Lab 2, Significance Testing: Solutions

Dave Harrington

May 2018

The lab files come in two versions: a PDF with the problem statements and an Rmd file that produces the PDF. In most cases, you can work in the Rmd file and enter in your solutions. For the purely algebraic questions, you may either use LaTeX commands to enter your solutions in the Rmd file or write out the answers on paper.

The solution files (a PDF with the solutions, and the Rmd file to produce them) are contained in a separate folder under each lab. Your learning experience with the labs will be more effective if you do not look at the solutions until after you finish working on the questions.

Problem 1: Tests for differences at a single time point

This problem uses the Cox and Oakes leukemia data mentioned in the slides for Units 2 and 3, which is in the `eventtimedata` package as `cox.oakes.leukemia`. You may use the code shown in the slides as needed.

a) In the Cox and Oakes leukemia data, find a 95% confidence interval for the difference in survival curves at time point $t^\star = 10$ weeks.

b) Does the confidence interval imply that the values of the survival functions at $t^\star = 10$ are significantly different at significance level $\alpha = 0.05$?

Problem 1 Solution.

a) A 95% confidence interval for the difference of two survival curves at a point $t^\star$ is given by

$$\left[\left(\widehat{S}_1(t^\star) - \widehat{S}_0(t^\star)\right) \pm 1.96 \times \sqrt{V_1(t^\star) + V_0(t^\star)}\right],$$

where $V_k(t^\star)$ is the estimated variance of $\widehat{S}_k(t^\star)$.

At the point $t^\star = 10$ for the control group, the value of the KM estimator and its standard error have the same value as at $t = 8$, since there are no events in that group until $t = 11$.

The 95% confidence interval is (-0.65, -0.09).

b) This confidence interval does suggest that the curves are significantly different at this point, since the interval does not contain 0.

This approach to testing for a difference between survival curves is far from optimal, but can be useful in some instances, especially in settings where, say, survival at 10 weeks is meaningful. There are several ways in which the method falls short. Most importantly, it ignores the pattern of failures throughout the curves.

```r
library(survival)
library(eventtimedata)

leukemia.group.0 = subset.data.frame(cox.oakes.leukemia, group == 0)
km.group.0 = survfit(Surv(time, relapse) ~ 1, data = leukemia.group.0)

leukemia.group.1 = subset.data.frame(cox.oakes.leukemia, group == 1)
km.group.1 = survfit(Surv(time, relapse) ~ 1, data = leukemia.group.1)

#extract values from summary and calculate CI
est.surv.0 = summary(km.group.0, time = 10)$surv
est.surv.1 = summary(km.group.1, time = 10)$surv
std.err.surv.0 = summary(km.group.0, time = 10)$std.err
std.err.surv.1 = summary(km.group.1, time = 10)$std.err

diff = est.surv.0 - est.surv.1
std.err.diff = sqrt(std.err.surv.0^2 + std.err.surv.1^2)
m = qnorm(0.975)*std.err.diff
diff - m; diff + m
```

```
## [1] -0.6527029
```

```
## [1] -0.09127471
```

Problem 2: The numerator of the log-rank statistic

Show that the $(o-e)$ terms in the numerator of the log-rank statistic can be written as

$$\frac{r_{0j}r_{1j}}{r_j}(\widehat{\lambda}_{1j} - \widehat{\lambda}_{0j}).$$

Problem 2 Solution.

The numerator of the log-rank statistic is

$$\left[\sum_{j=1}^{K}(d_{0j} - r_{0j} \times d_j/r_j)\right]^2.$$

Before squaring, the individual terms can be written as

$$\begin{aligned} d_{0j} - r_{0j}\frac{d_j}{r_j} &= d_{0j} - r_{0j}\frac{d_{0j}+d_{1j}}{r_j} \\ &= \frac{1}{r_j}[r_j d_{0j} - r_{0j}(d_{0j}+d_{1j})] \\ &= \frac{1}{r_j}[r_{1j}d_{0j} - r_{0j}d_{1j}] \\ &= \frac{r_{0j}r_{1j}}{r_j}\left[\frac{d_{0j}}{r_{0j}} - \frac{d_{1j}}{r_{1j}}\right] \\ &= \frac{r_{0j}r_{1j}}{r_j}\left[\widehat{\lambda}_{0j} - \widehat{\lambda}_{1j}\right]. \end{aligned}$$

This algebraic relationship shows that the log-rank statistic may be viewed as a squared sum of weighted differences of hazard functions at each of the failure times.

