

Q1:

The ending event is death of the patient, or end of the study.

The time origin is the moment they are admitted to hospital, which is when data started being collection.

Baseline cumulative hazard function for different age groups

Obs	id	age	gender	hr	sysbp	diasbp	bmi	cvd	afb	sho	chf	av3	miord	mitype	year	admitdate	disdate	fdate	los	dstat	lenfol	fstat
1	1	83	0	89	152	78	25.54051	1	1	0	0	0	1	0	1	01/13/1997	01/18/1997	12/31/2002	5	0	2178	0
2	2	49	0	84	120	60	24.02398	1	0	0	0	0	0	1	1	01/19/1997	01/24/1997	12/31/2002	5	0	2172	0
3	3	70	1	83	147	88	22.1429	0	0	0	0	0	0	1	1	01/01/1997	01/06/1997	12/31/2002	5	0	2190	0
4	4	70	0	65	123	76	26.63187	1	0	0	1	0	0	1	1	02/17/1997	02/27/1997	12/11/1997	10	0	297	1
5	5	70	0	63	135	85	24.41255	1	0	0	0	0	0	1	1	03/01/1997	03/07/1997	12/31/2002	6	0	2131	0

```
PROC IMPORT OUT= WORK.whas500
    DATAFILE= "C:\Users\cianb\Downloads\whas500.csv"
    DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

ods graphics on;

proc print data=whas500 (obs=5);
run;
```

Q2:

The SAS System

Obs	lenfol	_CENSOR_	SURVIVAL	SDF_LCL	SDF_UCL
1	0	.	1.000	1.00000	1.00000
2	1	0	0.984	0.96826	0.99197
3	2	0	0.968	0.94830	0.98027
4	3	0	0.962	0.94107	0.97559
5	4	0	0.958	0.93631	0.97241

Using the Univariate Chi-Squares for the Wilcoxon Test, seven variables were found to be significant ($p < 0.05$) or marginally significant ($p < 0.10$) factors or covariates: *age hr diasbp bmi sho chf los*. However, using the Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test, we found that the significant factors were *age hr diasbp bmi sho chf gender*, with *los* failing to make the cutoff.

The Forward Stepwise Sequence and Univariate Chi-Squares for log-rank tests mostly agreed with the above conclusions, except for *gender* being found significant and *los* being found non-significant in both tests.

Using the proportional hazards regression procedure, once with stepwise selection and once with backward selection, both models ended with the same variables: *age hr diasbp bmi sho chf gender*.

We included the variable *gender* in final model due its importance in medical issues, while the variable *los* was left out due to being found significant in only one of four tests as well as being dropped during the regression.

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
<i>age</i>	1	123.4	<.0001	123.4	<.0001
<i>chf</i>	2	160.4	<.0001	37.0212	<.0001
<i>sho</i>	3	177.8	<.0001	17.4349	<.0001
<i>hr</i>	4	187.1	<.0001	9.2229	0.0024
<i>diasbp</i>	5	199.2	<.0001	12.1028	0.0005
<i>bmi</i>	6	204.3	<.0001	5.1080	0.0238
<i>los</i>	7	208.8	<.0001	4.5049	0.0338
<i>gender</i>	8	210.5	<.0001	1.6917	0.1934
<i>mitype</i>	9	211.5	<.0001	1.0695	0.3011
<i>av3</i>	10	212.0	<.0001	0.5014	0.4789
<i>miord</i>	11	212.1	<.0001	0.0588	0.8084
<i>cvd</i>	12	212.1	<.0001	0.0457	0.8308
<i>sysbp</i>	13	212.2	<.0001	0.0530	0.8178
<i>afb</i>	14	212.2	<.0001	0.0275	0.8683

