BSTA 665 HW #3

Y. Paulsen

4/26/2021

Assignment:

Question a

Create a SAS dataset including all these variables and a new variable "Type" as the combination of GRAFT and Disease. Attach your SAS code here.

Below is the code to input the data and create a new variable.

The new variable "type" is as follows.

Table 1: Variable "Type"

Type	Graft	Disease
Type 1	Allo	NHL
Type 2	Allo	HOD
Type 3	Auto	NHL
Type 4	Auto	HOD

```
data leuk;
    input graft$ disease$ time status score wait;
    datalines;
1 1 28 1 90 24
1 1 32 1 30 7
1 1 49 1 40 8
1 1 84 1 60 10
1 1 357 1 70 42
1 1 933 0 90 9
1 1 1078 0 100 16
1 1 1183 0 90 16
1 1 1560 0 80 20
1 1 2114 0 80 27
1 1 2144 0 90 5
1 2 2 1 20 34
1 2 4 1 50 28
1 2 72 1 80 59
1 2 77 1 60 102
1 2 79 1 70 71
2 1 42 1 80 19
2 1 53 1 90 17
2 1 57 1 30 9
2 1 63 1 60 13
2 1 81 1 50 12
```

```
2 1 140 1 100 11
2 1 81 1 50 12
2 1 252 1 90 21
2 1 524 1 90 39
2 1 210 0 90 16
2 1 476 0 90 24
2 1 1037 0 90 84
2 2 30 1 90 73
2 2 36 1 80 61
2 2 41 1 70 34
2 2 52 1 60 18
2 2 62 1 90 40
2 2 108 1 70 65
2 2 132 1 60 17
2 2 180 0 100 61
2 2 307 0 100 24
2 2 406 0 100 48
2 2 446 0 100 52
2 2 484 0 90 84
2 2 748 0 90 171
2 2 1290 0 90 20
2 2 1345 0 80 98
run;

/*
Create a new variable 'type' which combines the information from the first two
columns.
*/
data leuk;
    set leuk;
    if graft='1' and disease='1' then type='1';
    else if graft='1' and disease='2' then type='2';
    else if graft='2' and disease='1' then type='3';
    else type='4';
run;
proc print data=leuk;
run;
```

Question b

Fit the Accelerated Failure Time (AFT) model including all covariates under the assumption of Log-normal survival times. Write the fitted AFT model and interpret the effect of covariates.

$$\log Y_x = (0.369 + 1.843 * Type1 - 1.481 * Type2 + 0.443 * Type3 + 0.054 * Score + 0.019 * wait) + 1.474 * W$$

Where $W \sim N(0, 1)$

The interpretation is like that of an OLS regression with log(Y) as the response variable. So for a given $\hat{\beta}$, the effect on survival time can be found by exponentiating.

For example, $\hat{\beta}_{Score} = 0.054$, so a person with score s+1 can expect their survival time to change by a factor of 1.05 over a person with score = s. In other words, they can expect a 5% increase in survival time.

```
/* log-normal */
proc lifereg data=leuk;
    class type;
    model time*status(0)=type score wait
        /dist=lnormal;
run;
```

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	0.3691	1.1447	-1.8746	2.6127	0.10	0.7472
type	1	1	1.8432	0.7579	0.3578	3.3286	5.91	0.0150
type	2	1	-1.4805	0.8770	-3.1994	0.2383	2.85	0.0914
type	3	1	0.4434	0.7049	-0.9383	1.8250	0.40	0.5294
type	4	0	0.0000	.	.	.	.	.
score		1	0.0535	0.0135	0.0269	0.0800	15.59	<.0001
wait		1	0.0192	0.0113	-0.0030	0.0415	2.88	0.0895
Scale		1	1.4742	0.2198	1.1007	1.9746		

Figure 1: SAS Output: Log-normal

Question c

Fit the (AFT) model including all covariates under the assumption of Log-logistic survival times. Write the fitted AFT model and interpret the effect of covariates.

