Stats for Clinical Trials, Math 150
Jo Hardin
Homework # 11
Due: Tues 12/7/10

Consider the following observational study on patients with benign breast lesions (conducted at Vanderbilt University and St. Thomas Hospital from 1950-1968). The goal of the study was to assess the importance of various risk factors for breast cancer in women with benign proliferative breast lesions. The study found that 70% of women who underwent a breast biopsy for a benign breast lesion were not at an increased risk of cancer. The variables are:

entage	=	age at entry biopsy
follow	=	years of follow up
pd	=	diagnosis of entry biopsy
		0: nonproliferative disease (No PD)
		1: proliferative disease without atypia (PDWA)
		2: atypical hyperplasia (AH)
fate	=	fate at end of follow up
		0: censored
		1: invasive breast cancer
fh	=	first degree family history of breast cancer
		0: no
		1: yes

1. Plot Kaplan-Meier breast cancer free survival curves for women with entry diagnosis of AH, PDWA, and No PD as a function of years since biopsy. Is this a useful graphic? If not, why not? (Note: if you don't want the censoring ticks on the curve, you can use the command `mark.time=FALSE` within the plot command.)

2. Calculate the breast cancer risk (hazard ratio) of women with AH relative to women without PD. Derive a 95% CI for the HR. Calculate the HR and associated 95% CI for women with PDWA compared to women without PD. [Use `as.factor(pd)` for categorical groups in the model.]

3. Adjust the survival model above for age by including age at entry biopsy as a covariate in your model. Calculate the 95% CI for the HRs. Does the new model give a better estimate of these risks than the model used above? Why or why not?

4. What are the mean ages of entry biopsy for these three diagnoses? Do they differ significantly from each other? Does your answer complicate the interpretation of the preceding results?
```
tapply(entage,pd,mean)
summary(aov(entage~pd))
```

5. Repeat question 3, but now adjust for age using a categorical variable that groups age at biopsy as follows: $\leq 30, 31-40, 41-50, 51-60, > 60$. Compare your answers to the two questions. What are the strengths and weaknesses of the different age adjustments? [Use `ifelse`.]

6. [Unrelated to the data question.] Show that proportional hazards lead to survival curves which are related by a power of R. That is,

$$\begin{aligned} \text{if } \lambda_1(t) &= R\lambda_0(t) \\ \text{then } S_1(t) &= (S_0(t))^R \end{aligned}$$

7. [Unrelated to the data question.] Note that the previous problem also says that proportional hazards leads to proportional log survival curves. Explain what I mean by that. Given the plot in 1, do you think proportional hazards is appropriate? Try plugging in a few values from the plot (just find approximate values by eye), and see if the log of the survival curves are proportional.