Univariate Chi-Squares for the Wilcoxon Test				
Variable	Test Statistic	Standard Error	Chi-Square	Pr > Chi-Square
<i>age</i>	-1799.3	162.0	123.4	<.0001
<i>gender</i>	-14.5401	5.6628	6.5929	0.0102
<i>hr</i>	-1418.8	270.7	27.4641	<.0001
<i>sysbp</i>	895.3	373.4	5.7493	0.0165
<i>diasbp</i>	1236.3	247.7	24.9105	<.0001
<i>bmi</i>	413.4	61.8269	44.7077	<.0001
<i>cvd</i>	-7.8213	4.9234	2.5236	0.1122
<i>afb</i>	-12.6357	4.2554	8.8169	0.0030
<i>sho</i>	-10.6383	1.9524	29.6910	<.0001
<i>chf</i>	-45.2565	5.3254	72.2196	<.0001
<i>av3</i>	-2.0814	1.6084	1.6747	0.1956
<i>miord</i>	-14.8515	5.5193	7.2406	0.0071
<i>mitype</i>	19.9591	5.1897	14.7911	0.0001
<i>los</i>	-44.4288	55.2089	0.6476	0.4210

Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test						Univariate Chi-Squares for the Log-Rank Test				
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment	Variable	Test Statistic	Standard Error	Chi-Square	Pr > Chi-Square
age	1	126.3	<.0001	126.3	<.0001	age	-2318.6	206.3	126.3	<.0001
chf	2	173.9	<.0001	47.6785	<.0001	gender	-19.7264	7.0759	7.7720	0.0053
sho	3	186.7	<.0001	12.7627	0.0004	hr	-1848.0	330.4	31.2839	<.0001
diasbp	4	197.1	<.0001	10.4088	0.0013	sysbp	932.0	460.7	4.0922	0.0431
hr	5	210.6	<.0001	13.5038	0.0002	diasbp	1489.7	308.1	23.3804	<.0001
