Survival Analysis

Linda Tang

9/27/2021

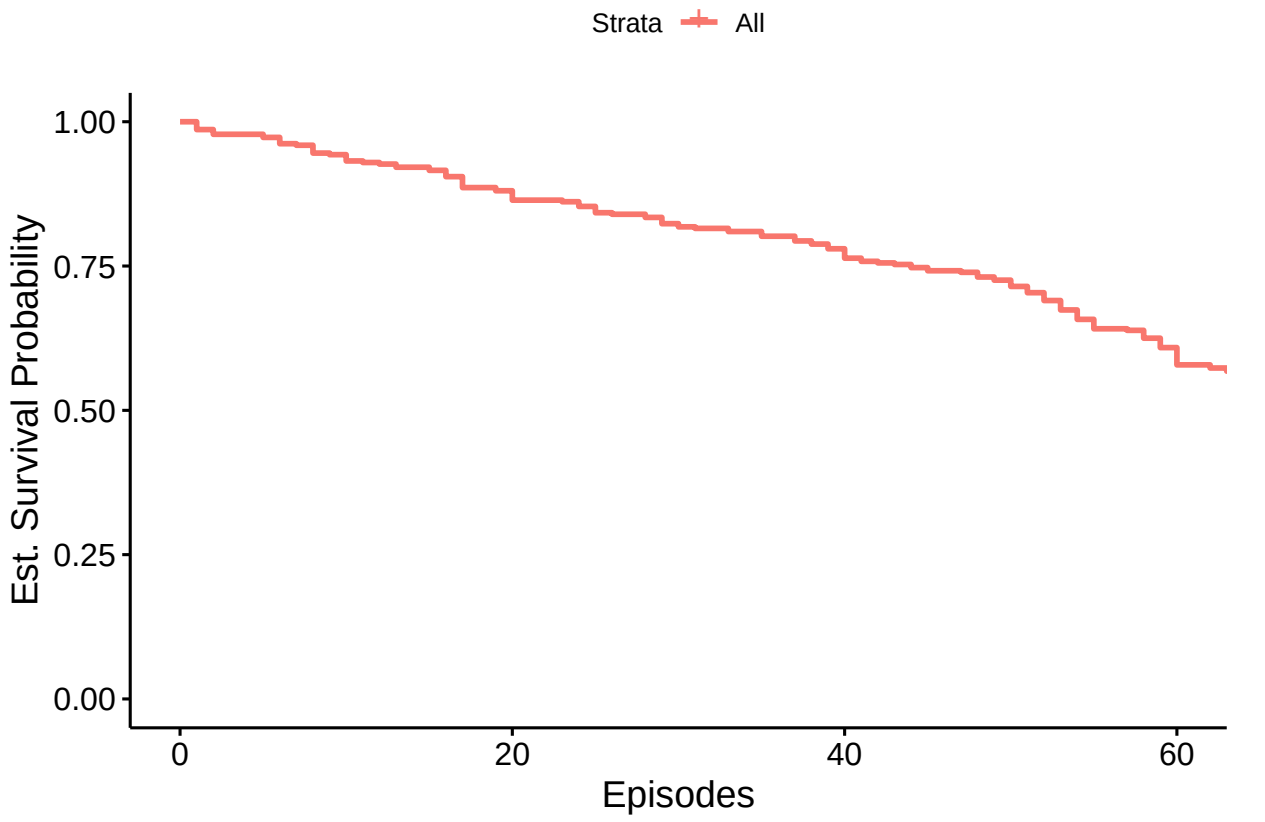
KM Curve Analysis

duration_in_episodes: Survival for characters

is_dead: 1, 0 indicator of death or not

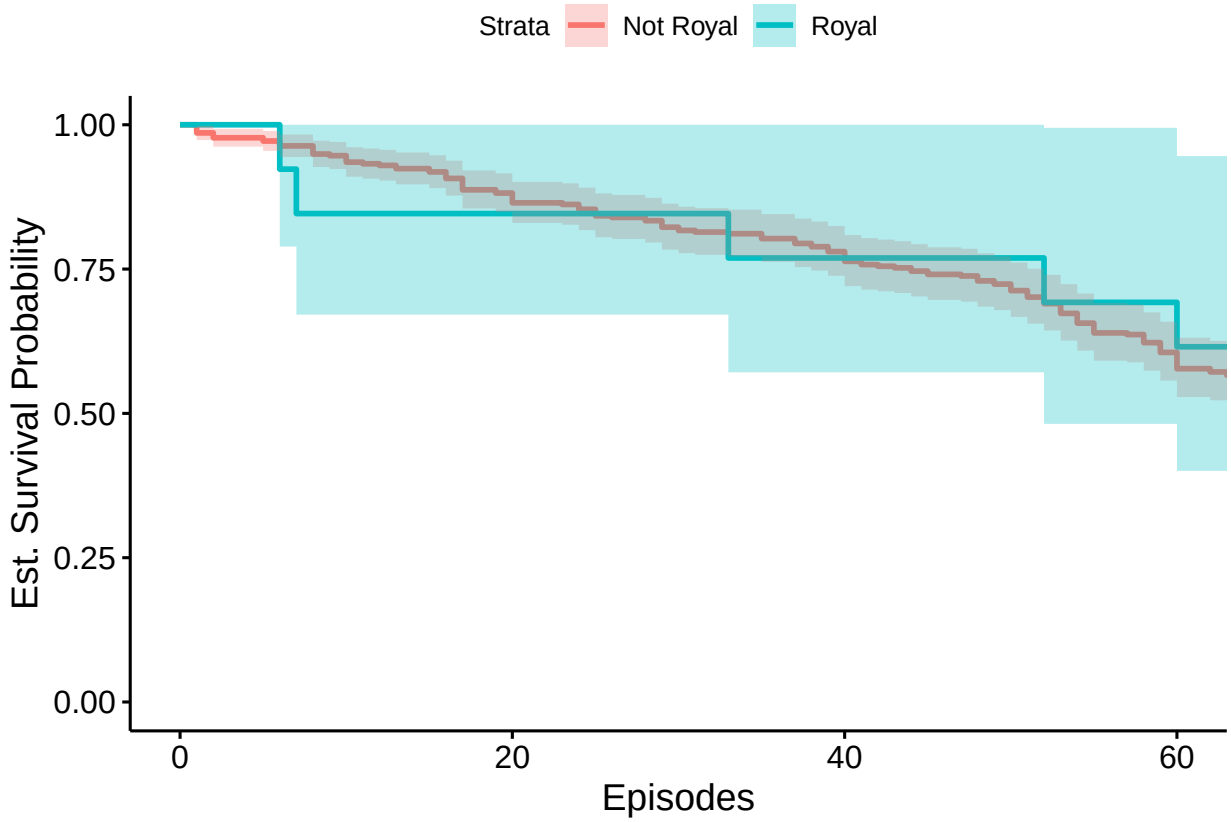
royal, house, gender: Stratas to explore

```
ggsurvplot(survfit(Surv(duration_in_episodes, is_dead) ~ 1, data = got),
     xlab = "Episodes", ylab = "Est. Survival Probability",
     conf.int = F)
```




```
ggsurvplot(survfit(Surv(duration_in_episodes, is_dead) ~ royal,
                   data = got),
     xlab = "Episodes", ylab = "Est. Survival Probability",
     ylim = c(0, 1),
```

```
    conf.int = T, censor = F,
    legend.labs = c("Not Royal", "Royal"))
```

