	Female	Fixed
Mixed	1	95 days	Female	Fixed
Mixed	1	4445 days	Male	Fixed
Mixed	1	745 days	Male	Fixed
Mixed	1	1764 days	Male	Fixed
Mixed	0	1794 days	Male	Fixed

Research Question

On average how much time do Labradors spend in animal shelters before they go to a home?

We will plot a Kaplan-Meier curve and find the median time that Labradors spent in this animal shelter.

Does the color, gender or neuter status of a Labrador affect it's chance for adoption?

We'll build a Cox Proportional Hazards (Cox PH) model with the covariates neuter status, sex, and color to answer this question. We will also explore an Accelerated Failure Time (AFT) model as our extension component, to look at a comparison of survival times.

Data Exploration

Before we start our analysis we must understand our data. We found that we had a total of 6,929 observations. Out of these, 3211 are Female (46%), 3713 have mixed fur pattern (53%), and 5726 were fixed (82%). Using the quantile function on the length of time spent in the shelter, we found that 50% of the dogs spent more than 395 days in the shelter, this is over one year!

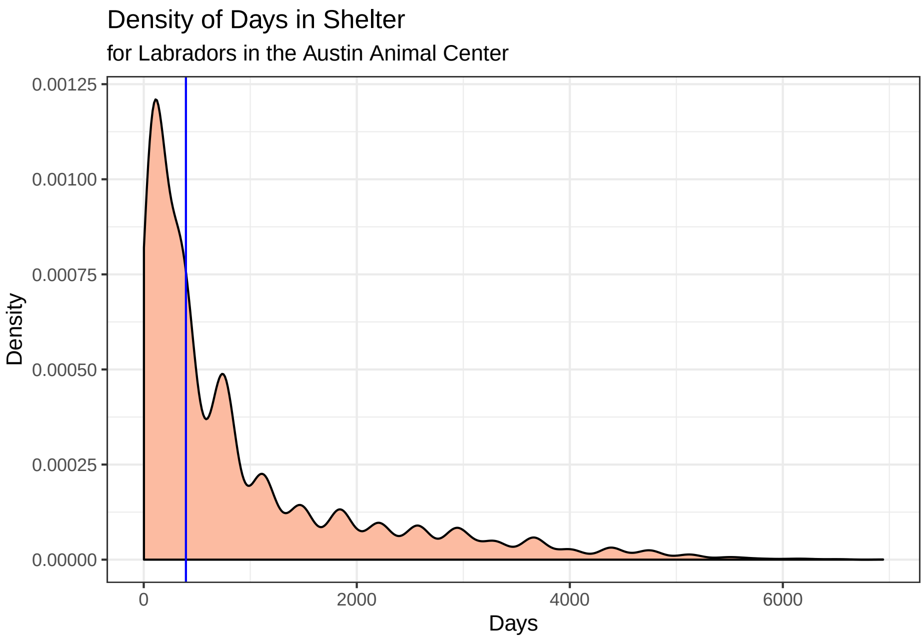


Figure 1: As a preliminary assesment of our data, and using the quantile function we find that half of the Labradors spend more than a year in the animal shelter (395 days, which is shown as the blue line). However, this form includes both censored and uncensored labradors which underestimates our value so we should take this figure lightly.

We plotted a Kaplan-Meier Curve without controlling for any variables. We noticed that the confidence interval curves were tight around the actual survival curve. This proves that the data is very detailed and the information is concrete because we have enough observations to do analysis on Labradors in the Austin Animal Center.

```
## Call: survfit(formula = labradors.surv ~ 1)
##
##        n  events  median 0.95LCL 0.95UCL
##     6929    5285     732     730     734
```

Kaplan-Meier

To visually see the survival curves we must plot Kaplan-Meier Curves for each covariate `color`, `sex` and `neuter` on time to adoption for labradors. The Kaplan-Meier curves for sex in Figure 3 shows that male Labs have higher estimated survival rates since the blue male line is above the red female line. As time increases, male and female Labs have the same or similar survival rates because the lines look like they begin to overlap after the 3,000 days mark. The Kaplan-Meier curves for neuter status in Figure 4 shows that intact Labs (labs that have not been neutered) have higher estimated survival rates than fixed labs. This means that more fixed labs are being adopted or going to homes (having the event) than labs that have not been neutered. The Kaplan-Meier curves for fur patterns in Figure 5 show labs with single colored fur patterns have higher survival rates than those with mixed fur patterns. The gaps immediately widen after the 50% survival rate mark and begin to close again nearer to the bottom.