Problem 3: Prognosis in lymphoma

The Unit 3 lecture slides used the lymphoma prognosis data to illustrate a four-group log-rank test, testing for survival differences in lymphoma by stage. This example examines a simpler two-group test but with an interesting complication. The data are in the dataset `lymphoma.prognosis` in the `eventtimedata` package.

a) Estimate the survival probability by status of bulky disease. Bulky disease is coded in the numeric variable `BULK`, with `1` denoting `not present` and `2` denoting `present`. Be careful about the coding of the status variable `SURVIVAL`. See the slides or the code chunk below for how to use it appropriately.

   Do the data appear to satisfy the proportional hazards assumption?

b) Using a log-rank statistic, test for significant differences in survival in patients with bulky disease versus those who do not have bulky disease. State precisely the null and alternative hypotheses that are being tested.

c) The validity of the p-value from a log-rank test does not require that the data satisfy proportional hazards, but the test does lose power for some settings in which the hazards are not proportional. Comment on the effect of absence of proportional hazards on the outcome of the test.

d) In the Fleming-Harrington tests, setting the parameters $\rho = 1, \gamma = 0$ produces a generalized Wilcoxon test which emphasizes early differences. Re-do part b) with such a test. How does it change the outcome? Explain why.

Problem 3 Solution.

a) We will examine more detailed diagnostics later, but it is clear in the survival curves that the data do not satisfy proportional hazards—the survival curves cross at about 6 years. Since the patients without bulky disease have better survival over the first 6 years, the hazard function in this group must initially be smaller than the hazard for patients with bulky disease. After 6 years, the survival curve for the patients without bulky disease begins dropping more rapidly than that for the patients with bulky disease, so the hazard for death in the patients without bulky disease must be larger than that for patients with bulky disease. This is an example of crossing hazards.

b) The log-rank statistic tests the hypotheses $H_0 : S_1(t) = S_2(t)$ for all $t \geq 0$ versus the alternative $H_A : S_1(t) \neq S_2(t)$ for at least one $t \geq 0$. The null and alternative are equivalent to $H_0 : \lambda_1(t) = \lambda_2(t)$ for all $t \geq 0$ versus the alternative $H_A : \lambda_1(t) \neq \lambda_2(t)$ for at least one $t \geq 0$. The p-value for the log-rank test is 0.07, which is non-significant using a traditional significance level of $\alpha = 0.05$.

c) The alternative form of the log-rank statistic shown in Problem 2 helps explain what is going on in this dataset. When hazards cross, some of the terms in the weighted differences of hazard functions will be positive, while others will be negative, perhaps pushing the numerator towards 0 before it is squared and producing the the statistically nonsignificant result.

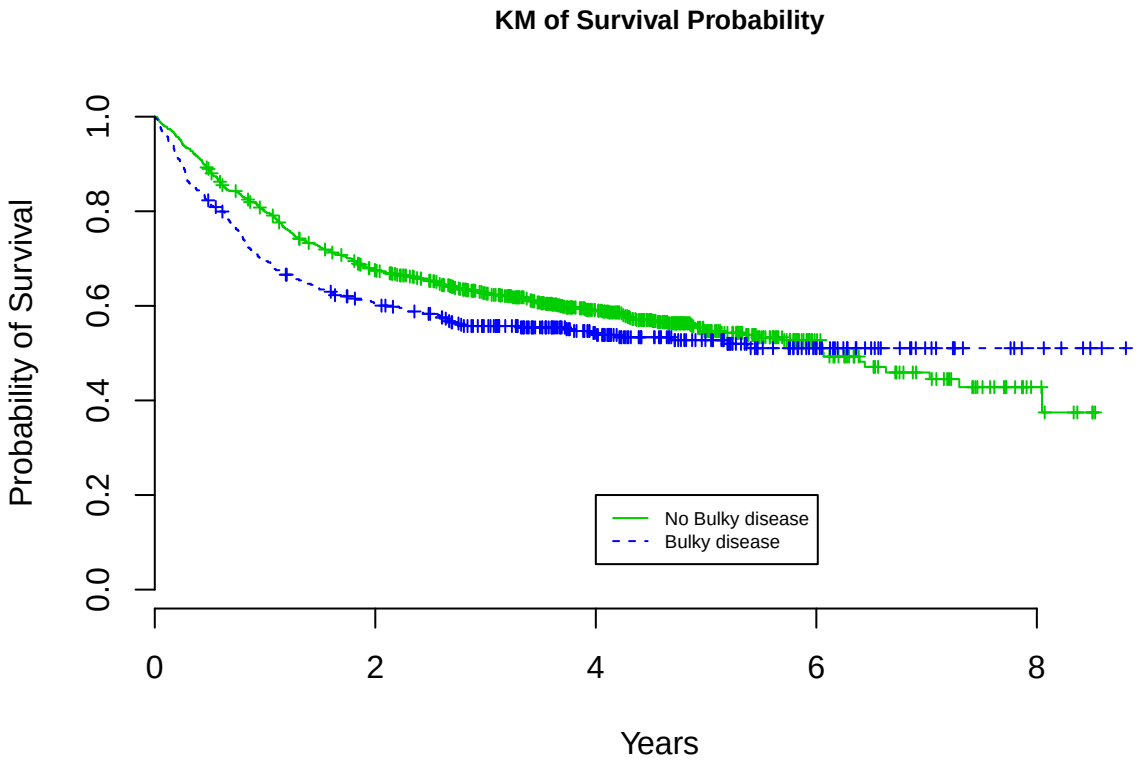
d) The generalized Wilcoxon test has p-value 0.01, which is significant at the 0.05 level. The generalized Wilcoxon test emphasizes differences in the survival curves when the curves are

