Survival Analysis of Lung Cancer Patients in Sweden

*Analysis of Survival Data/Survival Analysis (Joint Section) HT2025

Liang-Jen Huang
Department of Statistics, Uppsala University

Hadeel Elhassan
Department of Statistics, Uppsala University

I. Introduction

II. Data Description

The dataset used for this assignment is the *Lung Cancer Dataset*, obtained from Kaggle[1]. It contains information related to lung cancer mortality among the European population and provides a comprehensive collection of patient data, focusing on individuals who have been diagnosed with lung cancer.

After extracting the observations corresponding to Sweden, a total of 33161 observations were selected for analysis. Since the starting and ending times of observation differ among patients, the data are characterized by generalized right censoring. Among these patients, 7165 were still alive at the end of treatment, while 25996 people had died. Hence, the degree of censoring is approximately 21.6%, representing the proportion of patients whose event (death) had not yet occurred by the end of observation.

The dataset consists of 14 explanatory variables and one outcome variable survived. All variables are currently included in the analysis. The time-related variables are diagnosis_date, end_treatment_date, and age. The main grouping variables include cancer_stage, smoking_status, and treatment_type. Other variables are treated as potential covariates; however, in further analyses, the roles of grouping variables and covariates may be interchanged depending on the analytical results. A detailed description of all variables is provided in Table I.

III. Method

A. Method 1

B. Method 2

IV. Result

V. Conclusion

TABLE I
Description of variables used in the analysis

Variable	Description
diagnosis_date	Date of lung cancer diagnosis.
end_treatment_date	Date of treatment end or death.
survived	Survival status (0 = death, 1 = alive).
cancer_stage	Stage of lung cancer (0 = Stage I, 1 = Stage II, 2 = Stage III, 3 = Stage IV).
gender	Gender of the patient (0 = Female, 1 = Male).
age	Age at diagnosis (years).
bmi	Body Mass Index (kg/m^2).
cholesterol_level	Cholesterol level (mg/dL).
smoking_status	Smoking behavior (0 = Never Smoked, 1 = Passive Smoker, 2 = Former Smoker, 3 = Current Smoker).
family_history	Family history of cancer (0 = No, 1 = Yes).
hypertension	High blood pressure (0 = No, 1 = Yes).
asthma	Asthma condition (0 = No, 1 = Yes).
cirrhosis	Liver cirrhosis (0 = No, 1 = Yes).
other_cancer	History of other cancers (0 = No, 1 = Yes).
treatment_type	Type of treatment received (0 = Radiation, 1 = Chemotherapy, 2 = Surgery, 3 = Combined).

[1] https://www.kaggle.com/datasets/khwaishsaxena/lung-cancer-dataset?resource=download