$$\log Y_x = (0.244 + 1.976 * Type1 - 1.480 * Type2 + 0.474 * Type3 + 0.054 * Score + 0.021 * wait) + 0.812 * W$$

Where $W \sim Logistic(0,\ 1)$

Here the effect of the covariate on the probabitlty of experiencing the event is found as an odds ratio by exponentiating $\frac{-\beta}{\sigma}$. So, for instance, given that $\hat{\beta}_{Type1} = 1.976$ and $\hat{\sigma} = 0.812$; assuming that all other covariates are held constant, the effect of Being Type 1 (Allo, NHL) changes your odds of experiencing the event by a factor of $e^{-\frac{1.976}{0.812}} = 0.088$ over the baseline, Type 4 (Auto, HOD). That's an overall decrease of around 91.2% in a person's odds of experiencing the event (relapse or death).

Similarly, for the "wait" variable, which is waiting time in months, we can exponentiate to find that the odds ratio is changed by a factor of 0.975. In other words, assuming that all else is held constant, a person who waits w+1 months will have their odds of experiencing the event reduced by around 2.5% over a person who waits w months.

```
/* log-logistic */
proc lifereg data=leuk;
    class type;
    model time*status(0)=type score wait
        /dist=llogistic;
run;
```

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	0.2443	1.0214	-1.7577	2.2462	0.06	0.8110
type	1	1	1.9763	0.7323	0.5410	3.4117	7.28	0.0070
type	2	1	-1.4802	0.7758	-3.0008	0.0404	3.64	0.0564
type	3	1	0.4735	0.6636	-0.8272	1.7742	0.51	0.4755
type	4	0	0.0000	.	.	.	.	.
score		1	0.0543	0.0124	0.0299	0.0786	19.11	<.0001
wait		1	0.0207	0.0109	-0.0007	0.0420	3.61	0.0576
Scale		1	0.8116	0.1356	0.5849	1.1261		

Figure 2: SAS Output: Log Logistic

Question d

Fit the (AFT) model including all covariates under the assumption of Weibull survival times. Write the fitted AFT model and interpret the effect of covariates.

$$\log Y_x = (0.803 + 1.750 * Type1 - 2.049 * Type2 + 0.200 * Type3 + 0.057 * Score + 0.019 * wait) + 1.064 * W$$

Where $W \sim extreme\,value$

With a Weibull distribution, the Relative Risk is found as a ratio of hazard rates by exponentiating $\frac{-\beta}{\sigma}$. So, assuming all else is held constant, given that $\hat{\beta}_{Type1} = 1.750$ and $\hat{\sigma} = 1.064$; the effect of Being Type 1 (Allo, NHL) changes your hazard rate by a factor of $e^{-\frac{1.750}{1.064}} = 0.19$ compared to the baseline, Type 4. An overall decrease of around 81%. This means that a person's *risk* is lower if Type 1 than if Type 4.

Similarly, exponentiating the "wait" variable we find that the odds ratio is changed by a factor of 0.98. In other words, assuming that all else is held constant, a person who waits $w + 1$ months will have their risk reduced by around 2% relative to the person whose wait time is w.

```
/* Weibull */
proc lifereg data=leuk;
    class type;
    model time*status(0)=type score wait
        /dist=weibull;
run;
```

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	0.8028	0.9314	-1.0228	2.6283	0.74	0.3888
type	1	1	1.7503	0.6678	0.4415	3.0590	6.87	0.0088
type	2	1	-2.0491	0.7462	-3.5117	-0.5865	7.54	0.0060
type	3	1	0.2003	0.6004	-0.9764	1.3770	0.11	0.7387
type	4	0	0.0000	.	.	.	.	.
score		1	0.0572	0.0112	0.0353	0.0792	26.14	<.0001
wait		1	0.0190	0.0097	0.0000	0.0379	3.86	0.0494
Scale		1	1.0644	0.1722	0.7752	1.4615		
Weibull Shape		1	0.9395	0.1520	0.6842	1.2900		

Figure 3: SAS Output: Weibull

Question e

Which model is the best initial model among the three models in parts (b), (c) and (d)? Justify your answer. From the best initial model, conduct a model selection (at 5% level) and identify the final model.