near 1; that is, for smaller values of t. That feature is summarized by saying it 'emphasizes early differences'; those differences are evident in the survival plots.

Post-hoc reanalyses are always problematic, of course, especially so for a post-hoc change of a test statistic based on observed data. Essentially, the p-value loses its frequentist interpretation. In the analysis of a controlled clinical trial, the test statistic used to analyze the primary outcome should be the one that was specified in advance, except possibly in extraordinary circumstances.

It is also important to remember that in an observational dataset like this one, one cannot draw causal conclusions, and can only note an association that may be subject to confounding.

```
library(survival)
library(eventtimedata)
data("lymphoma.prognosis")

#adjusting coding of status variable
died = lymphoma.prognosis$SURVIVAL - 1
died[died == 2] = 0  #recoding those lost to follow-up as censored

#examine survival probability by status of bulky disease
lymphoma.km = survfit(Surv(SURVTIME, died) ~ BULK,
                      data = (lymphoma.prognosis))
plot(lymphoma.km, lty = 1:2, col = 3:4, mark.time = TRUE,
     xlab = "Years",
     ylab = "Probability of Survival",
     axes = FALSE,
     main = "KM of Survival Probability",
     cex = 0.6, cex.main = 0.8)
axis(1)
axis(2)
legend(4, .2, c("No Bulky disease", "Bulky disease"),
       lty = 1:2, col = 3:4, cex = 0.6)
```




```
#conduct log-rank test
survdiff(Surv(SURVTIME, died) ~ BULK, data = (lymphoma.prognosis))
```

```
## Call:
## survdiff(formula = Surv(SURVTIME, died) ~ BULK, data = (lymphoma.prognosis))
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## BULK=1 961      409      429     0.943      3.28
## BULK=2 424      194      174     2.327      3.28
##
##  Chisq= 3.3  on 1 degrees of freedom, p= 0.0703
```

```
#conduct generalized wilcoxon test
survdiff(Surv(SURVTIME, died) ~ BULK, rho = 1, data = (lymphoma.prognosis))
```

```
## Call:
## survdiff(formula = Surv(SURVTIME, died) ~ BULK, data = (lymphoma.prognosis),
##     rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## BULK=1 961      312      334      1.52      6.55
## BULK=2 424      158      136      3.73      6.55
##
##  Chisq= 6.5  on 1 degrees of freedom, p= 0.0105
```

Problem 4: Prognosis in lymphoma, continued

It is common in cancer studies to stratify treatment assignment by age, since older patients often have a much different prognosis than younger ones. Stratification in an analysis can also be useful in studying prognostic factors.

a) In the lymphoma data, assess the association between bulk disease and survival again, but this time by using a stratified log-rank test, stratifying on the binary variable `AGE60`. Is the result the same as in the earlier, unstratified log-rank test?

b) Try to explain the differences between the stratified and unstratified tests by exploring the data. This question is a bit open-ended, and so requires some thought. Are the assumptions for a stratified test met in this example?

c) Suppose you had only the variables for survival, censoring, bulky disease, and the binary variable `AGE60`. What do you think is the best way to analyze the relationship between survival and bulky disease, accounting for the possible confounder `AGE60`? Assume you cannot use a regression model.