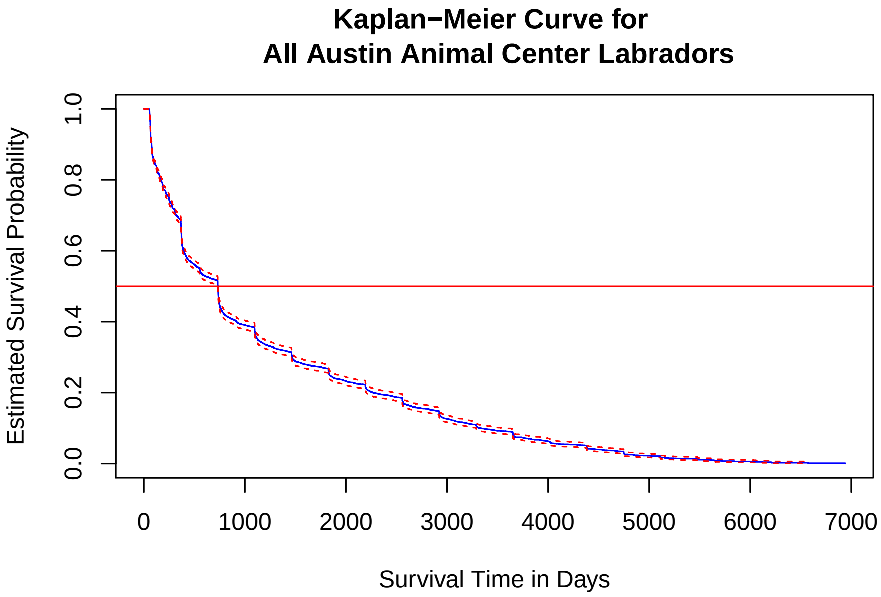


Figure 2: Kaplan-Meier curve

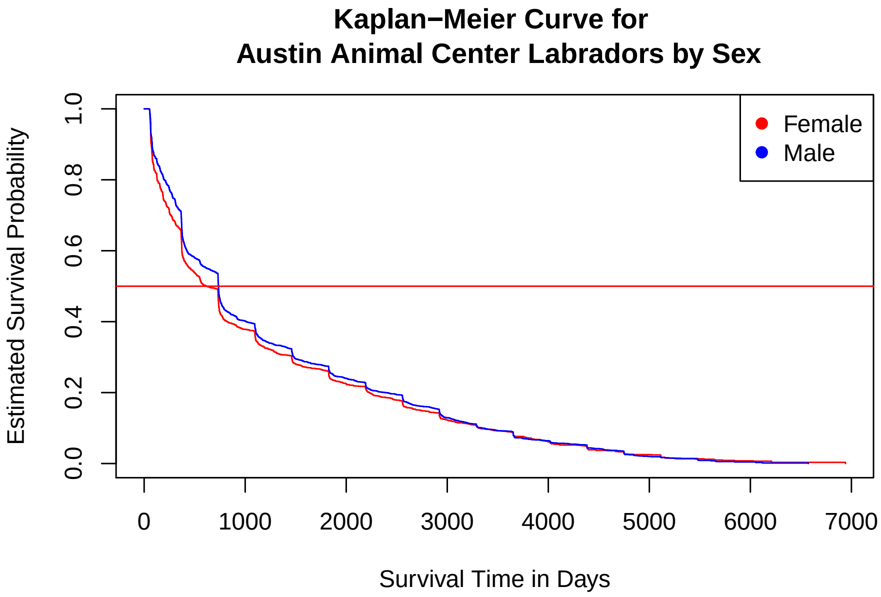


Figure 3: There is not a noticeable difference between the two curves so our pre-analysis hypothesis is that there is not going to be a significant difference.