The best model based on AIC is Weibull (AIC=126.405). Log-normal (AIC=129.396) and log-logistic (AIC=128.114) both have greater AIC's. However, I will note that the Generalized gamma model had a lower AIC than all other models (AIC=117.392), though I will not use the generalized gamma in my analyses as it serves only as a benchmark here.

(All AIC output is included in the appendix.)

Using the Weibull distribution for my final model I find that all p-values are significant at the $\alpha = 0.05$ level. So my final model is the same as the full Weibull model and it is given below.

i

Write the fitted final model.

$$\log Y_x = (0.803 + 1.750 * Type1 - 2.049 * Type2 + 0.200 * Type3 + 0.057 * Score + 0.019 * wait) + 1.064 * W$$

Where $W \sim extreme\, value$

ii

Does the effect of GRAFT differ for NHL and HOD patients?

Using Type 4 as the baseline we can see that an Allogenic transplant raises a Hodgkin's patient's relative risk by almost 700% over a Hodgkin's patient who received an Autologous transplant instead.

We can also see that a non_Hodgkin's patient fares better than the baseline regardless of the type of graft, but the effect of graft is opposite in in the NHL patients. An Auto graft reduces relative risk by around 18% compared to baseline while an Allo graft reduces the risk by 81%.

So while an Allogenic graft is quite bad for a Hodgkin's patient, it is quite good for a NHL patient and the opposite is true for Autologous grafts.

Type	Graft	Disease	Effect on RR
Type 1	Allo	NHL	0.19
Type 2	Allo	HOD	6.86
Type 3	Auto	NHL	0.82
Type 4	Auto	HOD	Baseline

Question f

Check the goodness of fit of the reduced model in part (e) by Probability plot and Cox-Snell residual plot. Does the reduced model fit the data adequately? If not, what methods/models would you use to analyze the data? Why?

From the plots on the following pages we can see that the goodness of fit is quite bad for this model. In the probability plot we would expect all of the points to fall within the confidence band and a few do fall outside of it.

In the Cox-Snell residual plot, if the model was a good fit, we would expect to see a straight line with a slope of 1. Instead, as can be seen below, the plot is far from linear and the overall trend is much greater than 1:1.

I will check the fit further by performing a likelihood ratio given the following null hypothesis:

H_0 : My final model above is good model for my data.

The likelihood ratio test statistic is given as:

$$\Lambda = (-2logL(H_0)) - (-2logL(H_1)) \sim \chi^2_{(df=Dropped\,parameters=1)}$$

And from the SAS output above we can see that:

$$\Lambda = 101.3 - 112.4 = -11.1$$

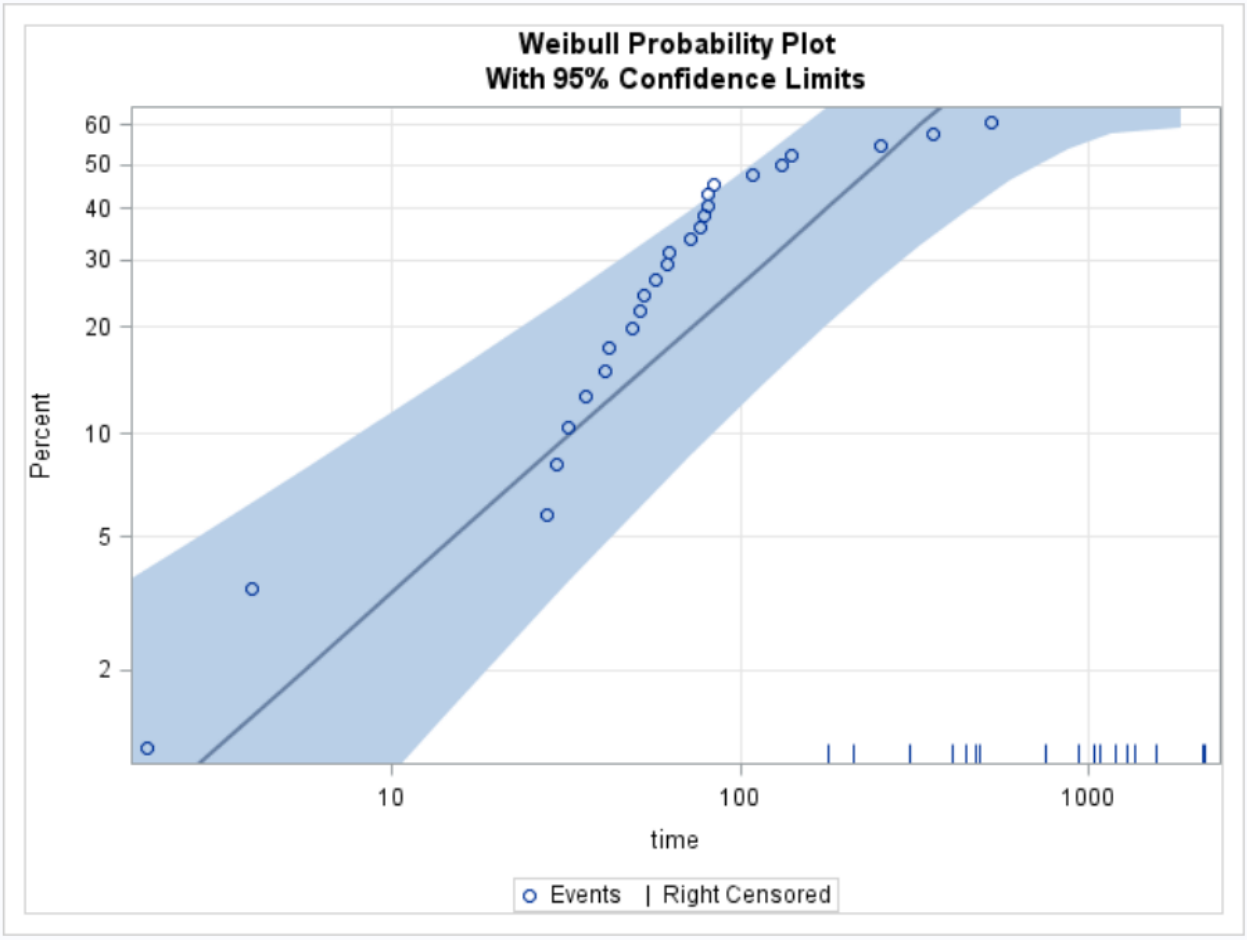
Using R to compute p:

```
paste("p =", pchisq(-11.1,1))
```

```
## [1] "p = 0"
```

Given that p is effectively zero, we can safely reject the null hypothesis and conclude that our model is insufficient to describe our data.

Having concluded that AFT is a bad model for these data I would continue my analysis by attempting to fit a Cox model and if that also failed then I would try non-parametric methods to assess survivorship.


Weibull Probability Plot
With 95% Confidence Limits
Percent
60
50
40
30
20
10
5
2
10
100
1000
time
Events | Right Censored