Problem 4 Solution.

a) The stratified log-rank test is highly significant, with p-value 0.00826. The result is clearly different from the unstratified test.

b) The two plots of survival probability by bulky disease, one in younger patients and one in older patients, help explain what is going on. In the younger patients, there is very little difference in survival by bulky disease. The log-rank p-value in this subgroup is 0.39. However, the difference by bulky disease status in older patients is highly significant, with $p = 0.001$.

   The numerator in the stratified log-rank test is made up of two components: the numerators of the separate log-rank statistics in each of the two strata. These two numerators contribute equally, and are scaled by a common variance term. In the stratified test, the difference in the older patients dominates the result, producing a significant difference.

   There are relatively few older patients with bulky disease, so when the analysis is not stratified (i.e., when the groups are combined), the larger number of younger patients with bulk disease dominate that group, causing the survival curves to move closer together.

   The assumptions for a stratified test are not met in this dataset, however. An important assumption in the stratified test is that, while the overall survival may differ between strata, the hazard ratios are assumed to be approximately equal. That is clearly not the case in this dataset. The full interpretation of a stratified test is a bit subtle, however. Significance levels are calculated under the assumption that the null hypothesis is true (that is, that the hazard ratio is 1 within each strata). So the p-value in this case is correct, and the observed differences in the hazard functions are larger than would be expected by chance. However, the more specific conclusion that the p-value is evidence of a large and constant hazard ratio across the two strata would be wrong. This is an example of why it is a very good idea to inspect the data whenever possible.

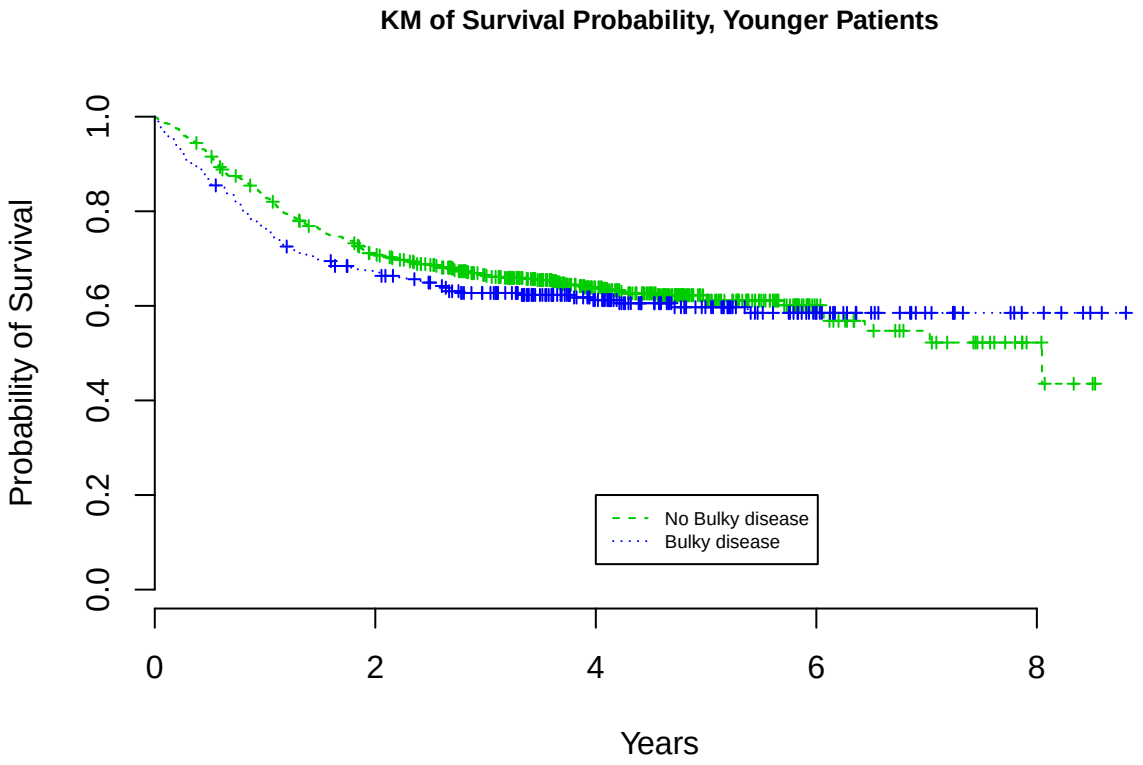
c) There are likely several answers to this question, but the best approach is probably to analyze the younger and older patients separately, as is shown with the separate survival curves and

log-rank tests. It would be possible to use a 4-group log-rank test, but that appraoch would not account for age as a confounder.