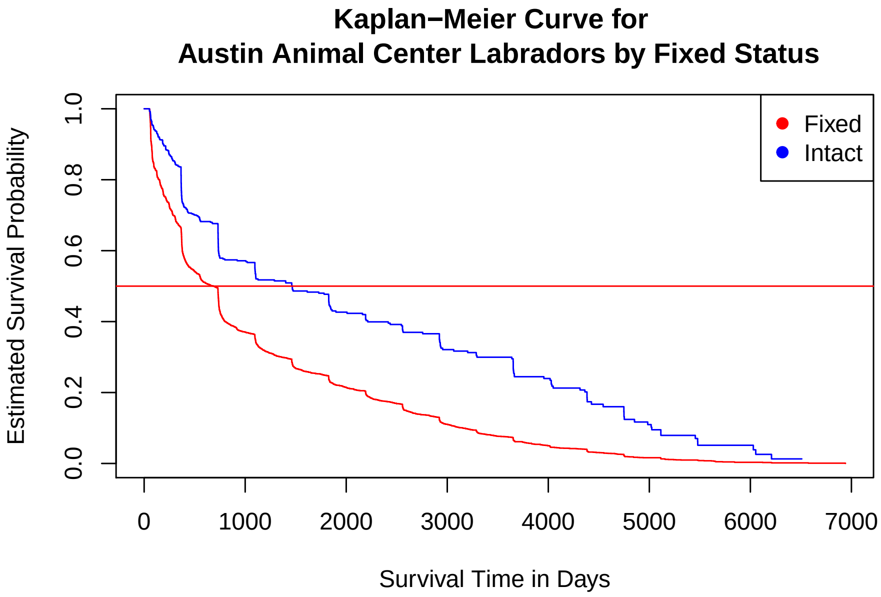


Figure 4: We see a bigger difference in the two survival curves between fixed and intact Labradors, mainly that fixed has a lower survival probability over time. However, this might be because there are many more observations on fixed labs than on intact labs, (fixed has about 4,000 more observations) so we will proceed with caution in our analysis because of this fact.

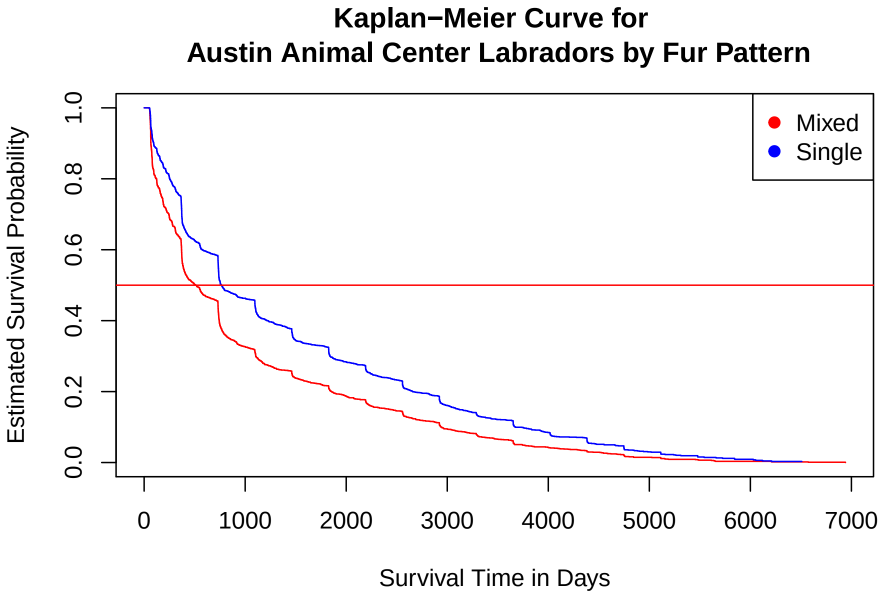


Figure 5: There appears to be an even amount of labs with mixed and solid fur patterns. There seems to be a slight difference in survival probability between the two fur patterns with mixed having a lower survival probability.

Log Rank Tests

The log-rank test allows us to compare 2 or more survival curves, more specifically, that there is a significant difference between the survival curves. The null hypothesis that we are testing is that there is no difference in survival curves. The low p-value means we reject our null hypothesis.

```
## Call:
## survdiff(formula = labradors.surv ~ labradors.df$sex)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## labradors.df$sex=Female 3211     2442     2347      3.87         7
## labradors.df$sex=Male   3718     2843     2938      3.09         7
##
##  Chisq= 7  on 1 degrees of freedom, p= 0.008
```

```
## Call:
## survdiff(formula = labradors.surv ~ labradors.df$neuter)
##
##                               N Observed Expected (O-E)^2/E (O-E)^2/V
## labradors.df$neuter=Fixed  5726     4884     4585      19.5       149
## labradors.df$neuter=Intact 1203      401      700     127.9       149
##
##  Chisq= 149  on 1 degrees of freedom, p= <2e-16
```

```
## Call:
## survdiff(formula = labradors.surv ~ labradors.df$color)
##
##                             N Observed Expected (O-E)^2/E (O-E)^2/V
## labradors.df$color=Mixed 3713     2853     2437      70.9       134
## labradors.df$color=Solid 3216     2432     2848      60.7       134
##
##  Chisq= 134  on 1 degrees of freedom, p= <2e-16
```

After performing log-rank tests for the 3 covariates (`sex`, fur pattern, fixed status) individually, we see that the p-value for each log-rank test is less than $\alpha = 0.05$. The null hypothesis for each test is that the survival curves for the covariate are not significantly different, but since the p-values for each test are less than $\alpha = 0.05$, we can conclude that each of the covariates' survival curves are statistically significantly different.