bmi	6	217.3	<.0001	6.7083	0.0096	bmi	512.1	77.0522	44.1654	<.0001
los	7	222.3	<.0001	4.9500	0.0261	cvd	-10.8399	6.4157	2.8547	0.0911
gender	8	226.4	<.0001	4.1767	0.0410	afb	-16.7749	5.0875	10.8722	0.0010
mitype	9	228.8	<.0001	2.3927	0.1219	sho	-11.3335	2.3251	23.7606	<.0001
sysbp	10	229.9	<.0001	1.0626	0.3026	chf	-57.3835	6.2473	84.3701	<.0001
cvd	11	230.5	<.0001	0.5964	0.4400	av3	-2.5435	2.0631	1.5200	0.2176
av3	12	230.5	<.0001	0.0459	0.8304	miord	-21.0555	6.8146	9.5467	0.0020
miord	13	230.6	<.0001	0.0323	0.8574	mitype	27.7870	6.9159	16.1432	<.0001
afb	14	230.6	<.0001	0.000037	0.9952	los	-79.1855	70.9393	1.2460	0.2643

```

proc lifetest data=whas500 method=km plots=(survival(c1),ls,lls)
  graphics outsurv=a;
  time lenfol*fstat(0);
  test age gender hr sysbp diasbp bmi cvd afb sho chf av3 miord mitype
  los;
run;

proc print data=a (obs=5);
run;

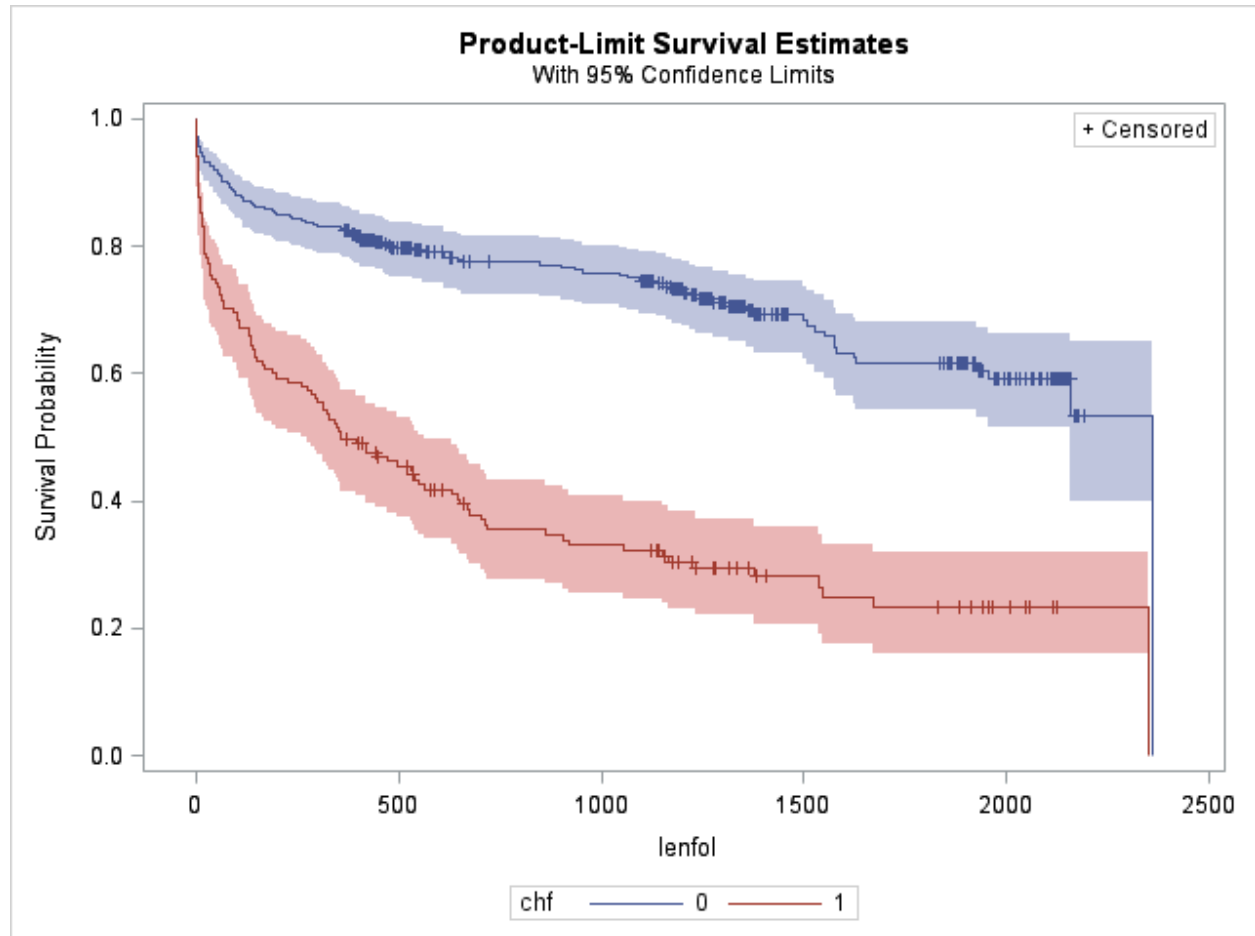
proc phreg data=whas500;
class gender;
  model lenfol*fstat(0)= age gender hr sysbp diasbp bmi cvd afb sho chf
  av3 miord mitype los/ties=exact selection=backwards;
run;

proc phreg data=whas500;
class gender;
  model lenfol*fstat(0)= age gender hr sysbp diasbp bmi cvd afb sho chf
  av3 miord mitype los/ties=exact selection=stepwise;
run;

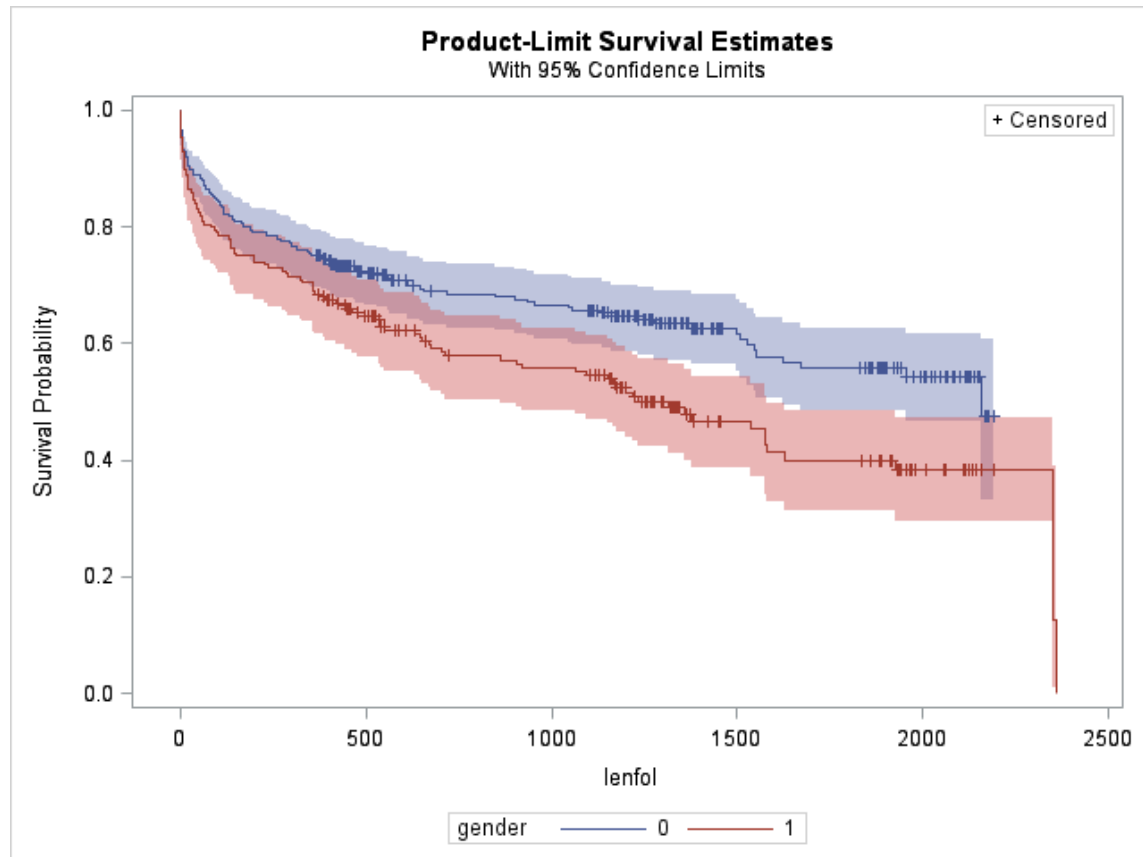
```