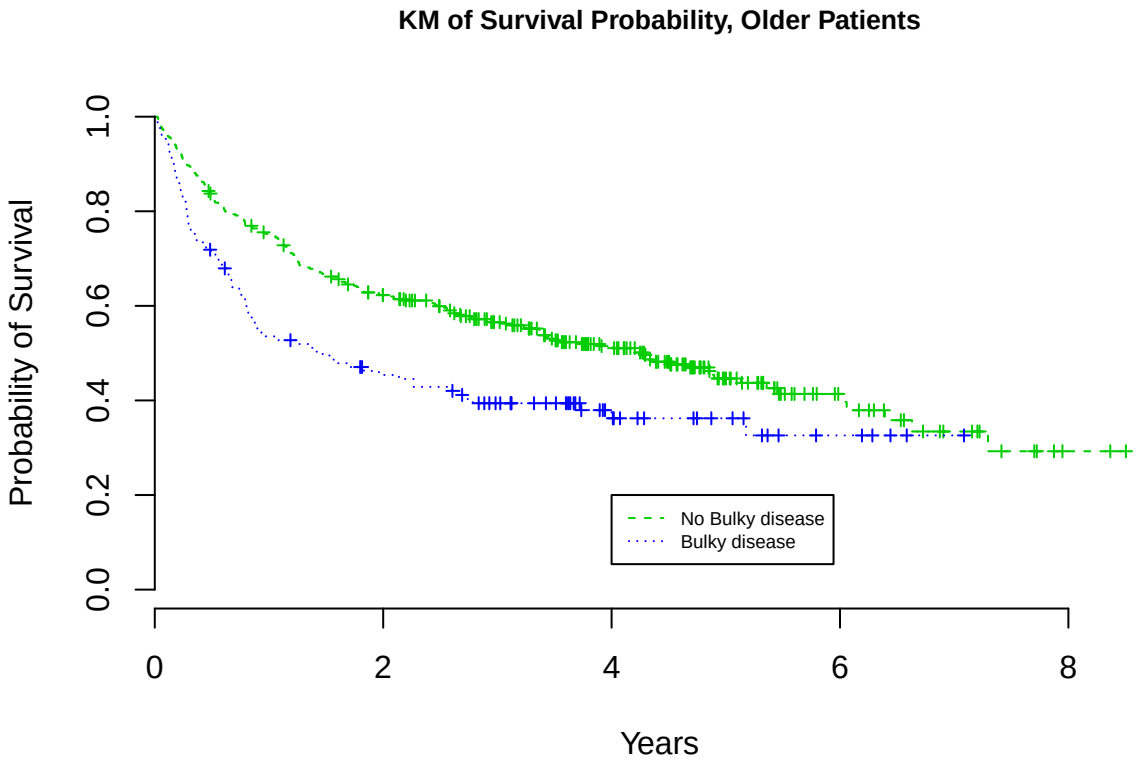
```
library(survival)
library(eventtimedata)
data("lymphoma.prognosis")

#adjusting coding of status variable in lymphoma.prognosis dataframe
lymphoma.prognosis$died = lymphoma.prognosis$SURVIVAL - 1
lymphoma.prognosis$died[lymphoma.prognosis$died == 2] = 0

#conduct stratified log-rank test
survdiff(Surv(SURVTIME, died) ~ BULK + strata(AGE60), data = (lymphoma.prognosis))
```

```
## Call:
## survdiff(formula = Surv(SURVTIME, died) ~ BULK + strata(AGE60),
##     data = (lymphoma.prognosis))
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## BULK=1 961      409      438      1.88      6.98
## BULK=2 424      194      165      4.97      6.98
##
##  Chisq= 7  on 1 degrees of freedom, p= 0.00826
```

```
#subset patients
younger.patients = subset(lymphoma.prognosis, AGE60 == 1)
older.patients = subset(lymphoma.prognosis, AGE60 == 2)

#survival plot, younger patients
lymphoma.young.km = survfit(Surv(SURVTIME, died) ~ BULK, data = (younger.patients))
plot(lymphoma.young.km, lty = 2:3, col = 3:4, mark.time = TRUE,
     xlab = "Years",
     ylab = "Probability of Survival",
     axes = FALSE,
     main = "KM of Survival Probability, Younger Patients",
     cex = 0.6, cex.main = 0.8)
axis(1)
axis(2)
legend(4, .2, c("No Bulky disease", "Bulky disease"),
       lty = 2:3, col = 3:4, cex = 0.6)
```




```
#survival plot, older patients
lymphoma.old.km = survfit(Surv(SURVTIME, died) ~ BULK, data = (older.patients))
plot(lymphoma.old.km, lty = 2:3, col = 3:4, mark.time = TRUE,
     xlab = "Years",
     ylab = "Probability of Survival",
     axes = FALSE,
     main = "KM of Survival Probability, Older Patients",
     cex = 0.6, cex.main = 0.8)
axis(1)
axis(2)
legend(4, .2, c("No Bulky disease", "Bulky disease"),
       lty = 2:3, col = 3:4, cex = 0.6)
```




```
#log-rank test, younger patients
survdiff(Surv(SURVTIME, died) ~ BULK, data = (younger.patients))
```

```
## Call:
## survdiff(formula = Surv(SURVTIME, died) ~ BULK, data = (younger.patients))
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## BULK=1 592      218      225     0.236      0.73
## BULK=2 296      115      108     0.493      0.73
##
##  Chisq= 0.7  on 1 degrees of freedom, p= 0.393
```

```
#log-rank test, older patients
survdiff(Surv(SURVTIME, died) ~ BULK, data = (older.patients))
```

```
## Call:
## survdiff(formula = Surv(SURVTIME, died) ~ BULK, data = (older.patients))
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## BULK=1 369      191    212.4      2.15      10.1
## BULK=2 128       79     57.6      7.93      10.1
##
##  Chisq= 10.1  on 1 degrees of freedom, p= 0.00145
```

