P8110: Applied Regression II

Camille Okonkwo

Uni: co2554 Spring 2024

Homework #5 [14 points]

Due on Feb 26, 11:59**AM**

**NOTE: Cut and paste relevant SAS/R output to appropriate places in the texts of your solutions. Attach the SAS/R codes to the end of your homework.**

1. We continue to use the “**hwdata1.csv**” dataset for breast cancer recurrence. We denote the model that we fit in HW #4 using the hormone therapy groups as the only covariate as Model 1.
   1. Fit a Cox model with hormone therapy groups, age, menopause status, tumor size, and number of nodes as covariates (Model 2). Is Model 2 significantly better than Model 1, at a significant level of *α* = 0*.*05? Write down Model 2. State the null and alternative hypothesis, test statistic (state which test was used), p-value, degrees of freedom, and conclusion. [4 points]

Model 2:

Our hypotheses:

Test Statistic: **Likelihood Ratio Test**

**A computer screen with white text

Description automatically generated**



Degrees of freedom:

P-value:

Conclusion:

**At the 0.05 level of significance, we reject H0. There is sufficient evidence to conclude that Model 2 is better than Model 1.**

* 1. Fit a Cox model with hormone therapy groups, age, menopause status, tumor size, number of nodes, and the interaction between hormone therapy groups and number of nodes as covariates (Model 3). Write down Model 3. Hand calculate the hazards ratio and 95% CI between patients under hormone therapy and patients not under hormone therapy who has 5 nodes involved (average number of nodes among all patients). Also obtain the 95% CI using SAS/R. Provide code and relevant output. [5 points]

Model 3:

Hazards Ratio and 95% CI:

Hazard Ratio Estimate and 95% CI:

**95% CI using R code:**

# Hazard ratio and 95% CI

```{r}

library(rms)

dd = datadist(hwdata1)

options(datadist="dd")

ctr = contrast(cph(Surv(rectime, censrec) ~ hormone + age + menopause + size + nodes + hormone\*nodes,

data = hwdata1,

ties = "efron"),

list(hormone = 1, nodes = 5),

list(hormone = 2, nodes = 5))

print(ctr, fun=exp)

```

**Output:**

A black screen with white text

Description automatically generated



* 1. Draw the survival curve for a patient under hormone therapy, at age53, being menopausal, having tumor of size 25mm, and having 3 nodes involved, based on Model 3. What is the probability that this patient survives more than three years (1095 days)? Provide code and relevant output. [3 points]

**R Code:**

# 1c) Survival curve and probability

```{r}

library(ggsurvfit)

# curve for hormone = 1, age = 53, menopause = 1, size = 25, nodes = 3, based on model3

survfit2(model3,

newdata = data.frame(

hormone = factor(1, levels = c("2", "1")),

age = 53,

menopause = factor(1, levels = c("2", "1")),

size = 25,

nodes = 3)) |>

ggsurvfit() +

scale\_y\_continuous(limits=c(0,1))

# output the survival table

model3.newdata = survfit(model3,

newdata = data.frame(

hormone = factor(1, levels = c("2", "1")),

age = 53,

menopause = factor(1, levels = c("2", "1")),

size = 25,

nodes = 3))

# what is the probability that this patient survives more than three years (1095 days)?

surv\_1095 = summary(model3.newdata, time = 1095)

prob\_surv = 1 - surv\_1095$surv

```

**Output:**

A graph showing a line

Description automatically generated

A computer code with white text

Description automatically generated



A black background with white text and orange letters

Description automatically generated



**There is a 29.46% probability that a patient under hormone therapy, at age 53, being menopausal, having a tumor size of 25mm, and having 3 modes will survive more than three years (1095 days).**

1. Read the paper [[link]](https://academic.oup.com/oncolo/advance-article/doi/10.1093/oncolo/oyad197/7243567) from *The Oncologist*. Compared to the general strategy for survival data analysis we discussed in Lecture 5, which steps have been included in this paper? What test procedures and statistical models have been used? [2 points]

Step 0: Data preparation

* Descriptive statistics analysis to summarize the analysis population.

Step 1: Kaplan-Meier curves and log-rank test statistics were used to compare overall survival time and progression free survival between patient groups

Step 2 & 3: “Cox proportional hazards models were fitted to investigate **univariable/multivariable** associations between covariates and OS and PFS. Covariates that were statistically significant in univariable analysis were included in subsequent multivariable analysis.”

Step 7: Model Evaluation

* Researchers stratified the Cox models by site to address variations in patient population and sequencing platforms across institutions. They also exclude patients to account for left truncation bias.

Test Procedures and Statistical Models Used:

Descriptive Statistics:

* Wilcoxon rank sum test, Fisher’s exact test, or Pearson’s chi-squared tests were used to calculate baseline characteristics.

Survival Analysis:

* Kaplan-Meier curves and log-rank tests for OS and PFS.
* Cox proportional hazards models for univariable and multivariable analysis.
* Stratification by site to address variations across institutions.
* Left truncation bias addressed by excluding certain patients.

Exploratory Analysis:

* Used a piece-wise structure to describe the OS hazard ratio between KRASG12C and non-G12C groups.

1