Q3:

The graph below of the survival function of patients stratified by *chf* (Congestive Heart Complications) shows that the survival probabilities of a patient with congestive heart complications is much worse than a patient who does not have congestive heart complications.



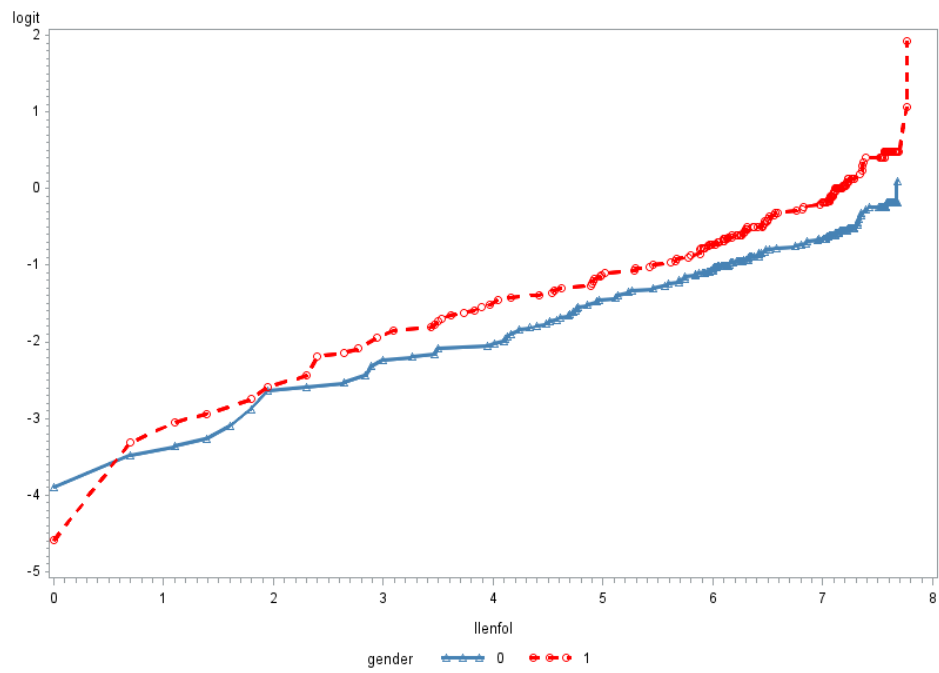
The graph below shows the kaplan-meier survival estimates stratified by *gender*, showing that women tend to have worse survival probabilities compared to men, with the probabilities getting worse as time goes on.



