Lin_ST625_HW6

Frances Lin

2/12/2022

1

Group A: 143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246, 265, 304, 216+, 244+

Group B: 142, 156, 173, 198, 205, 232, 232, 233, 233, 233, 233, 239, 240, 261, 280, 280, 296, 296, 323, 204+, 344+

1a

Survival table for Group A:

time	survival	failure	Survival.Std.Err	No.Left	No.Failed	No.Censored
0	1	0	0	19	0	0
143	0.9474	0.05263	0.05123	19	1	0
164	0.8947	0.1053	0.07041	18	1	0

Survival table for Group B:

time	survival	failure	Survival.Std.Err	No.Left	No.Failed	No.Censored
0	1	0	0	21	0	0
142	0.9524	0.04762	0.04647	21	1	0
156	0.9048	0.09524	0.06406	20	1	0

Survival function for Group A using KM method:

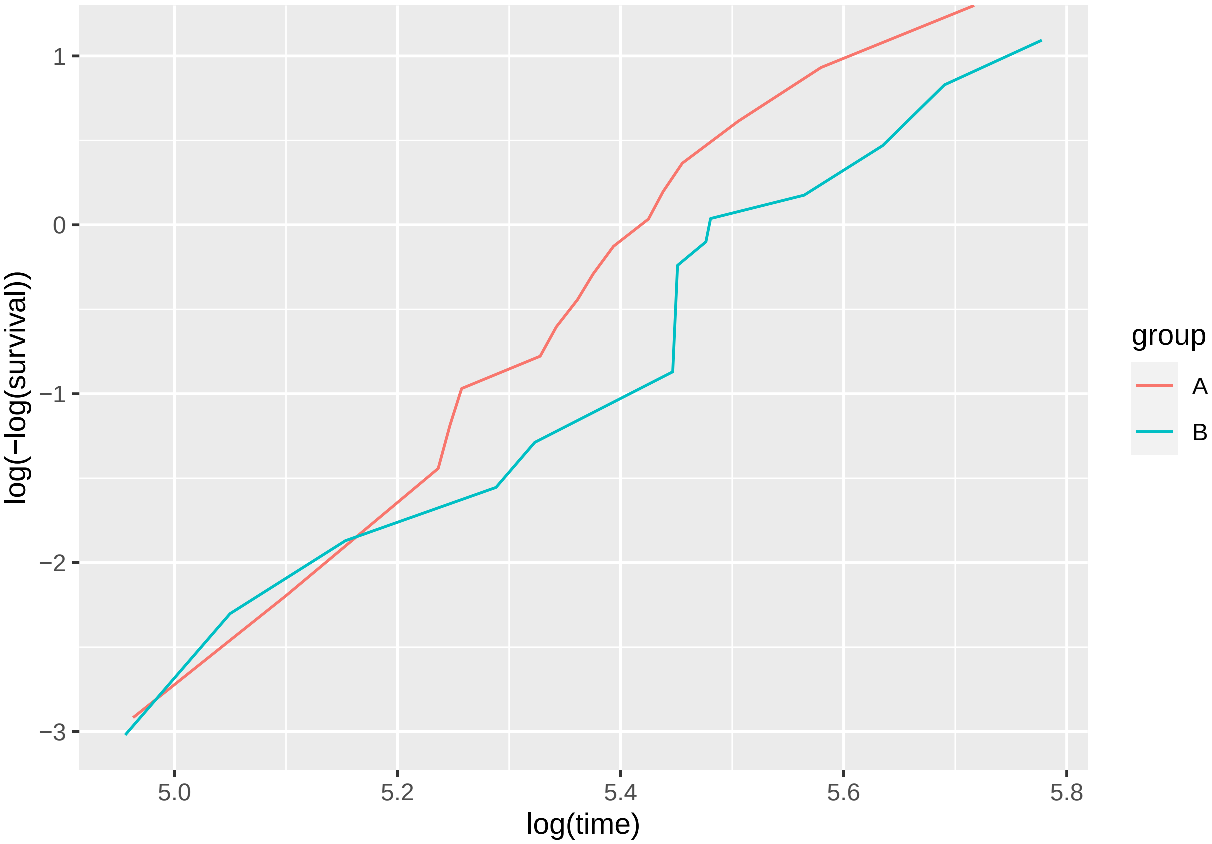
```
##  [1] 1.00000000 0.94736842 0.89473684 0.78947368 0.73684211 0.68421053
##  [7] 0.63157895 0.57894737 0.52631579 0.47368421 0.41447368 0.35526316
## [13] 0.29605263 0.23684211 0.23684211 0.15789474 0.07894737 0.00000000
```

Survival function for Group B using KM method:

```
##  [1] 1.00000000 0.95238095 0.90476190 0.85714286 0.80952381 0.80952381
##  [7] 0.75892857 0.65773810 0.45535714 0.40476190 0.35416667 0.30357143
## [13] 0.20238095 0.10119048 0.05059524 0.05059524
```

1b

My guess is that we need to check hazard $h(t)$ or cumulative hazard function $H(t)$. Let's find out! I went with plot of `log(-log(survival function))` vs `log(survival time)` instead.