Problem 5: UMARU Impact Study (UIS)

Unit 4, the unit on proportional hazards regression, uses data from the University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study. The data are stored as the dataset `uis` in the package `eventtimedata`.

The study was a 5-year collaborative project comprised of two concurrent randomized trials of residential treatment for drug abuse that enrolled a total of 628 patients.

- *Program A:* Randomized 444 subjects to a 3- or 6-month program of health education and relapse prevention. Clients were taught to recognize "high-risk" situations that are triggers to relapse, and taught skills to cope with these situations without using drugs.
- *Program B:* Randomized 184 participants to a 6- or 12-month program with highly structured lifestyle in a communal living setting.

The main outcome in the study was relapse into drug use during the follow-up period.

Variables in `uis`

`id`	Subject ID (1-628)
`age`	Age in years
`beck`	Beck depression score
`hercoc`	Heroine or cocaine use prior to entry
`ivhx`	IV drug use at admission
`ndrugtx`	Number of previous drug treatments
`race`	Race (`white`, `other`)
`treat`	Treatment assignment (`short`, `long`)
`site`	Treatment program
`los`	Length of stay in treatment (days)
`time`	Time to return to drug use (days)
`censor`	Indicator of drug use relapse (1 = `yes`, 0 = `censored`)

This problem examines possible differences in outcome by intervention group. Unit 4 examines those differences after adjusting for some of the covariates.

a) Plot the survival curves for the times to relapse by treatment group. The variable `treat` is a factor variable and is stored with value 1 for `short` and 2 for `long`.

b) Do the curves appear to come from distributions with proportional hazards? Plot the cumulative hazard functions for the two groups to confirm your answer.

c) Test for significant differences between the curves using the log-rank statistic, using a two-sided test with significance level $\alpha = 0.05$.

d) State precisely the null and alternative hypotheses being tested in part c). Does the significant p-value from the log-rank test reflect a systematic difference between the curves?

e) The short and long treatment groups both consist of a mixture of treatment programs. What might be a better analysis approach than the simple, unstratified log-rank test?

f) What do you notice about the time point at which the survival curves begin to diverge? Speculate on the possible reason for the pattern in the curves.

Problem 5 Solution.

a) See below for plot.

b) The survival curves clearly do not display proportional hazards, since the curves are equal to roughly 80 days or so, then diverge. The cumulative hazards show the same picture.

c) The p=value for the log-rank test is vanishingly small (< 0.0001), showing that if differences are extreme enough, the test will detect differences even when the hazards are clearly non-proportional.

d) The log-rank statistic tests the hypotheses $H_0 : S_1(t) = S_2(t)$ for all $t \geq 0$ vs. the alternative $H_A : S_1(t) \neq S_2(t)$ for at least one $t \geq 0$. As noted above, the curves are very different but not systematically so; there are not steadily increasing differences between the curves.

e) The short treatment is either 3 or 6 months, depending on the program (A or B), and long treatment is either 6 months (program A) or 12 months (program B). Treatment program is a natural stratification factor here, at least under the assumption that short versus long duration treatment has the same effect on reducing relapse rate, regardless of program. The stratified log-rank test produces an even more significant p-value than the unstratified test.

f) As noted above, the differences begin to emerge at approximately 80 days. The two groups labeled 'short' and 'long' each have a mixture of treatment periods, with the 'short' treatment being either 3 or 6 months, and the 'long' treatment being either 6 or 12 months.