Model Building

To pick the right covariates for our model we use the anova function on the full model which includes all three covariates that we are examining. The anova function below indicates that each of the covariates are significant to include in our model since each one has a p-value of less than 0.05. Since the last covariate also has a significant p-value then we can conclude that each previous covariate in the model is also included and thus also significant. In anova, order of covariates matter and each new covariate is added to the previous model and evaluated.

```
## Analysis of Deviance Table
##  Cox model: response is Surv(date_diff, event)
## Terms added sequentially (first to last)
##
```

```
##        loglik    Chisq Df Pr(>|Chi|)
## NULL   -41263
## neuter -41177 171.8988  1    < 2e-16 ***
## color  -41120 113.6214  1    < 2e-16 ***
## sex    -41117   6.0749  1    0.01371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also use the backward elimination method in the step function which also indicates that we should include all three of our covariates. The step function uses the Akaike Information Criterion (AIC) to measure each model quality and gives us the best model for the specificied stepwise search (i.e. backward, forward, or both).

```
## Start:  AIC=82239.47
## Surv(date_diff, event) ~ neuter + color + sex
##
##          Df   AIC
## <none>      82239
## - sex     1 82244
## - color   1 82351
## - neuter  1 82390
```

```
## Call:
## coxph(formula = Surv(date_diff, event) ~ neuter + color + sex,
##     data = labradors.df)
##
##                  coef exp(coef) se(coef)       z      p
## neuterIntact -0.59223   0.55309  0.05215 -11.357 <2e-16
## colorSolid   -0.29497   0.74455  0.02778 -10.617 <2e-16
## sexMale      -0.06811   0.93416  0.02761  -2.467 0.0136
##
## Likelihood ratio test=291.6  on 3 df, p=< 2.2e-16
## n= 6929, number of events= 5285
```

Model Checking

In order to build our Cox PH model we must check to see if the proportional hazard assumptions are met. We will do this by graphical evaluation of the log-log plots and by doing a residuals test using the cox.zph() function.