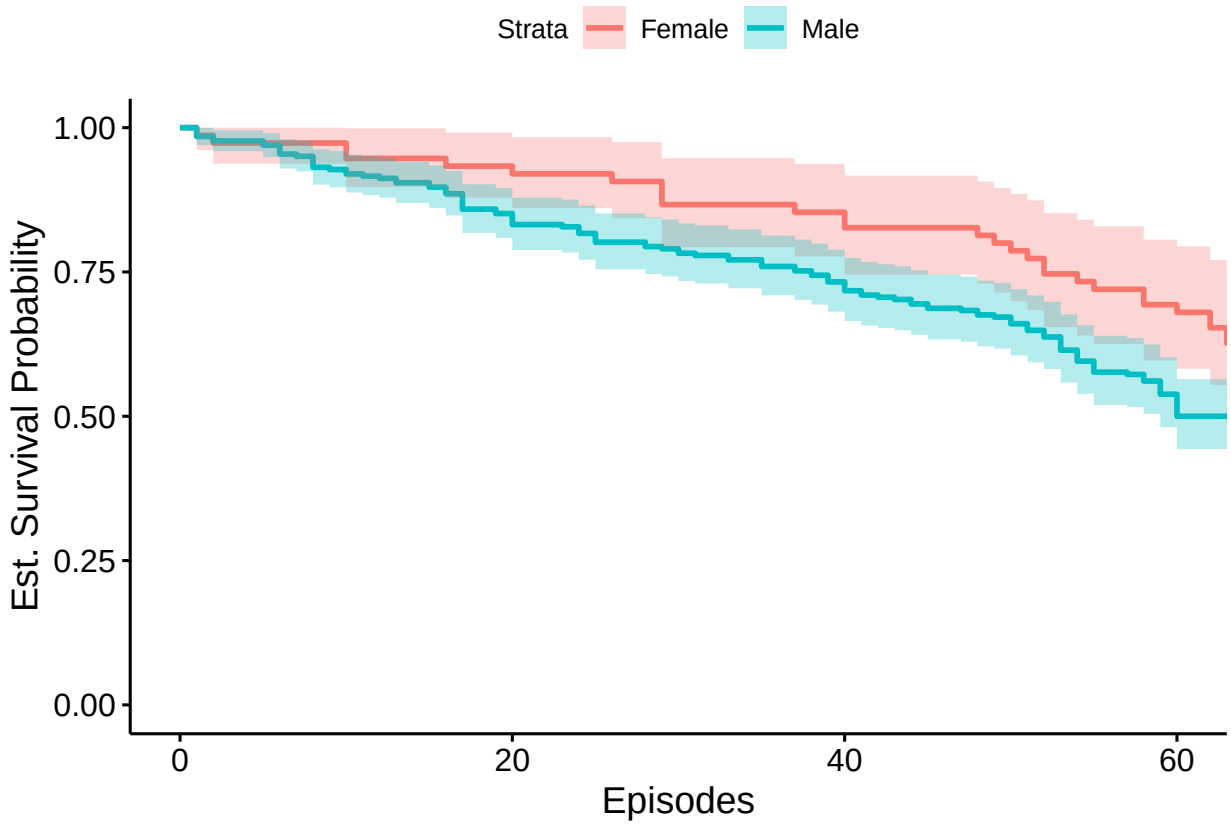



```
survdiff(Surv(duration_in_episodes, is_dead) ~ royal, data = got)
```

```
## Call:
## survdiff(formula = Surv(duration_in_episodes, is_dead) ~ royal,
##     data = got)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## royal=0 355      160   159.08    0.0053      0.15
## royal=1  13        5     5.92    0.1423      0.15
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.7
```

p_value = 0.7, there's no statistically significant difference.

```
ggsurvplot(survfit(Surv(duration_in_episodes, is_dead) ~ gender,
                   data = got),
    xlab = "Episodes", ylab = "Est. Survival Probability",
    ylim = c(0, 1),
    conf.int = T, censor = F,
    legend.labs = c("Female", "Male"))
```




```
survdiff(Surv(duration_in_episodes, is_dead) ~ gender, data = got)
```

```
## Call:
## survdiff(formula = Surv(duration_in_episodes, is_dead) ~ gender,
##     data = got)
##
## n=337, 31 observations deleted due to missingness.
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## gender=female  75       28     40.8      4.03      5.44
## gender=male   262      137    124.2      1.33      5.44
##
##  Chisq= 5.4  on 1 degrees of freedom, p= 0.02
```