The survival curve for the short treatment begins dropping rapidly around 80 days; the first patients to stop treatment in the short treatment group ended treatment at 90 days. The first patients to stop treatment in the long treatment group ended therapy at 180 days (6 months), and that is where the survival curve drops.

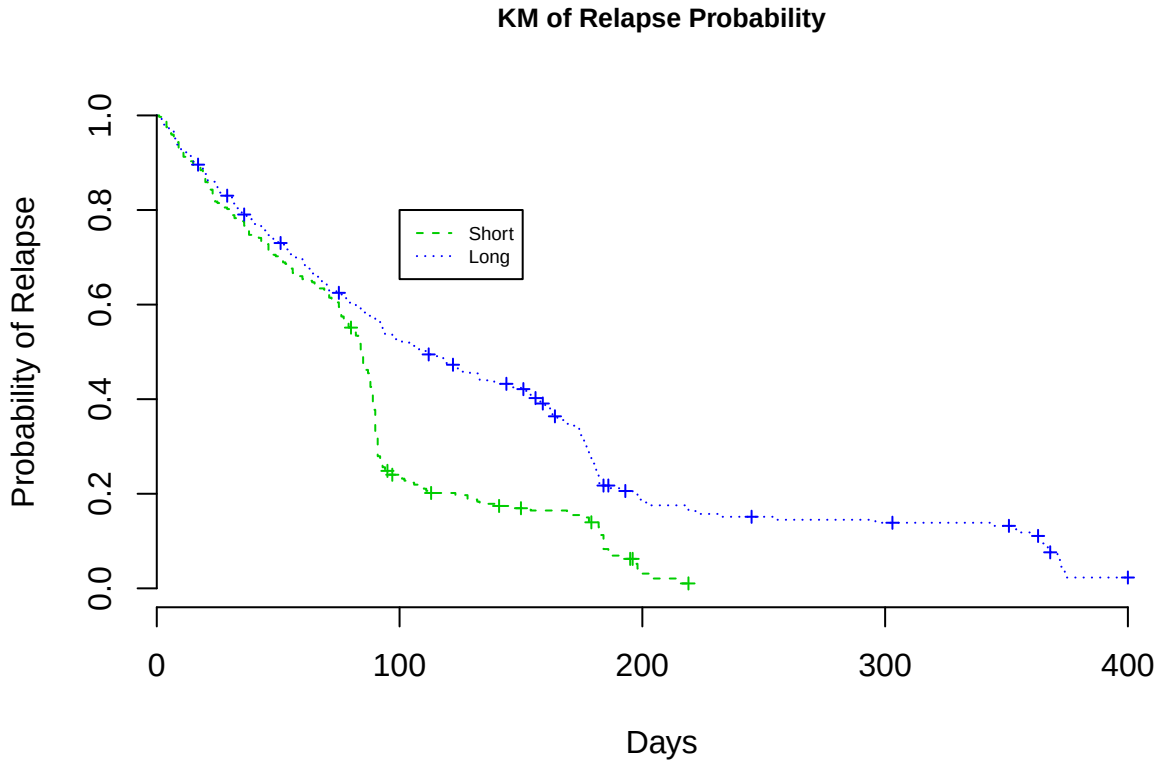
The survival curves suggest that the real benefit is simply being on treatment, not the duration of the treatment. We will examine that conjecture more formally with time-dependent covariates in proportional hazards model.

```
library(survival)
library(eventtimedata)
data(uis)

#determine ordering of coding, needed for labeling curves
levels(uis$treat)
```

