Bios 312: Modern Regression Analysis
Spring, 2012

Homework #4
March 13, 2012

Written problems due at the beginning of class, Thursday, March 22, 2012.

Part 1: Regression model theory

Let $Y_i \sim N(\beta_0 + \beta_1 * X_i, \sigma^2)$ as in simple linear regression.

1. Find the maximum likelihood estimates of β_0, β_1, and σ^2. Show your work.

2. Compare the MLEs obtained in (1) to the estimate of obtained for minimizing the residual sums of squares. You do not need to derive the least squares estimates (they are given in the notes).

3. Show that $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \widehat{\boldsymbol{B}}$ where

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \widehat{\boldsymbol{B}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the MLEs found in (1)

4. Let Y_{avg} be the mean of new observations taken at fixed, known covariate X_{avg} in a simple linear regression model where $Y_{avg} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{avg}$. Find the $Var[Y_{avg}]$. At what value of X_{avg} will $Var[Y_{avg}]$ be smallest?

5. Let Y_{new} be the predicted value of a new Y taken at fixed, known covariate X_{new} in a simple linear regression model where $Y_{new} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{new} + \epsilon$. Find the $Var[Y_{new}]$. At what value of X_{new} will $Var[Y_{new}]$ be smallest?

6. Use the $Var[Y_{avg}]$ to construct a (1-α)% confidence interval for the expected value of Y taken at X_{avg}. Use the $Var[Y_{new}]$ to construct a (1-α)% confidence interval for the predicted value of a new Y at X_{new}. Which of these two confidence intervals will always be wider? Why?

Part 2: Applied Survival analysis

The following 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"). This is the same dataset you used for homework 3

Perform analyses to determine whether the distribution of time to relapse differs across groups defined by bone scan score. Some code for survival analysis in Stata and R is provided after the questions.

1. Provide suitable descriptive statistics regarding the distribution of time to relapse according to bone scan score.
2. 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. Provide interpretation for the slope.
3. Create a new indicator variable (bss3) that takes on the value 1 if the bone scan score is 3 and 1 otherwise. Perform a proportional hazards regression comparing the instantaneous risk of relapse across groups defined by bss3. Provide interpretation for the slope.
4. 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)
5. 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. For models (2) and (3), 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 are built into the survival library in R. Loading the rms library also load the survival library

- "library(rms)" or "library(survival)" will load the needed library. I always have rms 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. Use robust=FALSE for classical 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. You could also use different colors, symbols, etc. to distinguish the lines.

- 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)