When evaluating the log-log plots we want to make sure that the survival curves don't cross and that the curves look proportional to one another through time. These criteria imply that the proportional hazards assumption is met. Figure 6 shows the log-log plot for the `sex` covariate. We see that the curves cross and overlap both at the beginning and end of the curves. The suggests that the PH assumption is violated. Figure 7 shows the log-log plot for the `neuter` covariate. We see that the survival curves cross in the beginning, but then seem to be proportional until the very end. Because of this, we will also assume the PH assumption is violated. In Figure 8, the log-log plot for the `color` covariate shows an increase in the gap between the curves, and then a decline in the size of the gap through time. This suggests that the hazard proportion is not constant and that the PH assumtions do not hold.

Using the Goodness-of-Fit test (table below) to further evaluate the PH assumptions, we test the null hypothesis that there is no correlation between the residuals and survival times. To do this we use the cox.zph() function in `R`. In the table below we see that for `color` and `sex` we reject the null hypothesis. This

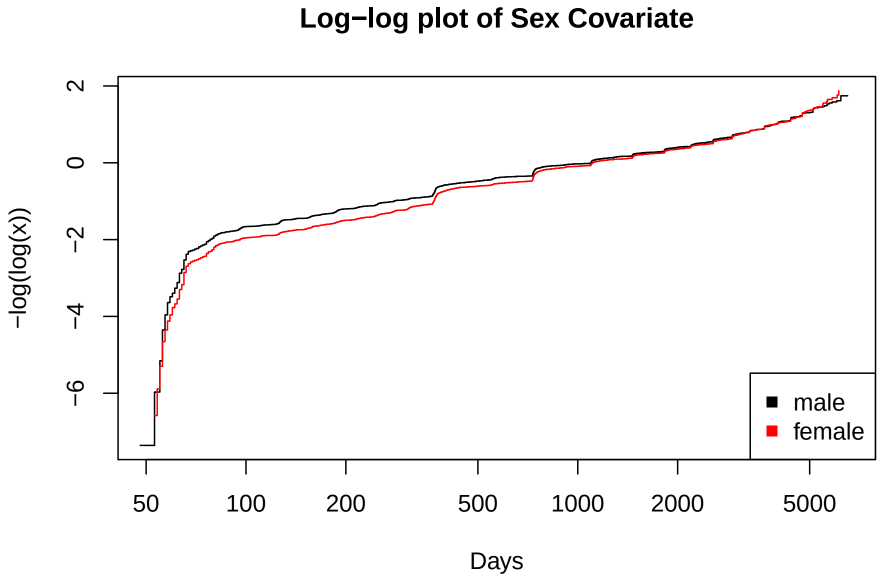


Figure 6: The log-log plot for the sex covariate violates the PH assumptions we are checking because the curves appear to be crossing in various places.

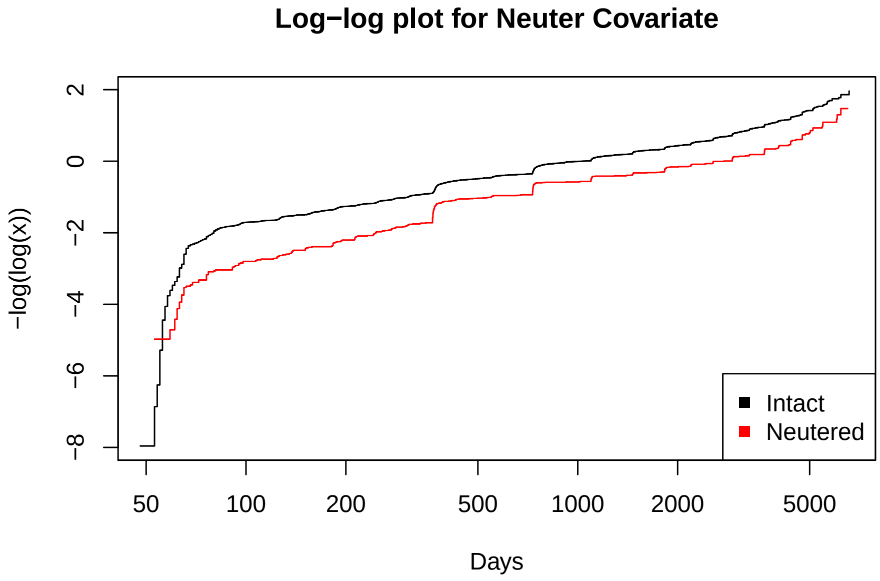


Figure 7: The log-log plot for the neuter covariate appears to be proportional except for the beginning of the curves where they cross a little and it doesn't appear to have constant hazard ratios through time since we see the the gap gets skinnier at the end.

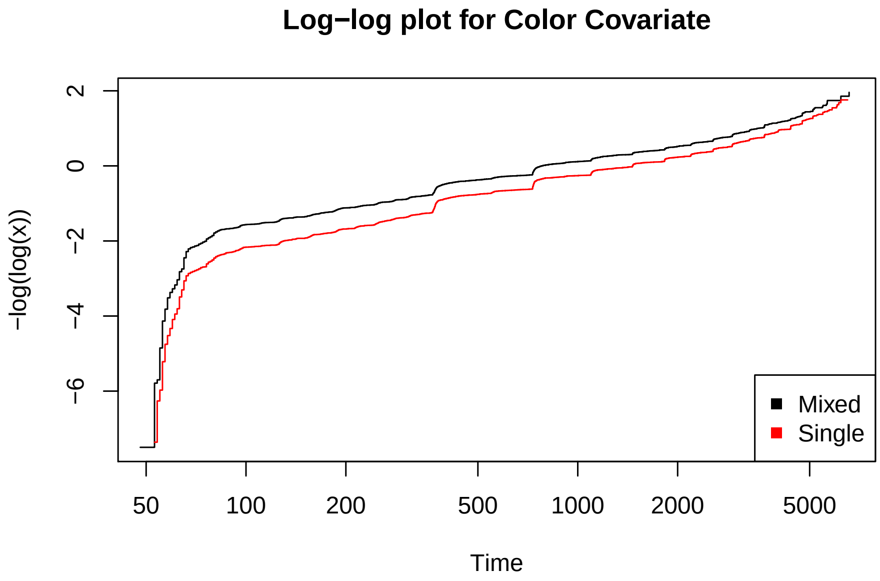


Figure 8: The survival curves appear to get closer to each other as time progresses. This means that the hazard ratio is no longer constant and it depends on time.

means that `color` and `sex` violate the PH assumptions. Notice that we do not reject the null hypothesis for the `neuter` variable which means that PH assumptions hold. Since this is a test and relied more on data than on visual analysis, we will follow the goodness-of-fit test conclusion that the PH assumptions hold for the `neuter` covariate. Our next step is to deal with the covariates that violate the PH assumption. We will do this by stratifying those covariates.

