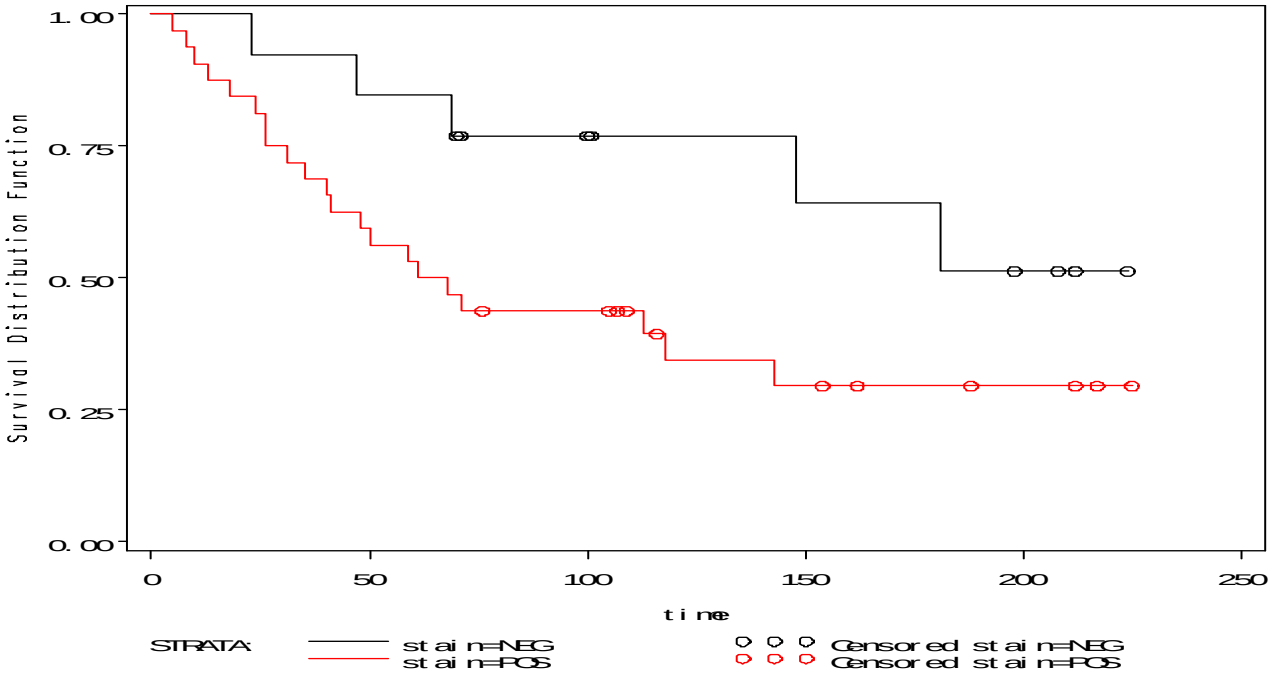
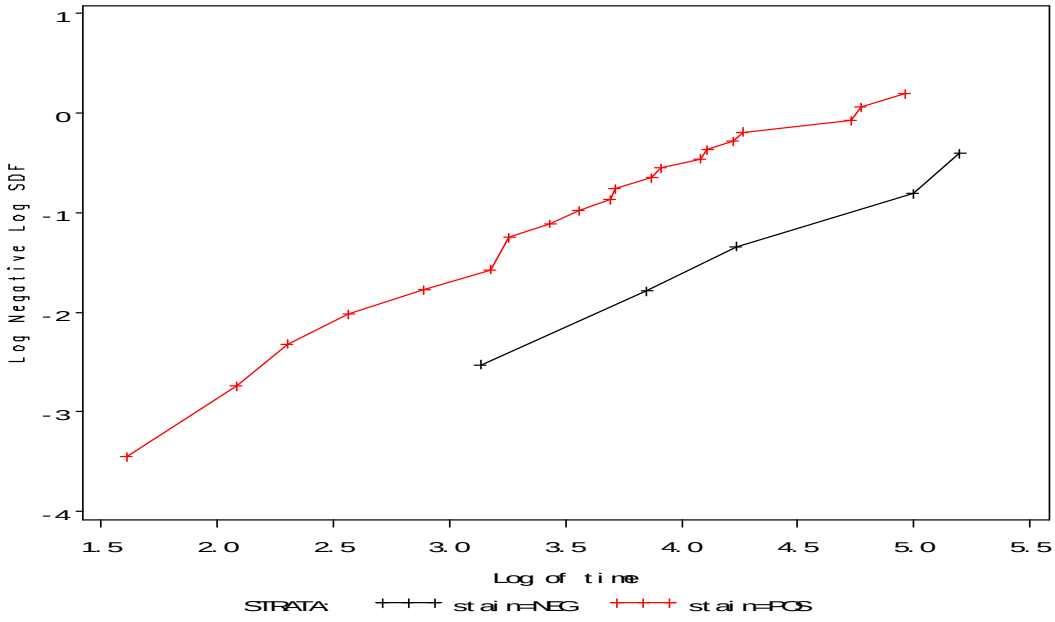
Class Notes for Survival Analysis (revised 3/16)

