Final Examination, Survival Analysis, 2025

Problem.

Primary biliary cirrhosis (PBC) of the liver is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 50-cases-per-million population. The primary pathologic event appears to be the destruction of interlobular bile ducts, which may be mediated by immunologic mechanisms. The following briefly describes data collected for the Mayo Clinic trial in PBC of the liver conducted between January 1974 and May 1984 comparing the drug D-penicillamine (DPCA) with a placebo. The first 312 cases participated in the randomized trial of DPCA versus placebo and contain largely compete data. An additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so there are data here on an additional 106 cases as well as the 312 randomized participants. The data consists of 14 variables with 308 observations after removing all missing values from the full dataset. Here comes the explanation of the variables in the dataset.

Col 1 fu_days - observed survival times
Col 2 status - (0=censored; 1=event)
Col 3 drug - treatement indicator (0=DPCA; 1=placebo)
Col 4 age - age at baseline (years)
Col 5 sex - (0=male; 1=female)
Col 6 ascites - presence of ascites (0=no; 1=yes)
Col 7 hepatom - presence of hepatomegaly (0=no; 1=yes)
Col 8 edema - edema status (0=no edema and no diuretic therapy; 1=edema present without diuretics; 2=edema despite diuretic therapy)
Col 9 bili - serum bilirubin (mg/dL)
Col 10 albumin - albumin (gm/dL)
Col 11 alk - alkaline phosphatase (U/L)
Col 12 sgot - SGOT (U/mL)
Col 13 platelet - platelets per cubic ml/1000
Col 14 protime - prothrombin time (seconds)

Tasks. Do survival analysis for the dataset with R software.

1. Preliminary Analysis & Kaplan-Meier Curves

1.1 Load and summarize the dataset. Provide descriptive statistics for all variables.

1.2 Plot a Kaplan-Meier(KM) survival curve for overall survival.

1.3 Construct KM curves for at least **two categorical covariates** (e.g., drug, ascites, edema). Perform log-rank tests and briefly summarize whether survival appears different across groups.

2. PH Assumption Checking

For **each covariate**, fit a *univariate Cox model* first and test PH assumption using:

- Goodness-of-fit test using Schoenfeld residual (cox.zph)
- log(-log(S(t))) plots

2.1 Report for each variable:

- Goodness-of-fit test p-value
- Whether PH assumption is satisfied (α= 0.05)

2.2 Categorize variables as follows:

- PH assumption satisfied → eligible for Cox PH model
- PH assumption violated → cannot be included in standard Cox model

3. Build a Survival Model Based on PH Results

Using Step 2 results:

3.1 Fit a multivariate Cox PH model including *only PH-satisfied variables.*

Report:

- coefficients (β)
- HR and 95% CI
- Wald p-values

Interpret at least **three** significant predictors.

3.2 Handle PH-violating variables using alternative methods

For variables violating PH, choose one of the following modeling strategies (and implement at least one):

Strategy 1: Stratified Cox Model (SC model)

- stratify by PH-violating variable(s)
- allows baseline hazard to differ by strata

Strategy 2: Extended Cox Model(Time-varying Coefficients)

- Example: Add an interaction with log(time) (e.g. edema*log(t))

Strategy 3: Use parametric modeling (e.g. PH model or AFT model)

3.3 Compare models (standard Cox model / SC model or extended Cox model) using at least one metric:

- Likelihood ratio test
- AIC comparison
- Improvement in residual patterns

Write a short interpretation on whether handling PH-violating terms improved model adequacy.

4. Model Diagnostics & Final Conclusions

4.1 Check residual diagnostics for the final chosen model:

- Martingale or deviance residuals for functional form
- dfbeta for influential observations

4.2 Provide overall conclusions:

- Which variables significantly impact survival?
- What is the effect direction? (clinical interpretation recommended)
- Which modeling strategy best handled PH-violating effects?