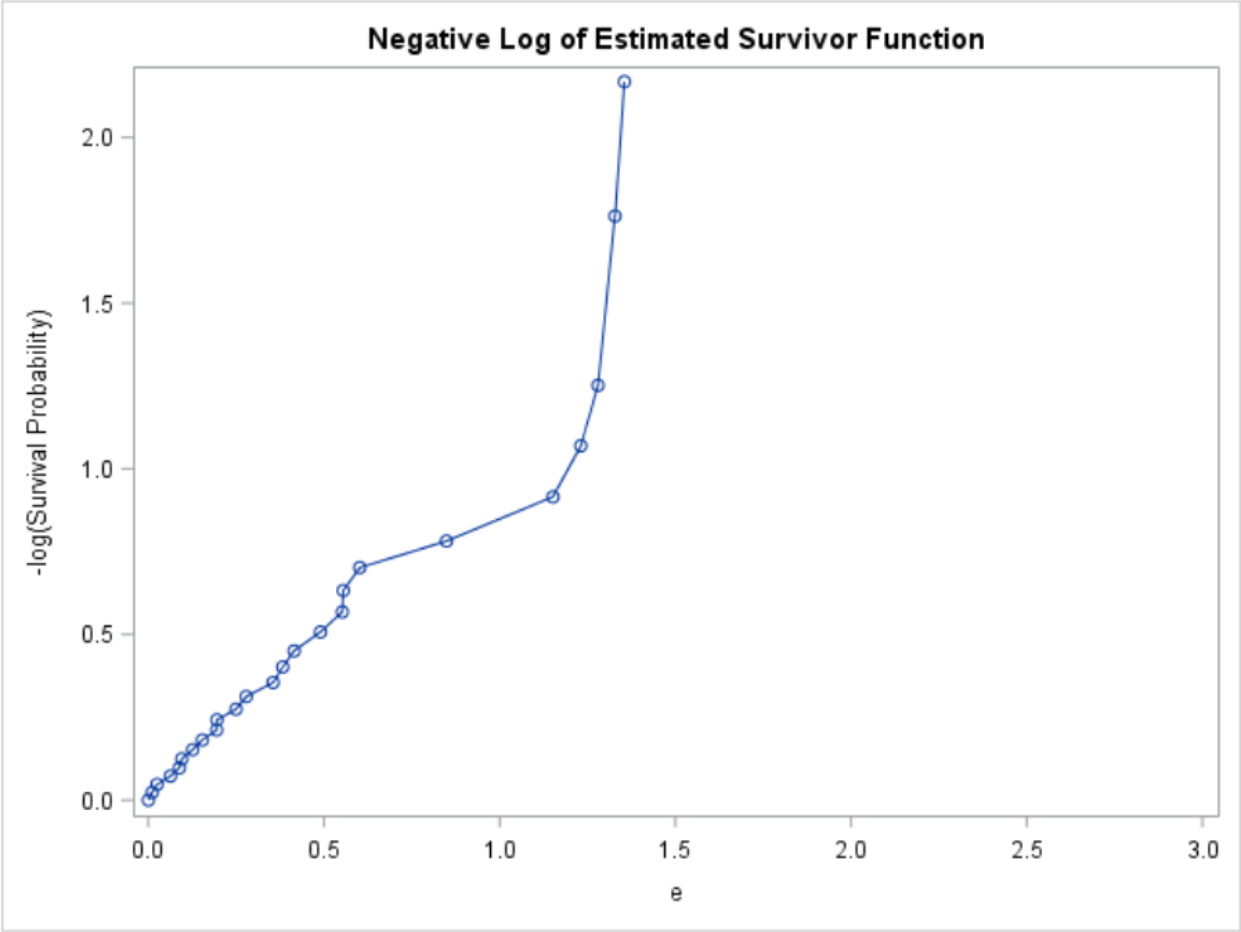


Figure 4: SAS Output: Cox-Snell residuals plot

Apendix

Fit Statistics	
-2 Log Likelihood	115.396
AIC (smaller is better)	129.396
AICC (smaller is better)	132.596
BIC (smaller is better)	141.724

Figure 5: SAS Output: Fit statistics for log-normal

Fit Statistics	
-2 Log Likelihood	114.114
AIC (smaller is better)	128.114
AICC (smaller is better)	131.314
BIC (smaller is better)	140.443

Figure 6: SAS Output: Fit statistics for log-logistic

Fit Statistics	
-2 Log Likelihood	112.405
AIC (smaller is better)	126.405
AICC (smaller is better)	129.605
BIC (smaller is better)	138.733

Figure 7: SAS Output: Fit statistics for Weibull