```
##                   rho  chisq        p
## neuterIntact 0.00847  0.379 5.38e-01
## colorSolid   0.07001 25.564 4.28e-07
## sexMale      0.06162 20.026 7.64e-06
```

Stratified Cox PH Model

```
## Call:
## coxph(formula = Surv(date_diff, event) ~ neuter + strata(color) +
##     strata(sex), data = labradors.df)
##
##   n= 6929, number of events= 5285
##
##                  coef exp(coef) se(coef)      z Pr(>|z|)
## neuterIntact -0.5964    0.5508   0.0522 -11.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## neuterIntact    0.5508      1.816    0.4972    0.6101
##
```

```
## Concordance= 0.527  (se = 0.002 )
## Likelihood ratio test= 154.3  on 1 df,   p=<2e-16
## Wald test            = 130.5  on 1 df,   p=<2e-16
## Score (logrank) test = 134.4  on 1 df,   p=<2e-16
```

Interaction Terms

We use the anova function

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(date_diff, event)
##  Model 1: ~ neuter + strata(color) + strata(sex)
##  Model 2: ~ neuter + strata(sex) * strata(color)
##   loglik Chisq Df P(>|Chi|)
## 1 -33898
## 2 -33898     0  0 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

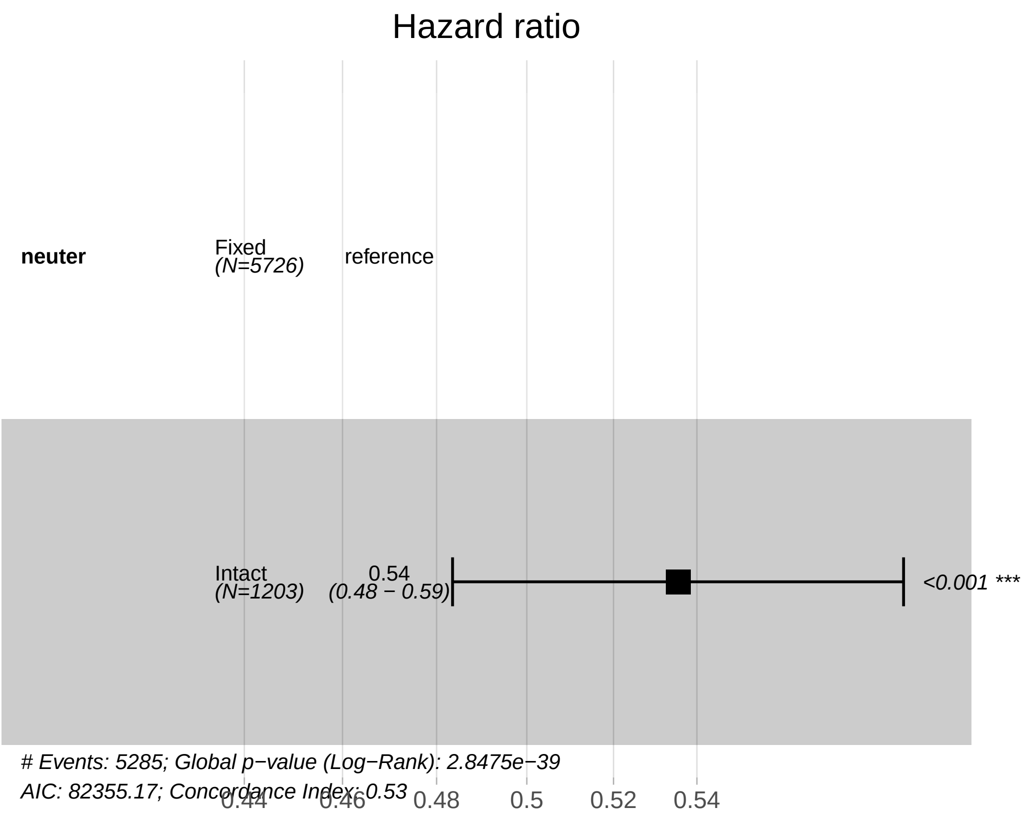
```
## Analysis of Deviance Table
##  Cox model: response is  Surv(date_diff, event)
##  Model 1: ~ neuter + strata(color) + strata(sex)
##  Model 2: ~ strata(color) + strata(sex) * neuter
##   loglik  Chisq Df P(>|Chi|)
## 1 -33898
## 2 -33894 7.5675  1  0.005943 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