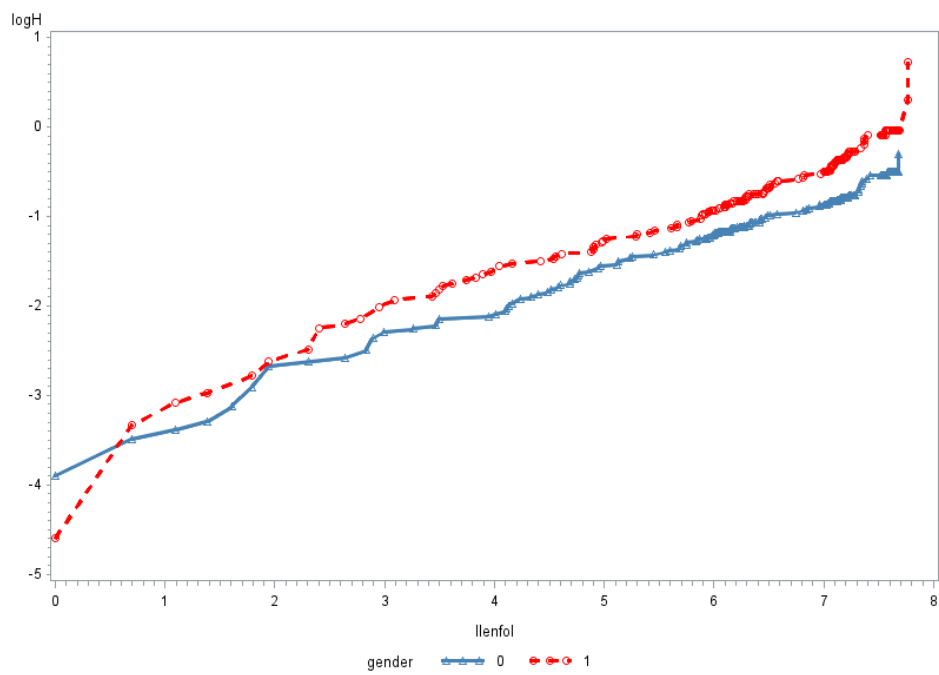
Q4:

The graphs below are the graphical checking of the log-normal, log-logistic, and Weibull survival times. All three, while they show some linearity, are not linear, leading us to conclude that the distributions do not work.