They don't appear quite linear to me so Weibull distribution may not be a good fit to the survival data. Recall that if the Weibull distribution can describe the survival time, then the `log(-log(survival function))` is a linear function of the `log(survival time)`.

1c

Weibull model results for Group A:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ 1, data = A, dist = "weibull")
##               Value Std. Error     z      p
## (Intercept)  5.4567     0.0412 132.6 <2e-16
## Log(scale)  -1.8055     0.1756 -10.3 <2e-16
##
## Scale= 0.164
##
## Weibull distribution
## Loglik(model)= -88.2   Loglik(intercept only)= -88.2
## Number of Newton-Raphson Iterations: 6
## n= 19
```

Weibull model results for Group B:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ 1, data = B, dist = "weibull")
##               Value Std. Error      z      p
## (Intercept)  5.578      0.047 118.57 <2e-16
## Log(scale)  -1.615      0.176  -9.16 <2e-16
##
## Scale= 0.199
##
## Weibull distribution
## Loglik(model)= -104.6   Loglik(intercept only)= -104.6
## Number of Newton-Raphson Iterations: 6
## n= 21
```

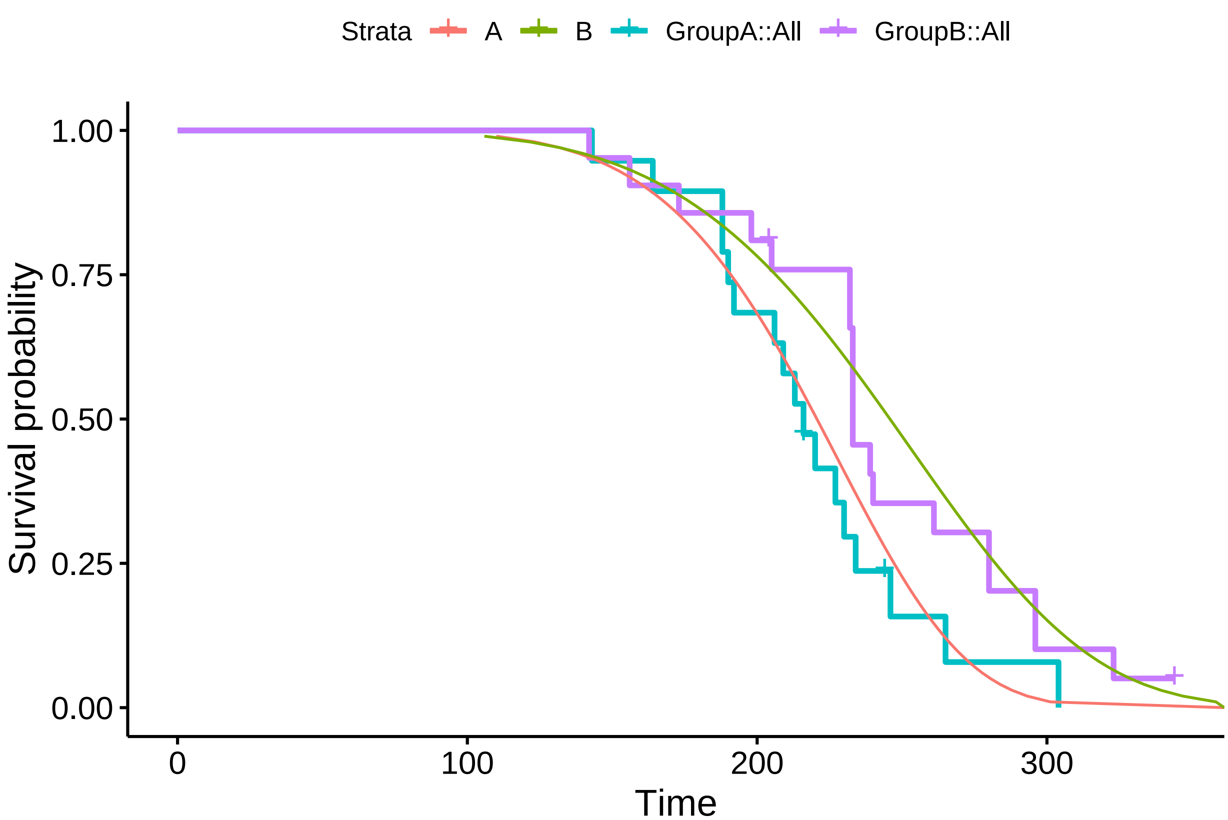
The Scale parameter for Group A is estimated as

```
## [1] 0.1643886
```

The Scale parameter for Group B is estimated as

```
## [1] 0.198977
```

1d



Weibull distribution appears to fit the data (both groups) well.