- See text Section 3.5 and Chapter 7 (pp. 54-62 & 211-252); ***for the quiz on Thursday 3/18, students should be familiar with all this material & know how to construct Kaplan Meier estimates.***
- **Survival data** (i.e., 'time until' data) is typically **skewed** and **censored**. Censored data means right (most common), left, and/or interval. We'll focus on right-censored data analysis; the text addresses interval- and left-censored data in §7.5.
- **Notation** – needed to get at an understanding of the survival and hazard functions. First, we discuss **PDF: f(t)** and **CDF: F(t)** (see in-class notes). Next, the **survival function: S(t)** gives the probability of "survival" beyond some time point. The **hazard function h(t)** gives the hazard of 'dying' in the next instance given survival up to time t – note that it is therefore 'conditional', and is not a probability (it can exceed one) – it is instead a rate.
- The **Exponential distribution** is considered and illustrated first since it is useful in applications (especially reliability and engineering); this distribution is generalized via the **Weibull distribution**. But these ***parametric*** approaches are criticized (see p.216 section 7.2.2) as being too demanding in requirements/ assumptions, and these points lead us to seek out ***nonparametric*** or ***semi-parametric*** methods (no distributional assumptions).
- **Leukemia example** on p.55 illustrates **Kaplan-Meier** method (to estimate the two survival curves; but note correct on HO#1 p.2) and the **log-rank test** to test for coincidence of the two curves.
- Another example: Collett (2003:7) presents data related to the prognosis for women with **breast cancer**. Here, the response variable is survival time (months) of women with tumors that were negatively or positively stained with a snail lectin (marker).

```
data one;
  input st time status @@; pt=_n_; posstain=st-1;
  stain='POS'; if st=1 then stain='NEG'; datalines;
1  23 1 1  47 1 1  69 1 1  70 0 1  71 0 1 100 0 1 101 0 1 148 1
```

```
1 181 1 1 198 0 1 208 0 1 212 0 1 224 0 2   5 1 2   8 1 2  10 1
2  13 1 2  18 1 2  24 1 2  26 1 2  26 1 2  31 1 2  35 1 2  40 1
2  41 1 2  48 1 2  50 1 2  59 1 2  61 1 2  68 1 2  71 1 2  76 0
2 105 0 2 107 0 2 109 0 2 113 1 2 116 0 2 118 1 2 143 1 2 154 0
2 162 0 2 188 0 2 212 0 2 217 0 2 225 0
;
proc lifetest plots=(s,lls);
  time time*status(0);
  strata stain;
run;
```





Parallelism above indicates CPH model may fit these data (p.239).

A. LIFETEST Output

Stratum 1: stain = NEG
Summary Statistics for Time Variable time

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	.	181.000	.
50	.	148.000	.
25	148.000	47.000	.

B. LIFETEST Output

Stratum 2: stain = POS
Summary Statistics for Time Variable time

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	.	113.000	.
50	64.500	40.000	143.000
25	28.500	18.000	50.000

C. LIFETEST Output

Testing Homogeneity of Survival Curves for time over Strata
Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	3.5150	1	0.0608
Wilcoxon	4.1800	1	0.0409
-2Log(LR)	4.3563	1	0.0369

- **Cox's Proportional Hazards (CPH) model** is a semi-parametric method/model used to relate hazard functions which vary with **x** to a baseline hazard (h_0) – the covariate vector (**x**) enters as in Equations (7.5) and (7.6) on p.216 in the text; these are equivalent manners to write the CPH model. When there are two treatments (such as drugs A and B), x is just a dummy variable.