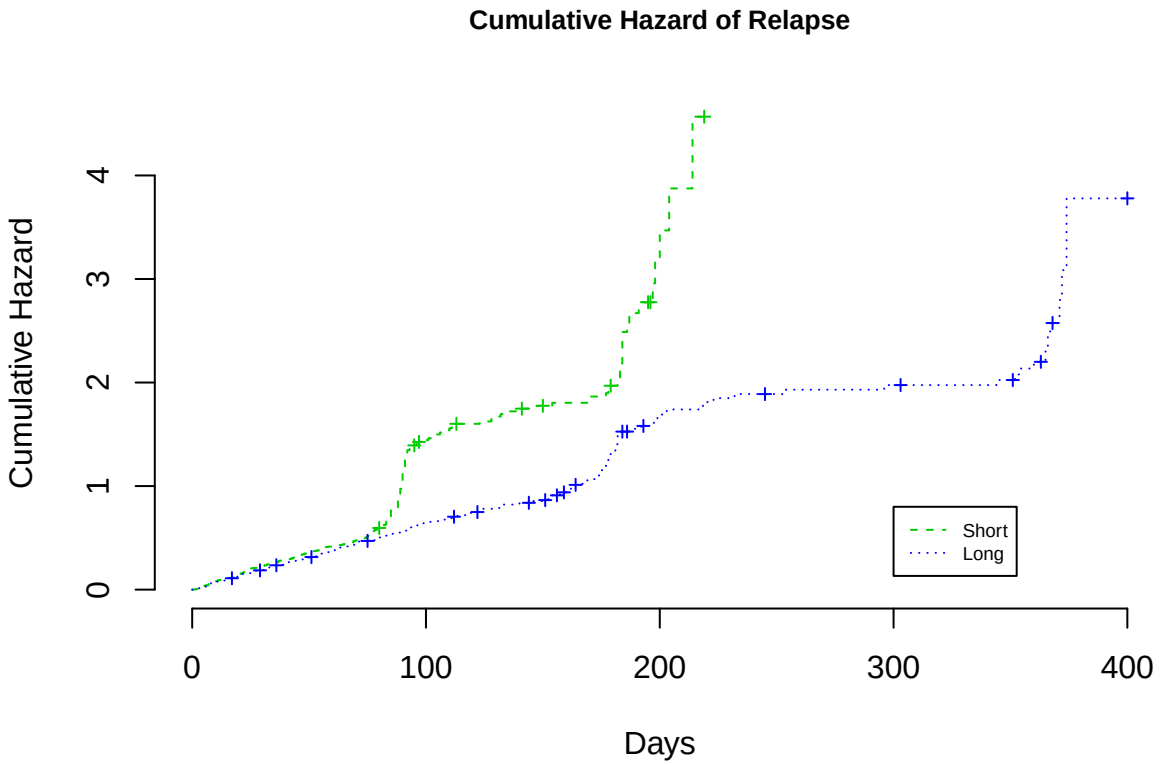
```
## [1] "short" "long"
```

```
#plot survival curves by treatment group
uis.km <- survfit(Surv(los, censor) ~ treat, data = uis)
plot(uis.km, lty = 2:3, col = 3:4, mark.time = TRUE,
     xlab = "Days",
     ylab = "Probability of Relapse",
     axes = FALSE,
     main = "KM of Relapse Probability",
     cex = 0.6, cex.main = 0.8)
axis(1)
```

```
axis(2)
legend(100, .8, c("Short", "Long"), lty = 2:3, col = 3:4, cex = 0.6)
```




```
#plot cumulative hazards by treatment group
plot(uis.km, lty = 2:3, col = 3:4, mark.time = TRUE,
     fun = "cumhaz",
     xlab = "Days",
     ylab = "Cumulative Hazard",
     axes = FALSE,
     main = "Cumulative Hazard of Relapse",
     cex = 0.6, cex.main = 0.8)
axis(1)
axis(2)
legend(300, .8, c("Short", "Long"), lty = 2:3, col = 3:4, cex = 0.6)
```




```
#log-rank test
survdiff(Surv(los, censor) ~ treat, data = uis)
```

```
## Call:
## survdiff(formula = Surv(los, censor) ~ treat, data = uis)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## treat=short  320      265      205      17.5        32
## treat=long   308      243      303      11.8        32
##
##  Chisq= 32  on 1 degrees of freedom, p= 1.55e-08
```

```
#stratified log-rank test
survdiff(Surv(los, censor) ~ treat + strata(site), data = uis)
```

```
## Call:
## survdiff(formula = Surv(los, censor) ~ treat + strata(site),
##     data = uis)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## treat=short  320      265      188      31.6      60.6
## treat=long   308      243      320      18.5      60.6
##
##  Chisq= 60.6  on 1 degrees of freedom, p= 6.77e-15
```