p_value = 0.02, there is a statistically significant difference.

```
survdiff(Surv(duration_in_episodes, is_dead) ~ house, data = got)
```

```
## Call:
## survdiff(formula = Surv(duration_in_episodes, is_dead) ~ house,
##     data = got)
##
## n=86, 282 observations deleted due to missingness.
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## house=Arryn;Tully          1        1    0.190     3.451     3.508
```

```
## house=Baratheon             9        8    2.944     8.684     9.657
## house=Bolton                2        2    0.882     1.416     1.482
## house=Bolton;Frey           1        1    0.388     0.967     0.998
## house=Frey                  6        3    4.447     0.471     0.547
## house=Greyjoy               5        2    3.254     0.483     0.540
## house=Lannister             9        6    4.316     0.657     0.756
## house=Lannister;Baratheon   1        0    0.815     0.815     0.871
## house=Martell               8        6    4.642     0.398     0.459
## house=Mormont               3        1    1.727     0.306     0.332
## house=Stark                20        8   12.362     1.539     2.160
## house=Stark;Targaryen       1        1    0.388     0.967     0.998
## house=Stark;Tully           1        1    0.136     5.497     5.580
## house=Targaryen            11        3    7.590     2.776     3.466
## house=Tarly                 1        0    0.815     0.815     0.871
## house=Tully                 3        2    1.339     0.326     0.348
## house=Tyrell                4        4    2.766     0.551     0.620
##
##  Chisq= 32.3  on 16 degrees of freedom, p= 0.009
```

p_value = 0.009 , there's statistical evidence that the house has an impact on survival of characters (at least one house is different).

Modeling

royal, house, gender: confounding to control

Exponential Model AFT model:

```
aft_exp <- survreg(Surv(duration_in_episodes, is_dead) ~ gender + royal + house,
                   data = got, dist = "exponential")
summary(aft_exp)
```

```
##
## Call:
## survreg(formula = Surv(duration_in_episodes, is_dead) ~ gender +
##     royal + house, data = got, dist = "exponential")
##                               Value Std. Error     z       p
## (Intercept)                 3.6109     1.0000  3.61 0.00031
## gendermale                 -0.8908     0.3937 -2.26 0.02366
## royal                       0.0546     0.6216  0.09 0.93003
## houseBaratheon              0.5964     1.1401  0.52 0.60091
## houseBolton                 1.2962     1.2865  1.01 0.31365
## houseBolton;Frey            0.3403     1.4142  0.24 0.80983
## houseFrey                   1.8916     1.1988  1.58 0.11457
## houseGreyjoy                2.1795     1.2747  1.71 0.08730
## houseLannister              1.4529     1.1404  1.27 0.20265
## houseLannister;Baratheon   18.5392  8103.0840  0.00 0.99817
## houseMartell                0.9862     1.1034  0.89 0.37147
## houseMormont                2.0547     1.4462  1.42 0.15538
## houseStark                  1.8626     1.1091  1.68 0.09307
## houseStark;Targaryen        1.2311     1.4680  0.84 0.40168
## houseStark;Tully           -0.2436     1.4142 -0.17 0.86323
```

```
## houseTargaryen              2.0194     1.2197  1.66 0.09779
## houseTarly                 18.5938  8103.0840  0.00 0.99817
## houseTully                  1.5839     1.2865  1.23 0.21825
## houseTyrell                 1.0321     1.1519  0.90 0.37023
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -256.9   Loglik(intercept only)= -266.6
##  Chisq= 19.45 on 18 degrees of freedom, p= 0.36
## Number of Newton-Raphson Iterations: 17
## n=78 (290 observations deleted due to missingness)
```