Note that there is no intercept in this model (it is absorbed into the h_0 term). The CPH model is fit to the **leukemia** data in the Addendum (attached at the end of HO#1) – **note the interpretation of the hazard ratio!** The proportionality assumption/ requirement is 'confirmed' by looking at the above LLS plot for parallelism (see p.239). See additional examples of the CPH model fit in the text in §7.2 noting especially the interpretations of the hazard ratios. When the CPH model is fit to Collett's **breast cancer** data, we get the following:

```
proc phreg;
  model time*status(0)=posstain;
run;
```

PHREG Partial Output

Model Information

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
45	26	19	42.22

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	173.968	170.096
AIC	173.968	172.096
SBC	173.968	173.354

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.8717	1	0.0491
Score	3.5081	1	0.0611
Wald	3.2858	1	0.0699

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
posstain	1	0.90801	0.50092	3.2858	0.0699	2.479

- Another model for survival data is the **Accelerated Failure Time (AFT) model** (not discussed in our text); it too brings

covariate(s) into the model. It appears that the CPH model is in widespread use in medicine and the AFT model in engineering/reliability applications, although there is evidence this is changing (see the DIJ article emailed to you).

- Some common SAS procedures for Survival Analysis are:
  (1) SAS **LIFETEST** procedure provides nonparametric (Kaplan Meier) estimates of quartiles and medians for the two groups, and tests of this question (null = no difference between the two curves). Often we use the **log-rank test**.
  (2) SAS **PHREG** procedure fits the CPH model; this model is the theme of Chapter 7 in the text.
  (3) SAS **LIFEREG** (with **Weibull** distribution) fits the ***parametric* Cox PH model**.
  (4) SAS' **LIFEREG** (with **Log-logistic** distribution) fits the **parametric AFT model**.
- For the above **breast cancer** example with two 'treatments' (Stain = NEG or POS), whereas the fitted CPH model gives $\hat{S}_{POS}(t) = \left[\hat{S}_{NEG}(t)\right]^{2.479}$, the fitted AFT model with log-logistic distribution gives $\hat{S}_{POS}(t) = \hat{S}_{NEG}(3.16t)$. This follows since the AFT model holds that $\mathbf{S_{POS}(t) = S_{NEG}(t/e^{\beta})}$ or equivalently that $\mathbf{h_{POS}(t) = e^{-\beta}\, h_{NEG}(t/e^{\beta})}$. The relevant output is the following:

```
proc lifereg;
  model time*status(0)=posstain/d=llogistic;
run;
```

LIFEREG Partial Output

Model Information

Name of Distribution	LLogistic
Log Likelihood	-59.24726035

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.4611	0.4604	4.5588	6.3634	140.72	<.0001
posstain	1	-1.1491	0.5202	-2.1687	-0.1294	4.88	0.0272
Scale	1	0.8047	0.1330	0.5821	1.1125		

- Today's handout (HO#3) emphasizes the interpretations of hazard ratios in the face of **interaction**.
- In HO#2, we examine the **Maryland recidivism** data. The **logistic analysis** at the end is less powerful than the **survival analysis** since the logistic analysis only measure whether or not the releasee was subsequently re-arrested, whereas the survival analysis measured this (censored or not) plus time until re-arrest. The interpretation associated with the logistic analysis would involve "**odds of re-arrest**" whereas that associated with the survival analysis involves "**hazard of re-arrest**" or "**survival**" (probability of no re-arrest up to a point in time). In terms of the survival analysis, interestingly, both the **AFT** (log-logistic) and the **CPH** models produce similar results, and both indicate that the '**fin*age**' interaction may be significant; this will complicate interpretation. It would be wise to write out one's interpretation for the CPH model. Also, note that the interpretation for the AFT model involves '**expected time to re-arrest**' whereas that for the CPH model involves '**hazard of re-arrest**'.
- In fact, the above recidivism data is actually a little more complicated as employment status is actually measure each week for each releasee; this means that 'employment status' is a '**time dependent covariate**', and this is discussed in §7.3.1 in the text. Handling TDC's in STATA is challenging but in SAS/PHREG it is not (p.234).
- One way around non-proportionality is to stratify (see p.235), although this is more so for 'blocking variables'.