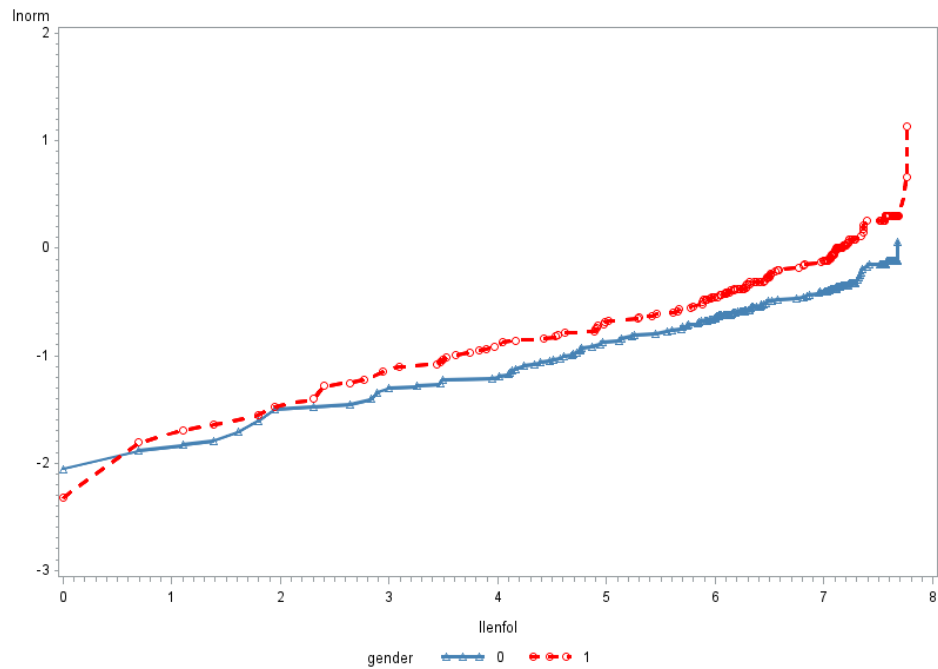
Graphical Checking of Distributions



Graphical Checking of Distributions



Graphical Checking of Distributions



Full SAS Code:

```
PROC IMPORT OUT= WORK.whas500
    DATAFILE= "C:\Users\cianb\Downloads\whas500.csv"
    DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

ods graphics on;

proc print data=whas500 (obs=5);
run;

proc lifetest data=whas500 method=km plots=(survival(c1),ls,lls)
    graphics outsurv=a;
    time llenfol*fstat(0);
    test age gender hr sysbp diasbp bmi cvd afb sho chf av3 miord mitype
    los;
run;

proc print data=a (obs=5);
run;
```

```

proc phreg data=whas500;
class gender;
    model lenfol*fstat(0)= age gender hr sysbp diasbp bmi cvd afb sho chf
av3 miord mitype los/ties=exact selection=backwards;
run;

proc phreg data=whas500;
class gender;
    model lenfol*fstat(0)= age gender hr sysbp diasbp bmi cvd afb sho chf
av3 miord mitype los/ties=exact selection=stepwise;
run;

proc phreg data=whas500;
class gender;
    model lenfol*fstat(0)= age gender hr diasbp bmi sho chf/ties=exact;
run;

/* Explain selection of factors/covariates and graph */

proc lifetest data=whas500 method=km nelson plots=(survival(cl),ls,lls)
    graphics;
    time lenfol*fstat(0);
    strata chf;
run;

proc lifetest data=whas500 method=km nelson plots=(survival(cl),ls,lls)
    graphics;
    time lenfol*fstat(0);
    strata gender;
run;

data distY;
    set a;
    s=survival;
    logH=log(-log(s));
    lnorm=probit(1-s);
    logit=log((1-s)/s);
    llenfol=log(lenfol);
run;

proc gplot data=distY;
    plot logit*llenfol logH*lenfol lnorm*llenfol;

```



```

        run;
quit;

proc lifetest data=whas500 method=km plots=(survival(c1),ls,lls)
    graphics outsurv=c;
    time lenfol*fstat(0);
    strata gender;
run;

proc lifetest data=whas500 method=km plots=(survival(c1),ls,lls)
    graphics outsurv=c;
    time lenfol*fstat(0);
    strata gender;
    test age gender hr sysbp diasbp bmi cvd afb sho chf av3 miord mitype
los;
run;

data distZ;
    set c;
    s=survival;
    logH=log(-log(s));
    lnorm=probit(1-s);
    logit=log((1-s)/s);
    llenfol=log(lenfol);
run;

proc gplot data=distZ;
    title "Graphical Checking of Distributions";
    symbol1 i=join width=2 value=triangle c=steelblue;
    symbol2 i=join width=2 value=circle c=red;
    plot logit*llenfol=gender logH*llenfol=gender lnorm*llenfol=gender;
run;
quit;

ods graphics off;

```