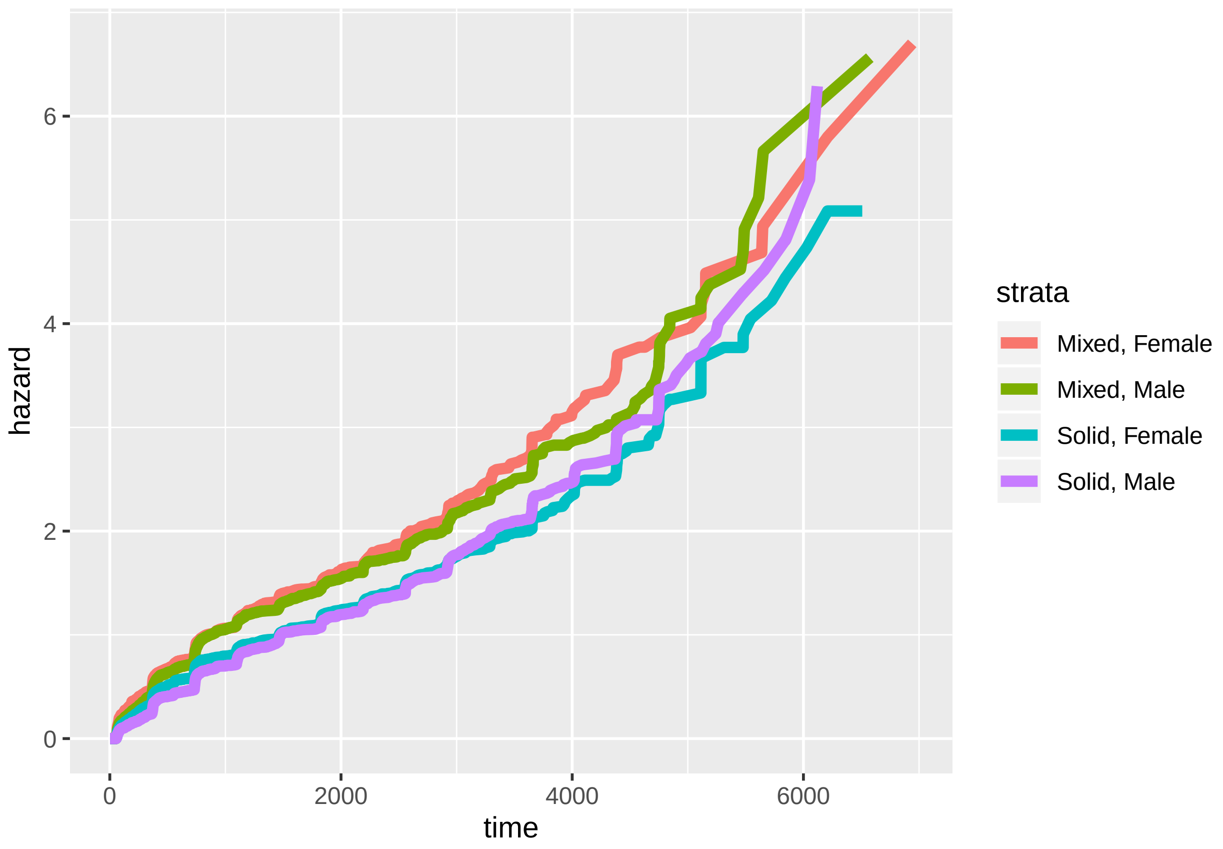
```
## Analysis of Deviance Table
##  Cox model: response is  Surv(date_diff, event)
##  Model 1: ~ neuter + strata(color) + strata(sex)
##  Model 2: ~ strata(sex) + strata(color) * neuter
##   loglik  Chisq Df P(>|Chi|)
## 1 -33898
## 2 -33898 0.0124  1    0.9112
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(date_diff, event)
## Terms added sequentially (first to last)
##
##                    loglik    Chisq Df Pr(>|Chi|)
## NULL               -33975
## neuter             -33898 154.2903  1  < 2.2e-16 ***
## neuter:strata(sex) -33894   7.5675  1   0.005943 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

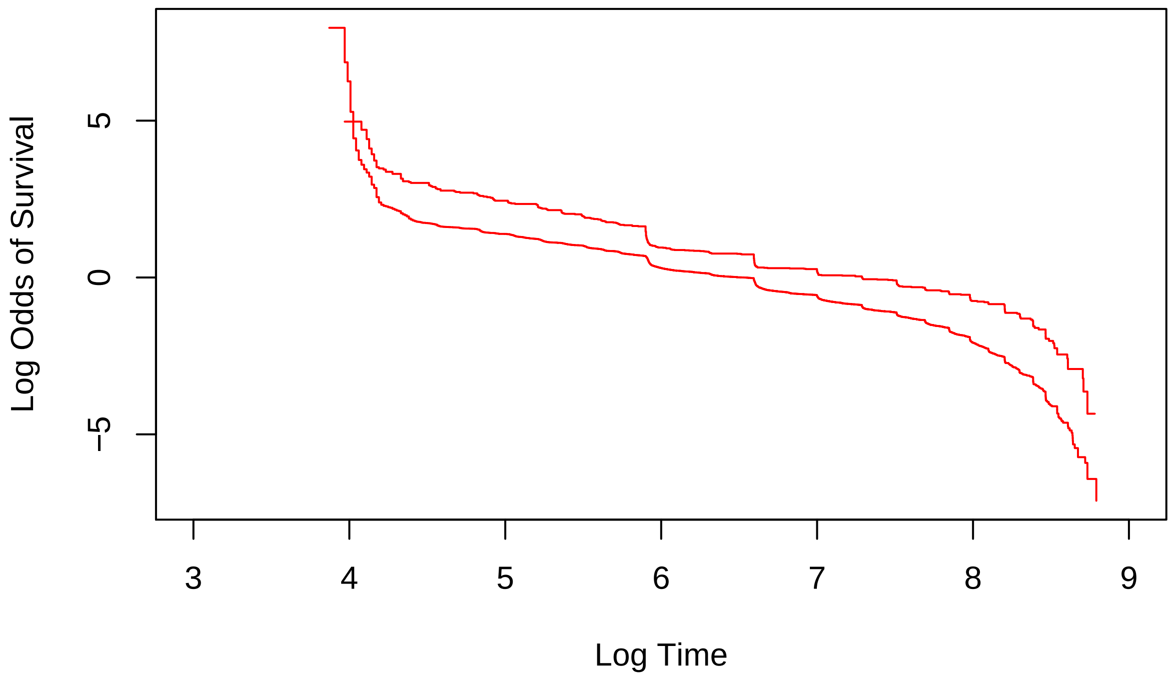
```
## Warning in .get_data(model, data = data): The `data` argument is not
## provided. Data will be extracted from model fit.
```