2

2a

Weibull model results:

```
##
## Call:
## survreg(formula = Surv(time, fstat) ~ factor(gender) + age +
##     factor(cvd), data = Data_1, dist = "weibull")
##                   Value Std. Error     z      p
## (Intercept)     17.5337     1.1010 15.93 <2e-16
## factor(gender)1  0.0741     0.2775  0.27   0.79
## age             -0.1276     0.0136 -9.40 <2e-16
## factor(cvd)1    -0.0632     0.3335 -0.19   0.85
## Log(scale)       0.6862     0.0602 11.40 <2e-16
##
## Scale= 1.99
##
## Weibull distribution
## Loglik(model)= -1684.4   Loglik(intercept only)= -1752.5
##  Chisq= 136.15 on 3 degrees of freedom, p= 2.6e-29
## Number of Newton-Raphson Iterations: 5
## n= 500
```

2b

It is estimated that the mean (or median) survival time will decrease by 12.76% (or $exp(-0.1276) = 0.8802054$ times) with one unit increase of `Age`. `Age` has significant effect on the survival time (p-val < 2e-16).

```
## [1] 0.8802054
```

2c

It is estimated that the mean (or median) survival time for Male (`gender = 0`) is $exp(0.0741) = 1.076914$ times of the mean (or median) survival time of Female (`gender = 1`). However, this improvement is not statistically significant (p-val = 0.79).

```
## [1] 1.076914
```

2d

It is estimated that the mean (or median) survival time for those with no Cardiovascular (`cvd = 0`) is $exp(-0.0632) = 0.9387557$ times of the mean (or median) survival time of those with Cardiovascular (`cvd = 1`). However, this reduction is not statistically significant (p-val = 0.85).

```
## [1] 0.9387557
```

3

group	time	status	DVAL	FVAL
Green	19	0	43	85
Blue	88	0	33	63
Green	23	0	45	77
Blue	89	0	38	41
Blue	24	0	45	51
Green	91	0	49	77

3a

Weibull model results:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ group + DVAL + FVAL, data = Data_2,
##     dist = "weibull")
##                Value Std. Error     z       p
## (Intercept)  6.22977    0.62828  9.92 < 2e-16
## groupGreen   0.37572    0.18361  2.05   0.041
## DVAL        -0.01732    0.01069 -1.62   0.105
## FVAL        -0.01452    0.00775 -1.87   0.061
## Log(scale)  -1.08943    0.21525 -5.06 4.2e-07
##
## Scale= 0.336
##
## Weibull distribution
## Loglik(model)= -95.6   Loglik(intercept only)= -99.8
##  Chisq= 8.5 on 3 degrees of freedom, p= 0.037
## Number of Newton-Raphson Iterations: 8
## n= 80
```

The Scale parameter is estimated as

```
## [1] 0.3364076
```

Or just the estimates:

```
## (Intercept)  groupGreen        DVAL        FVAL
##  6.22976810  0.37572496 -0.01732475 -0.01451836
```

3b

Failure times depend significantly only on the `group` (p-val = 0.041).

i.e. Neither `DVAL` (p-val = 0.105) nor `FVAL` (p-val = 0.061) is significant at the $\alpha = 0.05$ level.

3c

Blue `group` has the longer expected survival time. It is estimated that the mean (or median) failure time for Blue `group` is $exp(0.375725) = 1.456047$ times of the mean (or median) failure time of Green `group`.

```
## (Intercept)  groupGreen        DVAL        FVAL
##  6.22976810  0.37572496 -0.01732475 -0.01451836
```

```
## (Intercept)  groupGreen        DVAL        FVAL
## 507.6377482   1.4560466   0.9828245   0.9855865
```

3d

Exponential model results:

```
##
## Call:
## survreg(formula = Surv(time, status) ~ group + DVAL + FVAL, data = Data_2,
##     dist = "exponential")
##               Value Std. Error     z       p
## (Intercept)  8.6416     1.7230  5.02 5.3e-07
## groupGreen   0.5091     0.5369  0.95    0.34
## DVAL        -0.0356     0.0295 -1.21    0.23
## FVAL        -0.0288     0.0226 -1.27    0.20
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -103.9   Loglik(intercept only)= -105.8
##  Chisq= 3.73 on 3 degrees of freedom, p= 0.29
## Number of Newton-Raphson Iterations: 5
## n= 80
```

Or just the estimates:

```
## (Intercept)  groupGreen        DVAL        FVAL
##  8.64155621  0.50905096 -0.03559472 -0.02878394
```

3e

None of the covariates is significant at the $\alpha = 0.05$ level.

3f

Failure times no longer depends significantly on the `group` (p-val = 0.34).

Assuming exponential instead of Weibull distribution appears to decrease the significance of the covariates.

```
##  (Intercept)   groupGreen         DVAL         FVAL   Log(scale)
## 3.561182e-23 4.072363e-02 1.052011e-01 6.113148e-02 4.165508e-07
```

```
##  (Intercept)   groupGreen         DVAL         FVAL
## 5.289212e-07 3.430482e-01 2.269533e-01 2.026821e-01
```

3g

There is a strong evidence that Weibull distribution fits the data better than the Exponential model (test statistic = 16.68065 = 2 * ($LL_{Weibull} - LL_{Exponential}$) = $2*(-95.57189-(-103.9122))$, p-val = 4.422983e-05 = $P(\chi^2_1 \geq$ test statistic).) A Weibull regression model (full model) is preferred.

```
## [1] 16.68065
```

```
## [1] 4.422983e-05
```