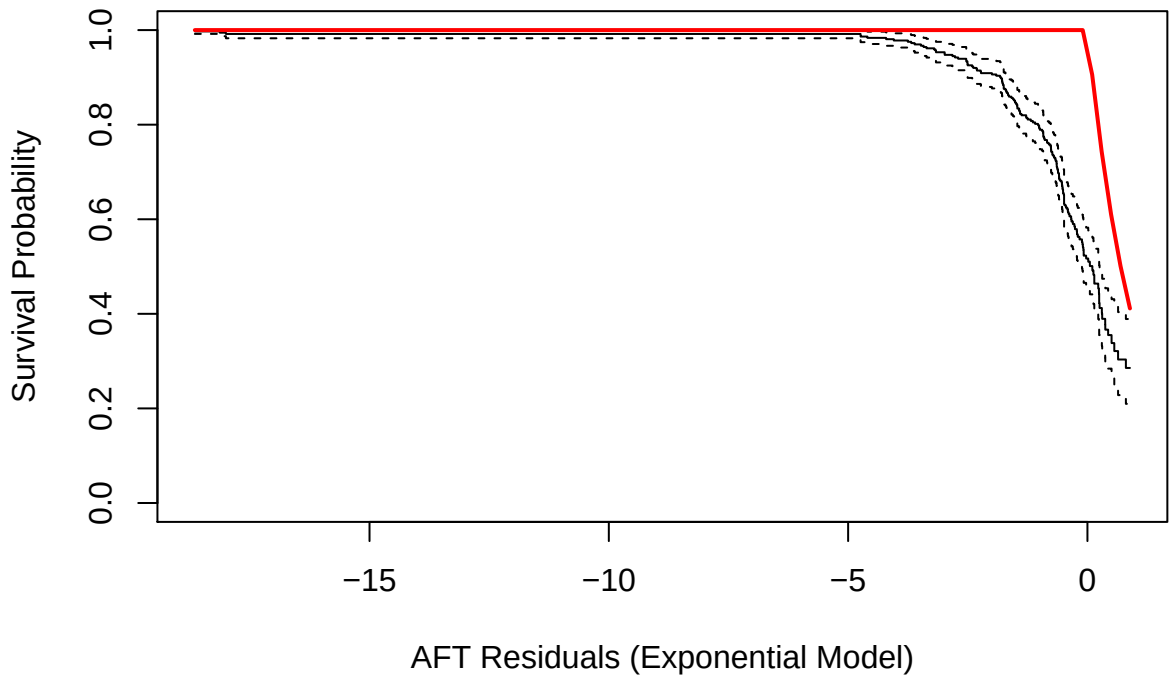
The p_value for I(gender = male) is 0.02366, which means it's a statistically significant predictor. Holding all else equal, a male is expected to survive approximately 0.41 times longer than a female. The probability that males survive to time 0.41t is the same females surviving to 1t.

```
resids <- (log(got$duration_in_episodes) - aft_exp$linear.predictors) /
  (aft_exp$scale)
```

```
## Warning in log(got$duration_in_episodes) - aft_exp$linear.predictors: longer
## object length is not a multiple of shorter object length
```

```
m1 <- survfit(Surv(resids, is_dead) ~ 1, data = got)
plot(m1, xlab = "AFT Residuals (Exponential Model)",
     ylab = "Survival Probability")

exp.x <- seq(min(resids), max(resids), length = 100)
exp.y <- pexp(exp.x, lower.tail = F) # F(t)
lines(exp.x, exp.y, col = "red", lwd = 2)
```



The residuals doesn't overlap well with the survival function of the assumed distribution, so the Exponential model doesn't fit very well.

Weibull Model AFT model:

```
aft_w <- survreg(Surv(duration_in_episodes, is_dead) ~ gender + royal + house,
                 data = got, dist = "weibull")
summary(aft_w)
```

