

All questions relate to the question of whether the bone scan score for prostate cancer is prognostic of time in remission independently of any information from other commonly used covariates. The data is posted on the class web pages (psa.csv, psa.dta). Note that the variable *inrem* is text (“yes” or “no”), which you will want to recode into a relapse indicator variable (relapse=1 if inrem is “no”).

Perform analyses to determine whether the distribution of time to relapse differs across groups defined by bone scan score. Some code for Stata and R is provide below

1. Before looking at the data, we should decide if we are going to robust standard errors or not.
  - a. What are the benefits of using robust standard errors over classical PH regression?
  - b. What are the benefits of using classical PH regression over PH regression with robust standard errors?
2. Provide suitable descriptive statistics regarding the distribution of time to relapse according to bone scan score.
  - a. Create a Kaplan-Meier estimate of the survival curves by bone scan score (bss)
  - b. From the plot, what is the (approximate) median survival time for subjects with a bone scan score of 2? What about a score of 3?
3. List the variables obstime and relapse for the 5 subjects with a bss==1.
  - a. Be able to explain what each row of data represents
  - b. From these points, calculate (by hand) the Kaplan-Meier estimate of the survival curve. Compare your answer to the plot in 2.a.
4. Perform a proportional hazards regression comparing the instantaneous risk of relapse across groups defined by bone scan score when modeled as a continuous, untransformed variable.
  - a. Provide an interpretation for the slope.
5. Create a new indicator variable (bss3) that takes on the value 1 if the bone scan score is 3 and 0 otherwise. Perform a proportional hazards regression comparing the instantaneous risk of relapse across groups defined by bss3.
  - a. Provide an interpretation for the slope for bss3
  - b. Why might you *a priori* prefer inference based on the continuous bss? What considerations would make you prefer inference based on indicator variable bss3 instead? (This question refers to scientific considerations)
  - c. After looking at the data and descriptive statistics (*a posteriori*), why might you prefer the model with bss included as a continuous predictor rather than bss3 as an indicator variable? (This question refers to statistical considerations)
6. Create a new indicator variables (bss2) that takes on the value 1 if the bone scan score is 2 and 0 otherwise. Perform a proportional hazards regression comparing the instantaneous risk of

relapse across groups define by bss2 and bss3.

- a. Provide an interpretation for the slopes: bss2 and bss3
  - b. Why might you *a priori* prefer inference based on the continuous bss? What considerations would make you prefer inference based on indicator variables bss2 and bss3 instead? (This question refers to scientific considerations)
  - c. After looking at the data and descriptive statistics (*a posteriori*), why might you prefer the model with bss included as a continuous predictor rather than bss2 and bss3 as indicator variables? (This question refers to statistical considerations)
7. For all three models, provide an interpretation (estimate, CI, verbal description) comparing the instantaneous risk of relapse across subjects with a bone scan score of 3 to bone scan score of 1.

### **Stata notes**

“stset obstime relapse” will define the outcome

“stcox predictor, [robust]” will run the Cox PH regression model, with or without robust SEs

“sts graph, by(bss)” will create the Kaplan-Meier estimate of the survival curve by bss group

### **R notes**

Robust standard errors is built into the survival package

- “library(Design)” or “library(survival)” will load this package. I always have Design loaded

“Surv(obstime, relapse)” defines the outcome, which you can use with the following functions

- “coxph(Surv(obstime, relapse) ~ predictor, robust=TRUE)” Cox PH regression model with robust standard errors
- “coxph(Surv(obstime, relapse) ~ predictor)” Classical Cox PH regression model
- “plot(survfit(Surv(obstime, relapse) ~ predictor))” Plot Kaplan Meier estimate of the survival curve

The R survival plot makes it difficult to discern groups by default. For the bone scan score question, I recommend adding a line type specification (e.g. “lty=2:4”) for the plot and a legend using a second command

- `plot(survfit(Surv(obstime, relapse) ~ bss, data=d1), lty=2:4, xlab="Time", ylab="Survival")`
- `legend("topright", c("BSS = 1", "BSS = 2", "BSS = 3"), lty=2:4, inset=0.05)`