```
## Warning: Removed 1 rows containing missing values (geom_errorbar).
```


Hazard ratio
neuter
Fixed
(N=5726)
reference
Intact
(N=1203)
0.54
(0.48 − 0.59)
<0.001 ***
# Events: 5285; Global p−value (Log−Rank): 2.8475e−39
AIC: 82355.17; Concordance Index: 0.53
0.44
0.46
0.48
0.5
0.52
0.54


6
4
2
0
hazard
0
2000
4000
6000
time
strata
Mixed, Female
Mixed, Male
Solid, Female
Solid, Male

Extension: Accelerated Failure Times



```
##        Df Deviance Resid. Df    -2*LL     Pr(>Chi)
## NULL   NA       NA      6927 85292.53           NA
## neuter  1 169.3364      6926 85123.20 1.033014e-38
```

```
##      Df Deviance Resid. Df    -2*LL    Pr(>Chi)
## NULL NA       NA      6927 85292.53          NA
## sex   1 7.577082      6926 85284.95 0.005911505
```

```
##       Df Deviance Resid. Df    -2*LL     Pr(>Chi)
## NULL  NA       NA      6927 85292.53           NA
## color  1 145.9612      6926 85146.57 1.323773e-33
```

```
##    Terms Resid. Df    -2*LL Test Df Deviance     Pr(>Chi)
## 1 neuter      6927 85335.44      NA       NA           NA
## 2 neuter      6926 85123.20    =  1  212.243 4.451633e-48
```

```
##   Terms Resid. Df    -2*LL Test Df Deviance     Pr(>Chi)
## 1   sex      6927 85494.30      NA       NA           NA
## 2   sex      6926 85284.95    =  1 209.3457 1.908173e-47
```

```
##   Terms Resid. Df    -2*LL Test Df Deviance     Pr(>Chi)
## 1 color      6927 85324.52      NA       NA           NA
## 2 color      6926 85146.57    =  1 177.9542 1.355545e-40
```

```
##
## Call:
## survreg(formula = Surv(date_diff, event) ~ neuter, data = labradors.df,
##     dist = "loglogistic")
##                Value Std. Error     z      p
## (Intercept)   6.3626     0.0195 326.7 <2e-16
## neuterIntact  0.7392     0.0620  11.9 <2e-16
## Log(scale)   -0.2045     0.0111 -18.5 <2e-16
##
## Scale= 0.815
##
## Log logistic distribution
## Loglik(model)= -42574   Loglik(intercept only)= -42649.5
##  Chisq= 151.06 on 1 degrees of freedom, p= 1e-34
## Number of Newton-Raphson Iterations: 4
## n= 6929
```

```
##
## Call:
## survreg(formula = Surv(date_diff, event) ~ sex, data = labradors.df,
##     dist = "loglogistic")
##               Value Std. Error      z       p
## (Intercept)  6.3480     0.0272 233.06 < 2e-16
## sexMale      0.1708     0.0368   4.64 3.4e-06
## Log(scale)  -0.2064     0.0110 -18.69 < 2e-16
##
## Scale= 0.814
##
## Log logistic distribution
## Loglik(model)= -42638.8   Loglik(intercept only)= -42649.5
##  Chisq= 21.51 on 1 degrees of freedom, p= 3.5e-06
## Number of Newton-Raphson Iterations: 3
## n= 6929
```

```
##
## Call:
## survreg(formula = Surv(date_diff, event) ~ color, data = labradors.df,
##     dist = "loglogistic")
##               Value Std. Error     z      p
## (Intercept)  6.2143     0.0248 250.6 <2e-16
## colorSolid   0.4791     0.0362  13.2 <2e-16
## Log(scale)  -0.2203     0.0110 -20.0 <2e-16
##
## Scale= 0.802
##
## Log logistic distribution
## Loglik(model)= -42563.1   Loglik(intercept only)= -42649.5
##  Chisq= 172.91 on 1 degrees of freedom, p= 1.7e-39
## Number of Newton-Raphson Iterations: 3
## n= 6929
```