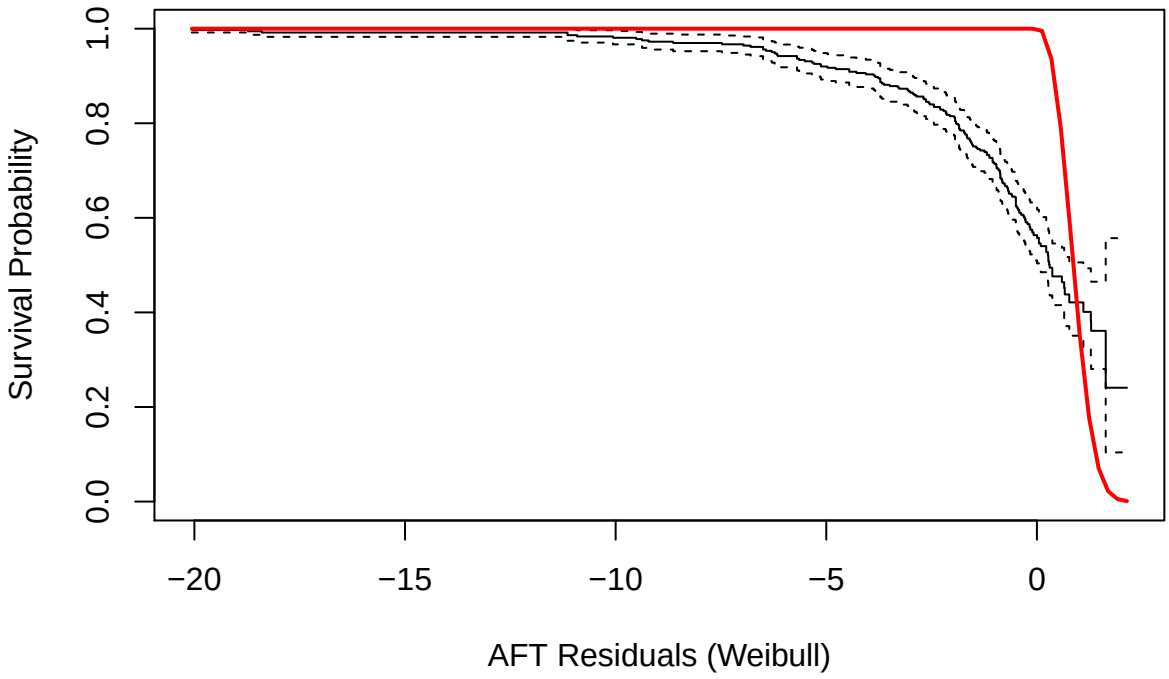
```
##
## Call:
## survreg(formula = Surv(duration_in_episodes, is_dead) ~ gender +
##     royal + house, data = got, dist = "weibull")
##                               Value Std. Error     z       p
## (Intercept)                  3.6109     0.3917  9.22 < 2e-16
## gendermale                  -0.4306     0.1623 -2.65   0.008
## royal                        0.0707     0.2461  0.29   0.774
## houseBaratheon               0.4219     0.4482  0.94   0.347
## houseBolton                  0.8392     0.5064  1.66   0.098
## houseBolton;Frey             0.3403     0.5540  0.61   0.539
## houseFrey                    1.1302     0.4756  2.38   0.017
## houseGreyjoy                 1.2392     0.5070  2.44   0.015
## houseLannister               0.8957     0.4491  1.99   0.046
## houseLannister;Baratheon     7.7323  3882.6306  0.00   0.998
## houseMartell                 0.6648     0.4316  1.54   0.123
## houseMormont                 1.1579     0.5712  2.03   0.043
## houseStark                   1.1096     0.4402  2.52   0.012
## houseStark;Targaryen         0.7710     0.5773  1.34   0.182
## houseStark;Tully            -0.2436     0.5540 -0.44   0.660
## houseTargaryen               1.1849     0.4836  2.45   0.014
## houseTarly                   7.8030  3882.6306  0.00   0.998
## houseTully                   0.9737     0.5071  1.92   0.055
## houseTyrell                  0.7616     0.4537  1.68   0.093
## Log(scale)                  -0.9371     0.1327 -7.06 1.6e-12
##
## Scale= 0.392
##
## Weibull distribution
## Loglik(model)= -238.6   Loglik(intercept only)= -253.6
##  Chisq= 29.96 on 18 degrees of freedom, p= 0.038
## Number of Newton-Raphson Iterations: 19
## n=78 (290 observations deleted due to missingness)
```

```
resids <- (log(got$duration_in_episodes) - aft_w$linear.predictors) /
  (aft_w$scale)
```

```
## Warning in log(got$duration_in_episodes) - aft_w$linear.predictors: longer
## object length is not a multiple of shorter object length
```

```
m1 <- survfit(Surv(resids, is_dead) ~ 1, data = got)
plot(m1, xlab = "AFT Residuals (Weibull)",
     ylab = "Survival Probability")

exp.x <- seq(min(resids), max(resids), length = 100)
exp.y <- pweibull(exp.x, shape = 1/aft_w$scale, lower.tail = F) # F(t)
lines(exp.x, exp.y, col = "red", lwd = 2)
```



The residuals doesn't overlap well with the survival function of the assumed distribution, so the Weibull model doesn't fit very well.